RESEARCH ARTICLE

Genetic Epidemiology
OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Multiple SNP Set Analysis for Genome-Wide Association Studies Through Bayesian Latent Variable Selection

Zhao-Hua Lu,[1] Hongtu Zhu,[1,2] Rebecca C. Knickmeyer,[3] Patrick F. Sullivan,[4] Stephanie N. Williams,[4] Fei Zou,[1]* and for the Alzheimer's Disease Neuroimaging Initiative†

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, North Carolina, United States of America; [2]Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, North Carolina, United States of America; [3]Department of Psychiatry, University of North Carolina at Chapel Hill, North Carolina, United States of America; [4]Department of Genetics, University of North Carolina at Chapel Hill, North Carolina, United States of America

Received 25 March 2015; Revised 23 July 2015; accepted revised manuscript 18 August 2015.
Published online 30 October 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21932

ABSTRACT: **The power of genome-wide association studies (GWAS) for mapping complex traits with single-SNP analysis (where SNP is single-nucleotide polymorphism) may be undermined by modest SNP effect sizes, unobserved causal SNPs, correlation among adjacent SNPs, and SNP-SNP interactions. Alternative approaches for testing the association between a single SNP set and individual phenotypes have been shown to be promising for improving the power of GWAS. We propose a Bayesian latent variable selection (BLVS) method to simultaneously model the joint association mapping between a large number of SNP sets and complex traits. Compared with single SNP set analysis, such joint association mapping not only accounts for the correlation among SNP sets but also is capable of detecting causal SNP sets that are marginally uncorrelated with traits. The spike-and-slab prior assigned to the effects of SNP sets can greatly reduce the dimension of effective SNP sets, while speeding up computation. An efficient Markov chain Monte Carlo algorithm is developed. Simulations demonstrate that BLVS outperforms several competing variable selection methods in some important scenarios.**
Genet Epidemiol 39:664–677, 2015. © 2015 Wiley Periodicals, Inc.

KEY WORDS: **Bayesian variable selection; GWAS; linkage disequilibrium blocks; imaging phenotypes**

Introduction

Genome-wide association studies (GWAS) have been fruitful in establishing the association between single-nucleotide polymorphisms (SNPs) and many complex traits. However, several key characteristics of SNP data greatly undermine the performance of GWAS in detecting causal genetic markers for various diseases. First, it is increasingly recognized that variation in complex traits represents, in part, the joint effect of many variants with individually small effect sizes, making them challenging to identify [Visscher et al., 2012]. Second, causal SNPs may not be genotyped directly. Their effects may be partly revealed by correlated SNPs that are genotyped, but the effects of these surrogate SNPs are even smaller, which makes them harder to detect. Third, SNPs may be highly correlated. Ignoring their correlation may increase the number of false positives and negatives. Finally, a group of SNPs with small or no marginal effects may have strong joint genetic effects on phenotype. Screening procedures, which usually use marginal information of SNPs to achieve dimension reduction, are likely to remove these SNPs with "weak" marginal effects. Marginal association approaches for a single SNP or SNP set have been widely adopted partly because joint association of multiple SNPs/SNP sets is computationally challenging. For these reasons, it is critical to develop much more efficient ways to simultaneously extract information from all SNPs in order to increase detection power, while capturing correlations and interactions among SNPs.

Various methods based on SNP sets have been proposed to improve the performance of GWAS studies [Fridley and Biernacka, 2011; Skarman et al., 2012]. The use of SNP sets [see Tzeng et al., 2011, for an overview] and gene sets/pathways [see Fridley and Biernacka, 2011; Wang et al., 2010, for overviews] instead of individual SNPs for gene-trait association mapping is attractive for the following reasons. First, combining information across similar SNPs may increase detection power. Second, combining multiple SNPs in linkage disequilibrium (LD) may recover the power of detecting the correlated latent causal SNPs better than single-SNP

Supporting Information is available in the online issue at wileyonlinelibrary.com.

†Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

*Correspondence to: Fei Zou, Department of Biostatistics, University of North Carolina, McGavran-Greenberg Hall CB 7420, Chapel Hill, NC 27599, USA. E-mail: feizou@email.unc.edu

analysis [Schaid et al., 2002]. Third, the correlations among SNP sets are usually small, and thus it alleviates the inflation of the false discovery rate.

Many methods have been proposed to carry out marker-set association, which include, but are not limited to, weighted sum of genotypes [Price et al., 2010; Wang and Elston, 2007], *U*-statistics approaches [Tzeng et al., 2003; Wei et al., 2008], and variance-component (VC) methods [Tzeng and Zhang, 2007; Wu et al., 2010]. In the VC methods, the genetic effects of SNPs within genes, pathways, or haplotype/LD blocks are modeled through a set of random variables, and testing the marginal association between an SNP set and a trait is equivalent to test whether the variance of the random variable is zero. The VC methods have been shown to have strong power in many situations when evaluating genetic main effects [Ballard et al., 2010; Fridley et al., 2010; Wu et al., 2010]. In addition, interaction effects of SNPs in an SNP set can be potentially represented even if they are not explicitly included in the model [Wu et al., 2010].

Despite the popularity and power of the SNP-set and gene-set association methods, further improvements can be made. Most of these methods [e.g., Chapman and Whittaker, 2008; Gauderman et al., 2007; Mukhopadhyay et al., 2010; Pan, 2011] only study the marginal association of each set with a trait. In the context of SNP association analysis, studying the association of multiple SNPs simultaneously has some advantages over studying marginal associations [Guan and Stephens, 2011; He and Lin, 2011]. Analogously, studying the association of multiple SNP sets jointly can be beneficial as well. Compared with single SNP set analysis, joint SNP-set association mapping accounts for the correlation among SNP sets. Given that the causal SNP sets are in the model, false-positive signals of SNP sets may be suppressed. Moreover, the joint SNP-set association dramatically reduces the burden of controlling for multiple comparisons, leading to improved power. Bayesian variable selection (BVS) [George and McCulloch, 1993] is one of the feasible methods for joint modeling all genetic variables [Bhadra and Mallick, 2013; Fridley, 2009; Hoggart et al., 2008; Logsdon et al., 2012]. However, in most cases, BVS is used for selecting SNPs instead of SNP sets or gene sets.

The goal of this article is to reformulate the association mapping of multiple SNP sets as a simultaneous regression of a large number (say 10^5) of SNP sets (or latent variables) on trait in a linear mixed-effects modeling framework. Our multiple SNP set mapping is based on a novel extension of BVS for selecting high-dimensional variables [Barbieri and Berger, 2004; Liang et al., 2013]. See O'Hara and Sillanpää [2009] for an overview of BVS methods. We propose a Bayesian latent variable selection (BLVS) procedure, which simultaneously selects "significant" latent variables. An efficient Markov chain Monte Carlo (MCMC) algorithm is developed to dramatically reduce the complexity of two time-consuming steps in each MCMC iteration from $O(qn^3)$ and $O(q|\boldsymbol{\delta}|^3)$ to $O(qn)$ and $O(q|\boldsymbol{\delta}|^2)$, where n is the number of subjects, q is the number of SNP sets, and $|\boldsymbol{\delta}|$ is the number of casual SNP sets. Simulation studies demonstrate that BLVS outperforms single-SNP association and Lasso [Friedman et al., 2010] in all scenarios examined, whereas BLVS outperforms the group Lasso [Yang and Zou, 2015] in most scenarios considered. BLVS is applied to the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset on brain volumetric measurements. BLVS is also applied to the Swedish Schizophrenia Study [Ripke et al., 2013]. BLVS is able to detect several key genes found to be associated with Alzheimer's disease (AD) and schizophrenia in the existing literature as well as several novel genes not reported before.

Materials and Methods

BLVS for SNP-Set Association

We propose the following linear mixed-effects model for joint association of multiple SNP sets with a quantitative trait. For the ith subject ($i = 1, \ldots, n$)

$$y_i = \mathbf{x}_i^T \mathbf{a} + \sum_{j=1}^{q} \gamma_j b_{ij} + \epsilon_i, \tag{1}$$

where y_i is a phenotype of interest, $\mathbf{a}$ and $\mathbf{x}_i$ are $d \times 1$ vectors of regression coefficients and covariates, and $\epsilon_i \sim N(0, \psi)$ is a Gaussian random error. The random variable b_{ij} represents the genetic information of the jth SNP set, and γ_j is the corresponding SNP-set effect. More specifically, let $\mathbf{u}_{jk}$ be an $n \times 1$ vector containing the kth normalized SNP in the jth block of all subjects, $\mathbf{U}_j = (\mathbf{u}_{j1}, \ldots, \mathbf{u}_{jr_j})$ be all SNPs in the jth block of all subjects, and $\mathbf{b}_j = (b_{1j}, \ldots, b_{nj})$. The vector of random effects $\mathbf{b}_j$ is derived from $\mathbf{U}_j$ through

$$\mathbf{b}_j \sim N(\mathbf{0}, \boldsymbol{\Sigma}_j), \tag{2}$$

where $\boldsymbol{\Sigma}_j = \mathbf{U}_j \mathbf{U}_j^T$ characterizes all subjects' correlation structure in the jth SNP set. It is also assumed that $\mathbf{b}_j$ is independent of ϵ_i and $\mathbf{b}_{j'}$ for $j' \neq j$. The $\mathbf{b}_j$ characterizes the correlation structure and the average direction of the $\mathbf{u}_{jk}$s, and thus it represents the joint information of $\mathbf{u}_{jk}$s. Extracting random effects from SNPs has been used in the literature. For instance, Kang et al. [2008] and Zhou et al. [2013] used a single random effect to characterize the correlation among subjects, which can handle relatedness due to population structure and family structure, among others. Wu et al. [2010] used a single random effect to test the marginal effect of a single SNP set. In model (1), however, we use multiple random effects to represent a large number of SNP sets and study their joint association with the complex trait. Through selecting nonzero γ_js, model (1) can lead to the mapping of all q SNP sets on the trait.

Model (1) may increase the power of detecting causal SNP sets through accounting for the correlations among SNP sets. Although $\mathbf{b}_j$ and $\mathbf{b}_{j'}$ are assumed to be independent for $j' \neq j$ in the prior distributions (2), the posterior means of causal $\mathbf{b}_j$ and $\mathbf{b}_{j'}$ are correlated given that $\mathbf{U}_j$ and $\mathbf{U}_{j'}$ are correlated. A small simulation study is presented in the supplementary material for demonstration.

We propose a BLVS procedure, which can be viewed as an extension of BVS for selecting fixed covariates. Our selection of $\mathbf{b}_j$ is achieved by imposing a sparse structure on $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)$ through the spike-and-slab prior as follows:

$$\gamma_j | \delta_j, \sigma_j^2 \sim (1 - \delta_j) I_0 + \delta_j N(0, \sigma_j^2), \tag{3}$$

$$\delta_j \sim \text{Bernoulli}(\pi), \quad \sigma_j^2 \sim \text{Inverse Gamma}(a_{\sigma 01}, a_{\sigma 02}), \tag{4}$$

where π, $a_{\sigma 01}$ (shape), and $a_{\sigma 02}$ (rate) are hyperparameters. The prior of γ_j is a mixture of a normal distribution and a point mass distribution at 0. The indicator variable δ_j is equal to either 0 or 1, indicating the mixture component affiliation of γ_j. We use δ_j to test the jth SNP set because $P(\gamma_j \neq 0) = P(\delta_j = 1)$. In the BVS literature, the posterior inclusion probability $P(\gamma_j = 1|\mathbf{y})$ has been widely used as the selection criteria of γ_j. For instance, one widely used criterion is to retain all $\mathbf{b}_j$ with $P(\gamma_j = 1|\mathbf{y}) > 0.5$, which results in the median probability model [Barbieri and Berger, 2004]. Compared with other Bayesian shrinkage priors [Park and Casella, 2008], the spike-and-slab prior has the important advantage that it shrinks many γ_js to zero exactly. The "significant" γ_js in Equation (1) usually suggest correlation instead of causal relation. However, in GWAS, it is usually reasonable to interpret identified SNPs as causal SNPs [Guan and Stephens, 2011]. For convenience, we follow this convention and term the identified SNP sets as causal SNP sets.

From the computational perspective, the complexity of many operations, such as matrix multiplication and inversion, depends on the dimension of nonzero γ_js, which leads to a huge computational saving. Our formulation also reduces the computational complexity from $O(q^\alpha)$ to $O(|\boldsymbol{\delta}|^\alpha)$, where $|\boldsymbol{\delta}|$ is the number of nonzero γ_js and α is a positive number. Given that $|\boldsymbol{\delta}| << q$, the computation involved here is feasible.

The hyperparameter π controls the prior belief of sparsity of $\boldsymbol{\gamma}$, and a small number is often given in high-dimensional problems. The σ_j represents the prior information of the scale of γ_j. Instead of fixing σ_j^2, we learn σ_j^2 from data through Equation (4). Let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_q)$ and $\boldsymbol{\sigma}^2 = (\sigma_1, \ldots, \sigma_q)$ for $j = 1, \ldots, q$. We assume that γ_j, δ_j, and σ_j are independent of each other such that $P(\boldsymbol{\gamma}|\boldsymbol{\delta}, \boldsymbol{\sigma}^2) = \prod_{j=1}^q P(\gamma_j|\delta_j, \sigma_j^2)$, $P(\boldsymbol{\delta}) = \prod_{j=1}^q P(\delta_j)$, and $P(\boldsymbol{\sigma}^2) = \prod_{j=1}^q P(\sigma_j^2)$.

Among studies using BVS, our proposed BLVS is closely related to Guan and Stephens [2011], which studied various aspects of BVS in GWAS extensively. They used a model and a prior setting similar to Equations (1), (3), and (4) with MCMC algorithms. In comparison, they performed association based on SNPs instead of SNP sets. Hence, b_{ij} in model (1) is the observed SNPs instead of latent effects of SNP sets. After model fitting, the inference about the inclusion of SNPs is done through moving windows, which is conceptually related to SNP sets. In comparison, we directly consider SNP sets in the model estimation and inference. In the presence of highly correlated local SNPs, the signals of an SNP set may be stronger than each SNP in the SNP set. Hence, SNP sets may be easier to detect compared with SNPs, and combining information of SNPs in an SNP set may improve detection power. Moreover, the MCMC algorithms are different due to the latent variables.

It is worth noting that selecting random effects in the linear mixed-effects model was previously studied by Chen and Dunson [2003] in a different context. Random effects were used to account for the within-group correlation motivated by a longitudinal study. They considered selecting from tens of random effects with unknown covariance structures. In contrast, our model is motivated from GWAS and the dimension of random effects is much higher. Each latent variable is used to represent the joint effect of SNPs in an SNP set, and the correlation structure is assumed known and determined by the SNP genotypes.

An Efficient MCMC Algorithm

Let $\boldsymbol{\theta} = \{\mathbf{a}, \boldsymbol{\gamma}, \psi, \boldsymbol{\delta}, \boldsymbol{\sigma}^2\}$. It is assumed that $P(\boldsymbol{\theta}) = P(\mathbf{a})P(\psi)P(\boldsymbol{\gamma}|\boldsymbol{\delta}, \boldsymbol{\sigma}^2)P(\boldsymbol{\delta})P(\boldsymbol{\sigma}^2)$,

$$\mathbf{a} \sim N(\mathbf{a}_0, \boldsymbol{\Sigma}_{a0}), \text{ and } \psi \sim \text{IG}(a_{01}, a_{02}), \tag{5}$$

where $\mathbf{a}_0$, $\boldsymbol{\Sigma}_{a0}$, a_{01} and a_{02} are hyperparameters. To develop Bayesian inference, we augment $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_q)$ to the observed data $\mathbf{y} = (y_1, \ldots, y_n)$, and use an MCMC algorithm [Gelfand and Smith, 1990] to draw samples from $P(\boldsymbol{\theta}, \mathbf{B}|\mathbf{y}) \propto P(\mathbf{y}, \mathbf{B}|\boldsymbol{\theta})P(\boldsymbol{\theta})$, where

$$\begin{aligned} P(\mathbf{y}, \mathbf{B}|\boldsymbol{\theta}) &= P(\mathbf{y}|\mathbf{a}, \boldsymbol{\gamma}, \psi, \mathbf{B})P(\mathbf{B}) \\ &= \prod_{i=1}^n P(y_i|\mathbf{a}, \boldsymbol{\gamma}, \psi, \mathbf{B}) \prod_{j=1}^q P(\mathbf{b}_j). \end{aligned} \tag{6}$$

Moreover, $\boldsymbol{\theta}$ and $\mathbf{B}$ are further divided into blocks, each of which is sampled from its full conditional distribution iteratively [Gelfand and Smith, 1990]. The main challenge is how to efficiently sample $\mathbf{b}_j$ and $(\boldsymbol{\gamma}, \boldsymbol{\delta})$ given their high dimensionality. Details of the MCMC algorithm are given in the Appendix. We briefly elaborate the efficiency of our algorithm.

There are two time-consuming steps for sampling $\mathbf{b}_j$. The first factorizes an $n \times n$ covariance matrix, which requires $O(n^3)$ operations. It would be very time-consuming if it was done in each MCMC iteration for all SNP sets. Due to the assumption of $\mathbf{b}_j$ in Equation (2), factorization is only needed in the first iteration. Similar strategies have been used in different settings [Lippert et al., 2011; Zhou et al., 2013]. The second is a matrix multiplication, which transforms a standard multivariate Gaussian vector to Equation (2) and requires $O(n^2)$ operations. Given the low-rank structure of Equation (2) and $r_j << n$, the number of operations is reduced to $O(n)$.

Let $|\boldsymbol{\delta}|$ be the number of nonzero elements in $\boldsymbol{\delta}$. Sampling $(\boldsymbol{\gamma}, \boldsymbol{\delta})$ is implemented by simulating (γ_j, δ_j) iteratively for all j, which requires $O(|\boldsymbol{\delta}|^3)$ operations for matrix inversion and determinant calculation. By using some matrix manipulations including binomial inverse theorem and matrix determinant lemma, we are able to reduce the computational

complexity from $O(|\boldsymbol{\delta}|^3)$ to $O(|\boldsymbol{\delta}|^2)$. Such computational saving is considerable due to the large number of SNP sets and MCMC iterations.

Simulation Study

We evaluated the finite-sample performance of our multiple SNP set association method in a variety of simulation settings. Specifically, we compared our method with the following competing methods: (1) single-SNP association analysis of each SNP [Purcell et al., 2007]; (2) joint association of multiple SNPs through Lasso [Friedman et al., 2010]; (3) joint association of multiple SNP sets through group Lasso [Yang and Zou, 2015]; and (4) marginal SNP-set association through the SKAT method [Wu et al., 2010]. We did not report some other commonly used methods [Chapman and Whittaker, 2008; Mukhopadhyay et al., 2010; Pan, 2011] that perform similarly to SKAT under certain kernel functions. We tried all the six kernel functions provided by SKAT and reported the one with the best performance.

Simulation Scenarios

We used LD blocks defined by the default method [Gabriel, 2002] of Haploview [Barrett et al., 2005] and PLINK [Purcell et al., 2007] to form SNP sets. The SNPs in the adjacent LD blocks may have small or modest correlation. To calculate LD blocks, 1,000 subjects were simulated by randomly combining haplotypes of HapMap CEU subjects. We used PLINK to determine the LD blocks based on these subjects. We randomly selected q blocks, and combined haplotypes of HapMap CEU subjects in each block to form genotype variables for n subjects. We assumed that the causal SNPs are not directly genotyped, and their association with phenotype is measured by the observed SNPs in the same LD block. Specifically, for a causal SNP set j ($j = 1, \ldots, q^*$) with r_j SNPs, r_j^* SNPs were randomly selected to generate y_{ij}. Define the genotypes of these SNPs as u_{ijk}^* ($k = 1, \ldots, r_j^*$). The remaining $r_j - r_j^*$ SNPs were used as u_{ijk} in Equation (2). Moreover, u_{ijk}^* and u_{ijk} were standardized so that their mean and standard deviation equal to 0 and 1, respectively.

We considered different structures of the casual genetic effects with different signal strengths. In case 1, we assumed that the genetic effect of SNPs in an SNP set is additive and homogeneous, such that y_i were generated from the following:

$$y_i = \mathbf{x}_i^T\mathbf{a} + \sum_{j=1}^{q^*}\sum_{k=1}^{r_j^*}\gamma_{jk}^* u_{ijk}^* + \epsilon_i, \tag{7}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2})$, $x_{i1} = 1$, $x_{i2} \sim \text{Uniform}[0, 1]$, $\mathbf{a} = (1, 1)^T$, and $\psi = 3$. Moreover, three scenarios with $\gamma_{jk}^* =$ 0.014, 0.021, and 0.028 were considered. The corresponding average heritability of each set is 0.5%, 1.0%, and 1.8%, respectively. In case 2, the setting is the same as that of case 1 except that we adopted an alternating structure of SNP effects such that $\boldsymbol{\gamma}_j^* = (\gamma_{j1}^*, \ldots, \gamma_{jr_j^*}^*)^T = (c, -c, c, -c, \ldots, c, -c)$, and $c = 0.025$, 0.035, and 0.07. The average heritability of each block is 0.3%, 0.6%, and 2%, respectively. In case 3, we considered a haplotype effects model inspired by Pan [2010]. Let $(u_{ij1}^*, \ldots, u_{ijr_j^*}^*) = (h_{ij11}^*, \ldots, h_{ijr_j^*1}^*) + (h_{ij12}^*, \ldots, h_{ijr_j^*2}^*)$, where $(h_{ij1l}^*, \ldots, h_{ijr_j^*l}^*)$ ($l = 1, 2$) are the two haplotypes of individual i at block j. The phenotype is then generated through

$$y_i = \mathbf{x}_i^T\mathbf{a} + \sum_{j=1}^{q^*}\gamma_j^* f(u_{ij1}^*, \ldots, u_{ijr_j^*}^*) + \epsilon_i, \tag{8}$$

where $f(u_{ij1}^*, \ldots, u_{ijr_j^*}^*)$ equals standardized $(g_{ij1} + g_{ij2})$, and $g_{ijl} = |\sum_{k=1}^{r_j^*} h_{ijkl}^* - r_j^*/2|/r_j^*$, for $l = 1, 2$. We set the average heritability of each block to 0.5%, 1%, and 2% through different γ_j^*, respectively.

Settings with different numbers of q, q^*, r_j^*, and n were also investigated. In setting 1, we used $q = 1,000$, $q^* = 10$, $r_j^* = 20$, and $n = 500$. In setting 2, we set $q = 2,000$, $q^* = 10$, $r_j^* = 10$, and $n = 1,000$. In setting 3, we set $q = 2,000$, $q^* = 10$, $r_j^* = 4$, and $n = 1,000$.

As a result, 27 combinations were tested (3 cases × 3 scenarios × 3 settings). For each combination, 100 datasets were simulated. For each dataset, 5,000 MCMC samples were used as burn-in, and 20,000 samples were acquired to form the empirical posterior distribution. The hyperparameters in the prior distributions were set as follows: $\mathbf{a}_0 = \mathbf{0}$, $\boldsymbol{\Sigma}_{a0} = 10^4\mathbf{I}_d$, $a_{01} = a_{02} = 0.001$, $\pi = 0.005$, $a_{\sigma 01} = 2.1$, and $a_{\sigma 02} = 0.5 * \nu$, where ν is the sample variance of y_i. Additional simulation results with different SNP-set effects and larger numbers of $q(= 10,000)$, $q^*(= 20)$, and $n(= 2,000)$ were reported in the supplementary material.

Methods for Comparison

For comparison, we reanalyzed the datasets with the four competing methods stated above. First, we performed the single-SNP analysis where the association between the phenotype and each SNP is tested by PLINK [Purcell et al., 2007]. Second, we used Lasso to fit

$$y_i = \mathbf{x}_i^T\mathbf{a} + \sum_{j=1}^{q}\sum_{k=1}^{r_j}\beta_{jk}u_{ijk} + \epsilon_i \tag{9}$$

with the glmnet package [Friedman et al., 2010]. For both methods, an SNP- set is regarded as informative, when at least one of the SNPs in the set is informative. Third, we used group Lasso to fit Equation (9) with the gglasso package [Yang and Zou, 2015]. The group structure corresponds to the SNP-sets. Lastly, we applied the SKAT method [Wu et al., 2010] to each SNP- set. We compared the receiver operating characteristic (ROC) curves of our proposed method with the four methods. For SKAT, six different kernels are available in their R package, and we tried all kernels and reported the one with the largest area under curve (AUC) in favor of SKAT.

The ADNI

Imaging genetics evaluates associations between genetic factors and imaging measurements of brain structure and/or function [Medland et al., 2014]. The ADNI is a public-private partnership that combines genetic, structural and functional neuroimaging, and clinical data to measure the progression of mild cognitive impairment (MCI) and early AD. ADNI subjects have been recruited from over 50 sites across the United States and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2 with additional 200 and 650 subjects, respectively. Structural brain MRI data and corresponding clinical and genetic data from baseline and followup were obtained from the ADNI public database (adni.loni.usc.edu, downloaded on May 6, 2009). For up-to-date information, see www.adni-info.org.

We performed GWAS for imaging phenotypes related to AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. Advocates of this approach hypothesize that, compared with traditional case-control GWAS, using imaging measurements may improve the identification of pathogenic genes if imaging phenotypes are closer to the underlying biological etiology of many neurodegenerative and neuropsychiatric diseases (e.g., Alzheimer) [Cannon and Keller, 2006; Chiang et al., 2011; Scharinger et al., 2010].

Subjects from ADNI-1 were used in this study. Briefly, the MRI data were collected using 1.5-T MRI scanners with protocols individualized for each scanner, included standard T1-weighted images obtained using volumetric three-dimensional sagittal MPRAGE or equivalent protocols with varying resolutions. The MRI data were preprocessed by standard steps including anterior commissure and posterior commissure correction, skull-stripping, cerebellum removing, intensity inhomogeneity correction, segmentation, and registration [Shen and Davatzikos, 2004]. Subsequently, automatic regional labeling was performed by labeling the template and transferring the labels following the deformable registration of subject images [Wang et al., 2011]. Ninety-three Regions of Interest (ROIs) were labeled, and the volume of each ROI for each subject was computed. For phenotype variables, we considered nine imaging biomarkers: whole gray matter volume, whole white matter volume, and whole brain volume plus ROIs that are biomarkers for AD (left and right hippocampal volumes, left and right lateral ventricular volumes, and left and right amygdala volumes).

Genotypes for 818 ADNI subjects were generated using Human 610-Quad BeadChips (Illumina, San Diego, CA, USA), where there are 193 AD patients and 397 subjects with MCI. The original SNPs data are based on the human reference sequence build hg18, which were lifted over to hg19 in our analysis. We only considered the 760 Caucasian subjects. Our quality control included call-rate check per subject and per SNP, sex check, relatedness identification, Hardy-Weinberg equilibrium test, SNP minor allele frequency, and ancestry outlier determination. We removed SNPs with (1) more than 5% missing values, (2) minor allele frequency smaller than 5% , and (3) Hardy-Weinberg equilibrium P-value $< 1 \times 10^{-6}$. Remaining missing genotype variables were imputed as the modal value. We removed subjects with (1) outliers in population stratification, (2) sex check failure, and (3) more than 10% missing SNPs. In addition, we removed subjects with invalid volume measurements of ROIs for the ADNI dataset. After quality control, there were 745 Caucasian subjects and 501,666 SNPs left for the analysis.

We also considered two structures of genotype variables. First, LD blocks were calculated through PLINK [Gabriel, 2002; Purcell et al., 2007]. We studied the association between volumes and all LD blocks in each chromosome. Second, we selected SNPs belonging to the top AD candidate genes listed in the AlzGene database (http://www.alzgene.org), and used these genes to form SNP sets. Specifically, we combined the list of the top 40 candidate genes published on June 10, 2010, with the 10 newly updated genes on April 18, 2011. Genes located on chromosome X and those with no genotyped SNPs were removed, resulting in 41 candidate genes for the analysis.

Swedish Schizophrenia Dataset

Schizophrenia is an often devastating neuropsychiatric disorder with considerable morbidity, mortality, and personal and social costs. A large recent GWAS meta-analysis identified 108 common variant associations meeting genome-wide significance [Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014]. We analyzed five subsamples from the Swedish Schizophrenia Study, genotyped with Affymetrix 6.0 (sw2–sw4) and Illumina OmniExpress arrays (sw5–sw6) as described at length in Ripke et al. [2013]. The SNPs data are based on the human reference sequence build hg19. sw5–sw6 contains 2,895 cases and 3,835 controls, whereas sw2–sw4 contains 2,075 cases and 2,341 controls. We applied the quality control procedures used for ADNI to the datasets except that with subjects with more than 5% missing SNPs. For sw2–sw4 and sw5–sw6, 600,745 and 539,883 SNPs remain after quality control, from which 92,771 and 105,670 SNP sets were calculated through the default method [Gabriel, 2002] of Haploview [Barrett et al., 2005]. A total of 2,875 cases and 3,814 controls remain for sw5–sw6, and 2,075 cases and 2,341 controls remain for sw2–sw4.

We studied the association between the case/control status and all LD blocks in each chromosome by extending model (1) with the probit link used in generalized linear model. Gender and the first 10 principal components calculated using EIGENSOFT [Price et al., 2010] are included as covariates. We compared our SNP-set association method with the previous meta-analysis of Ripke et al. [2013]. For comparison, we also used logistic regression with Wald test to study association between case/control status and each SNP. It is of interest to investigate if our method is capable of identifying SNP sets missed by single-SNP and single SNP set

analysis. Also, we studied whether our top-ranked SNP sets were previously reported for schizophrenia and other mental disorders.

Results

Simulation Results

The power of our proposed method to detect SNP sets and SNPs is compared with the other methods in all combinations of simulation settings through ROC curves. Figure 1 shows the ROC curves in case 1 with total counts of false positives vs. true positives under different thresholds in 100 datasets. Group Lasso performs best in combinations with weak and modest heritabilities. For case 1, phenotype is generated from the additive model (7), which resembles model (9). Given that $\boldsymbol{\gamma}_j^*$ is homogeneous in the same direction, and the correlation among SNPs in LD blocks is positive, the L2 norm penalty used by group Lasso [Yang and Zou, 2015] gives a good representation of the signal structure in SNP sets. SKAT with the linear weighted kernel gives similar performance to group Lasso in these combinations. In contrast, our model does not assume additive SNP effects, leading to a slight power loss in the scenarios with weak and modest heritabilities. However, for data with larger heritability and sample sizes, the power loss becomes less severe.

Figure 2 shows the ROC curves in case 2. Our method outperforms group Lasso in most combinations with different heritabilities and dimensions. Moreover, group Lasso performs worse than Lasso and single-SNP analysis in most combinations of this case. Given the alternating structure of SNP effects, SNP signals within each SNP set are largely cancelled out given the strong correlation among SNPs. In this case, the correlation structure of SNPs is an important factor to determine the overall signal strength of each SNP set. Our approach uses SNP information $\mathbf{U}_j$ in the prior specification of $\boldsymbol{\gamma}_j$. In contrast, large absolute SNP effects γ_{jk}^* may lead to small joint effect under certain structure of $\mathbf{U}_j$, making

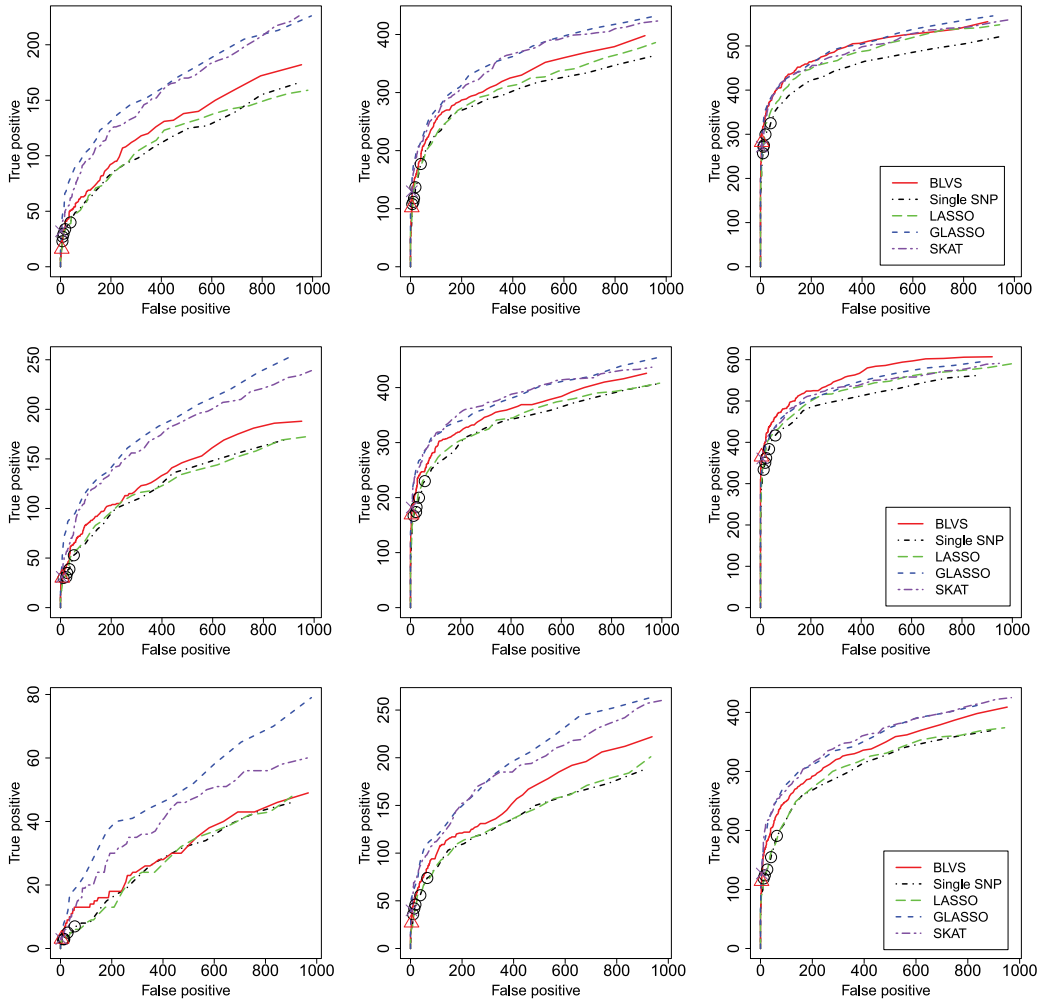

Figure 1. ROC curves in case 1. From left to right are scenarios with different increasing heritability. From top to bottom are settings 1–3 with different dimensions. The red solid, black dotdash, green longdash, blue dashed, and purple twodash lines represent our method, single SNP analysis, Lasso, group Lasso, and SKAT with the best kernel, respectively.

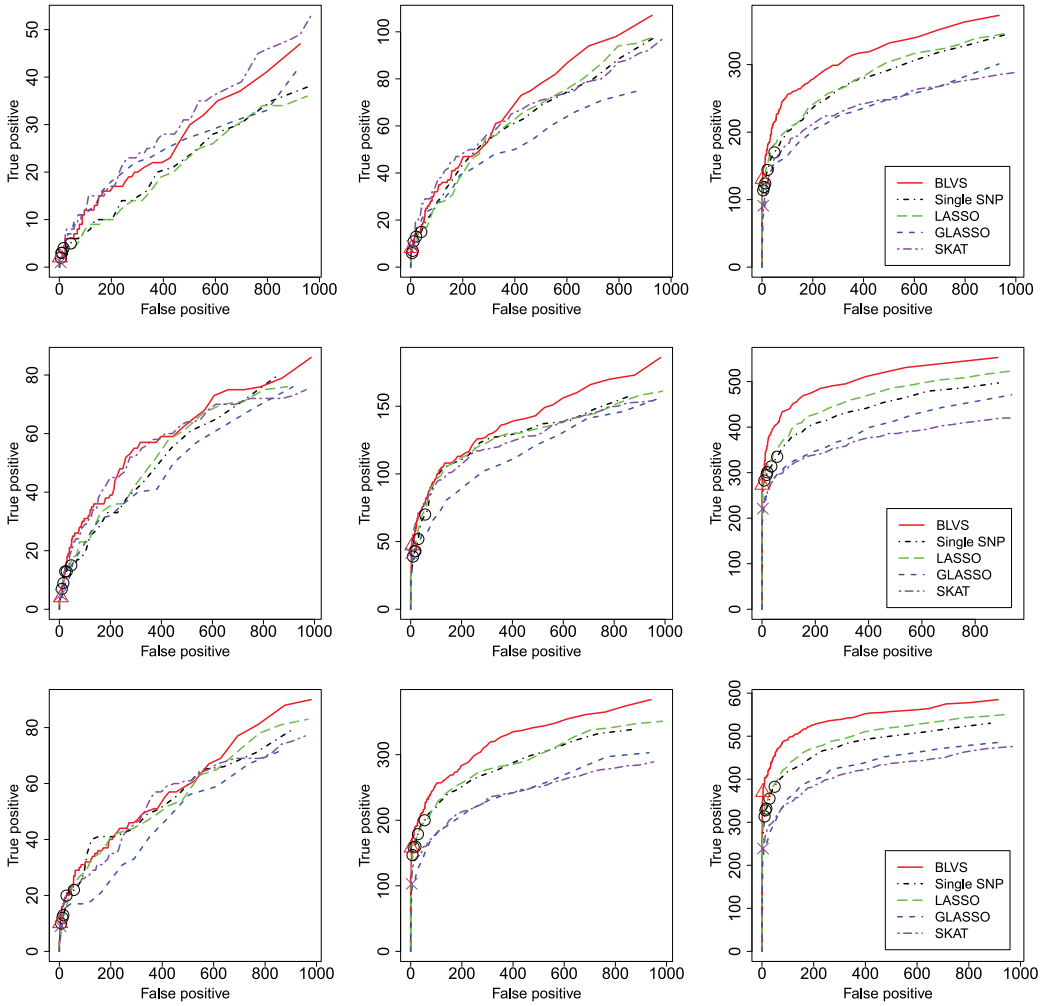

Figure 2. ROC curves in case 2. From left to right are scenarios with different increasing heritability. From top to bottom are settings 1–3 with different dimensions. The red solid, black dotdash, green longdash, blue dashed, and purple twodash lines represent our method, single SNP analysis, Lasso, group Lasso, and SKAT with the best kernel, respectively.

the group Lasso penalty [Yang and Zou, 2015] inefficient in measuring the joint effects of SNP sets. In addition, given that the SNPs in an SNP set are highly correlated, using all SNPs as basis functions to explain the joint effect of SNP set may be inefficient. The abundant parameters cause the fixed effects models (e.g., Lasso and group Lasso) to lose power. In comparison, random effects models (e.g., our approach) use fewer parameters, leading to improved power. In the context of accounting for population structure, Zhang and Pan [2015] provide similar arguments through comparing principal component regression and linear mixed model. Linear weighted kernel is also the best kernel for SKAT in this case. SKAT performs similar to our method when the heritability is small and the number of blocks is small. When the number of blocks increases, the performance of SKAT may decrease, which may be due to the characteristic that SKAT associates each SNP set marginally. The ROC curves in case 3 are depicted in supplementary Figure S2. The results are similar to those in Figure 2.

We tried to investigate the false-positive levels for different methods empirically. In the figures with ROC curves, we have added symbols on the ROC curves to indicate the true- and false-positive counts (TP/FP) corresponding to certain thresholds. The *P*-value threshold of SKAT is 0.05 divided by the number of SNP sets, i.e., Bonferroni corrected. For the single-SNP association, using the total number of SNPs for Bonferroni correction is too conservative because the SNPs in the SNP sets are highly correlated. On the other hand, using the number of SNP sets may be too liberal. We showed TP/FP at five levels, which assumes the number of independent tests equals one to five times of the number of SNP sets. Specifically, the red triangle is the TP/FP of our method with posterior inclusion probability (PIP) >0.5 as threshold. The purple cross is the TP/FP of SKAT with Bonferroni correction of 0.05. The number of tests is the number of SNP sets. The black circles represent TP/FP of marginal single-SNP association at different thresholds. In these figures, the thresholds of our method that use PIP >0.5

produce similar FP compared to the marginal single-SNP association and the SKAT with certain kinds of Bonferroni corrections.

In summary, in the presence of highly correlated SNPs, our method outperforms Lasso and the single-SNP analysis in all combinations. Lasso performs slightly better than the single-SNP analysis. However, group Lasso has unbalanced performance in different cases and settings. Group Lasso performs better in situation where the SNP effects are weak or modest, and the model is correctly specified. In contrast, our model is more robust to model misspecification, e.g., non-additive SNP effects, or the penalty function does not fit the structures of SNP effects within each SNP set. Also, our prior distribution of $\boldsymbol{\gamma}$ accounts for the SNP structure, and thus the proposed method is more flexible to data with different SNP structures.

In terms of computational efficiency, our method takes about 30 min to analyze one dataset from setting 3 with a Linux server (3-GB memory and one core of a Intel X5560 processor); single-SNP association takes 5 min; Lasso (400 tuning parameter values) takes 10 min; group Lasso (400 tuning parameter values) takes 20 min; SKAT takes 120 min.

Although we have formulated our model and simulation studies on quantitative trait, the proposed multiple SNP set method can be extended to case-control studies. Let z_i denote the binary phenotype of the ith subject. A common approach to handle binary response variables in BVS is to introduce a latent variable y_i^* following model (1) and relate z_i and y_i^* with a Probit link: $z_i = 1$ if $y_i^* \geq 0$, $z_i = 0$ if $y_i^* < 0$ [Albert and Chib, 1993; Guan and Stephens, 2011]. The y_i^* is treated as missing data and we include an additional step to sample y_i^* to the MCMC algorithm for quantitative traits. Larger sample sizes are needed for binary phenotypes to achieve similar power of quantitative traits. We conducted a small simulation study for binary data, and results are summarized in supplementary Section 2 and Figure S5. The ROC curves indicate that the new results agree well with the results from the quantitative traits.

ADNI Results

We studied the associations between autosomal LD blocks and nine MRI phenotypes. The Manhattan plots of the inclusion probability of LD blocks for the right hippocampal formation are shown in the left panel of Figure 3. The Manhattan plots for the other ROIs and the total volumes are shown in the supplementary Figure S5.

For comparison, we also calculated the P-values based on single-SNP analysis. The Manhattan plots of $-\log 10(P\text{-values})$ are shown in the right panel of Figure 3 and supplementary Figure S6. We showed the thresholds 5×10^{-8} commonly used in GWAS (the upper one). In addition, we also plotted the thresholds based on Bonferroni correction with the number of SNP sets instead of SNPs (the lower one). Only SNPs near *TOMM40* (adjacent to *APOE*) on chromosome 19 pass the thresholds. In contrast, our SNP set method identifies more informative regions. SNP sets associated with volumes of ROIs and total volumes are reported in Table 1, together with their locations and inclusion probabilities. In each SNP set, the most significant SNP and its P-value of single-SNP analysis are also recorded in Table 1. We also reported the nearest gene of each SNP set within ±400-kb flanking regions. The genes overlapping with the SNP sets are in bold. Regions with rich inclusion of genes are depicted in supplementary Figure S8. The SNP set containing the *APOE* and *TOMM40* on chromosome 19 is among the identified SNP sets. There are other identified SNP sets near genes that have not previously been associated with AD. Names of these genes are shown with italic fonts below.

First, some genes are known to be related to the development of the AD: (1) *GALR1* (chr18, left amygdala) [Stelzer et al., 2011]; (2) *GAS6* (chr13, right amygdala) [Yagami et al.,

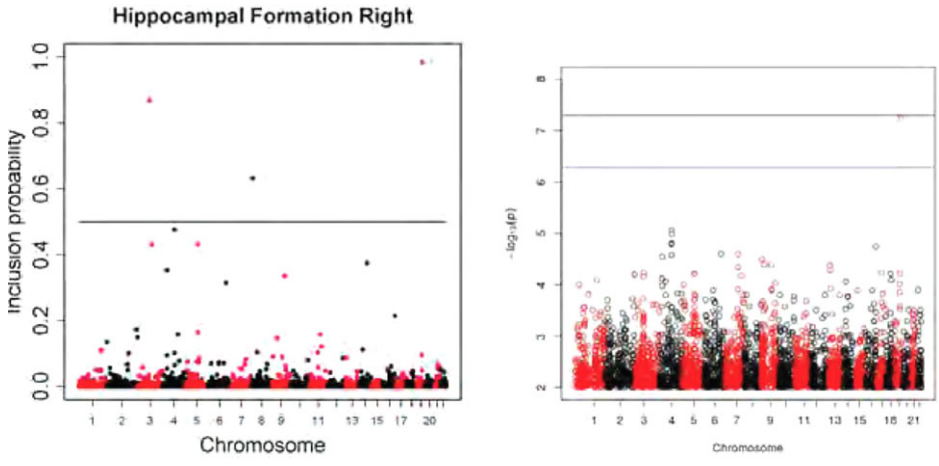

Figure 3. Manhattan plots of inclusion probabilities of all SNP-sets in all autosomes for the Hippocampal Formation Right. The two lines in the right panel are the thresholds corresponding to 0.05 divided by the number of SNPs (upper) and SNP-sets (lower), respectively.

Table 1. SNP sets associated with ROIs and global volumes

ROIs	CHR	Begin BP	End BP	Inc *P*	Best SNP	*P*-value	Gene
LH	19	45395619	45408836	0.990	rs2075650	**2.5E-08**	***TOMM40***
	2	71958480	71975921	0.776	rs4123814	5.6E-05	*DYSF*
RH	19	45395619	45408836	0.966	rs2075650	1.2E-07	***TOMM40***
	8	3494151	3501785	0.642	rs1482203	5.8E-02	***CSMD1***
LA	19	45395619	45408836	0.863	rs2075650	3.0E-06	***TOMM40***
	18	74715269	74722567	0.679	rs470330	6.8E-04	**MBP**, *GALR1*
	18	13516214	13520356	0.547	rs2027683	3.7E-05	***LDLRAD4***
RA	19	45395619	45408836	0.998	rs2075650	**7.0E-09**	***TOMM40***
	5	14301803	14326046	0.684	rs42204	4.0E-05	***TRIO***
	13	114876429	114883790	0.540	rs9805752	3.8E-03	***RASA3***
LL	15	88040269	88048242	0.770	rs2679098	5.5E-05	*NTRK3*
	14	26573361	26590340	0.723	rs12436472	2.5E-06	*NOVA1*
	6	6851538	6858630	0.560	rs9405316	5.8E-06	*LY86*
RL	6	154622914	154627294	0.832	rs1534446	3.4E-07	***PCEF1***
	5	44150391	44221712	0.706	rs4296809	1.2E-06	*FGF10*
	19	11256285	11256887	0.583	rs10402592	3.3E-06	**SPC24**, *LDLR*
	10	132059763	132061197	0.539	rs2480271	6.9E-06	*GLRX3*
	3	1836718	1840129	0.501	rs10510217	3.6E-05	*CNTN4*
WB	13	113838015	113854560	0.911	rs553316	3.8E-05	***PCID2***
	2	226116965	226140372	0.836	rs6728230	2.2E-06	*KIAA1486*
	6	70847385	70868449	0.510	rs3806042	2.9E-05	***COL19A1***
	8	33470899	33471273	0.510	rs7840674	4.4E-05	*DUSP26*

The genes overlapping with the SNP sets and *P*-values that are smaller than 5×10^{-8} are in bold. The coordinates are based on hg19. LH/RH, left/right hippocampal volumes; LA/RA, left/right amygdala volumes; LL/RL, left/right lateral ventricle volumes; WB, whole brain volume.

2002]; (3) *LAMP-1* and *ADPRHL1* (chr13, whole brain volume)[Barrachina et al., 2006; Stelzer et al., 2011]; and (4) *LDLR* (chr19, right lateral ventricular) [Bu, 2009; Kim et al., 2009].

Second, some genes are related to genes associated with AD. It was reported that rare variants in *APP*, *PSEN1*, and *PSEN2* increase risk for AD in late-onset AD families [Cruchaga et al., 2012]. The mutation of *APP* (amyloid beta [A4] precursor protein) protects against AD and age-related cognitive decline [Jonsson et al., 2012]. The following genes are known to interact with *APP* [Stark et al., 2006]: (1) *LY86* (chr6, left lateral ventricle) and (2) *UPF3A* (chr13, right amygdala). Also, *ATP11A* (chr13, whole brain volume) is an interacting gene with *CTNNA3* [Vardarajan, 2013], which is associated with the late-onset AD in females [Miyashita et al., 2007]. In addition, *KANK2* and *SPC24* (chr19, right lateral ventricular) were reported as interacting genes with *PSEN1* and *PSEN2*, respectively [Soler-López et al., 2011].

Third, some genes were reported in other brain dysfunction studies. *CSMD1* (chr8, right hippocampal) is related to schizophrenia [Håvik et al., 2011]. *GAS6* (chr13, right amygdala) is related to cerebrovascular disorders [Allen et al., 1999].

In the SNP set association analysis that used top AD candidate genes as SNP sets, *APOE* is found to be associated with hippocampal formation left and right and amygdala left and right, which agrees with the results in Table 1. All the other candidate genes are not associated with the ROIs or total volumes in the analysis.

Results of Swedish Schizophrenia Dataset

The Manhattan plots of the inclusion probability of SNP sets for sw2–sw4 and sw5–sw6 are shown in the first row of Figure 4. We also calculated the *P*-values based on the single-SNP analysis. The Manhattan plots of $-\log 10$(*P*-values) of sw2–sw4 and sw5–sw6 are shown in the second row of Figure 4. The posterior inclusion probabilities of two SNP sets in sw5–sw6 are greater than 0.5. In comparison, no SNP passes the genome-wide threshold commonly used in GWAS (5×10^{-8}) and the Bonferroni correction with the number of SNP sets. The top SNP sets together with their locations and inclusion probabilities are shown in Tables 2 and 3. We also reported the SNP sets with posterior probabilities larger than

Table 2. Top SNP sets associated with the schizophrenia case/control status based on sw2–sw4

Index	Chr	Start BP	End BP	Inc *P*	Best SNP	–log10*P*	Size	Gene
1	16	4317811	4324293	0.456	rs251740	0.336	2	*TFAP4*
2	10	35267367	35267810	0.419	rs4934692	0.017	3	
3	2	155226677	155246109	0.368	rs799790	0.323	9	*GALNT13*
4	8	139947883	139970267	0.357	rs4736062	0.033	4	
5	4	102549706	102568458	0.34	rs17249850	0.483	6	*BANK1*
6	17	49398988	49441059	0.327	rs11655149	0.006	12	
7	1	241132016	241152090	0.322	rs4659586	0.406	9	*RGS7*
8	9	94822540	94873324	0.309	rs7848874	0.034	6	*SPTLC1*
9	2	221952002	221967358	0.307	rs12694558	0.008	6	
10	13	109907417	109928508	0.297	rs7322967	1.138	6	
11	8	19532200	19543788	0.257	rs17481221	0.807	5	*CSGALNACT1*
12	5	152203199	152310761	0.248	rs7722574	0.016	15	
13	5	151961446	152088546	0.23	rs11750746	0.018	31	
14	5	115515014	115533065	0.214	rs4921076	1.035	7	*COMMD10*
15	3	2357019	2357058	0.214	rs12494110	0.043	2	*CNTN4*
16	6	15477030	15478095	0.214	rs2179168	2.001	2	*JARID2*
17	16	84741149	84741343	0.202	rs3764286	3.403	2	*USP10*

Index shows the ranks of the inclusion probabilities. The chromosome, start BP, end BP, and inclusion probability of each SNP set are listed. The SNP with the smallest *P*-value based on the single-SNP analysis in each SNP set (Best SNP), its $-\log 10$(*P*-value) (–log10*P*) and the genes that overlap with the SNP sets are also reported. The coordinates are hg19.

Table 3. Top SNP sets associated with the schizophrenia case/control status based on sw5–sw6

Index	Chr	Start BP	End BP	Inc *P*	Best SNP	–log10*P*	Size	Gene
1	16	77976316	77983488	0.865	rs436035	0.9	7	*VAT1L*
2	7	2220053	2220092	0.547	rs1637759	0.399	2	*MAD1L1*
3	11	44089161	44091399	0.476	rs178514	1.33	2	*ACCS*
4	4	185137596	185149599	0.461	rs6819977	0.355	3	*ENPP6*
5	10	53317666	53318238	0.443	rs10998097	3.42	2	*PRKG1*
6	9	88999220	89025919	0.425	rs187136	0.935	8	
7	12	72348006	72396996	0.424	rs1843809	0.698	11	*TPH2*
8	5	179805370	179806356	0.407	rs6601116	4.707	2	
9	18	36686072	36786114	0.342	rs4800080	0	14	
10	1	243662027	243662773	0.337	rs9428576	0.874	2	*AKT3* *SDCCAG8*
11	22	19147441	19149580	0.314	rs2096376	0.111	3	
12	7	86415987	86422232	0.288	rs2228595	0.24	4	*GRM3*
13	6	29605935	29648506	0.271	rs9257936	0.065	24	*MOG*, *ZFP57*
14	1	22937818	22952386	0.266	rs186037	3.213	2	
15	3	26622845	26633278	0.236	rs6551116	1.314	5	
16	4	130539982	130555339	0.227	rs10001198	0.468	4	
17	10	13393621	13398428	0.222	rs7923713	2.881	2	
18	20	23528536	23565778	0.221	rs4815220	3.935	6	*CST9L*
19	21	44598827	44602306	0.219	rs2839642	0.088	3	

Index shows the ranks of the inclusion probabilities. The chromosome, start BP, end BP, and inclusion probability of each SNP set are listed. The SNP with the smallest *P*-value based on the single-SNP analysis in each SNP set (Best SNP), its $-\log 10$ (P-value) (–log10P) and the genes that overlap with the SNP sets are also reported. The coordinates are based on hg19.

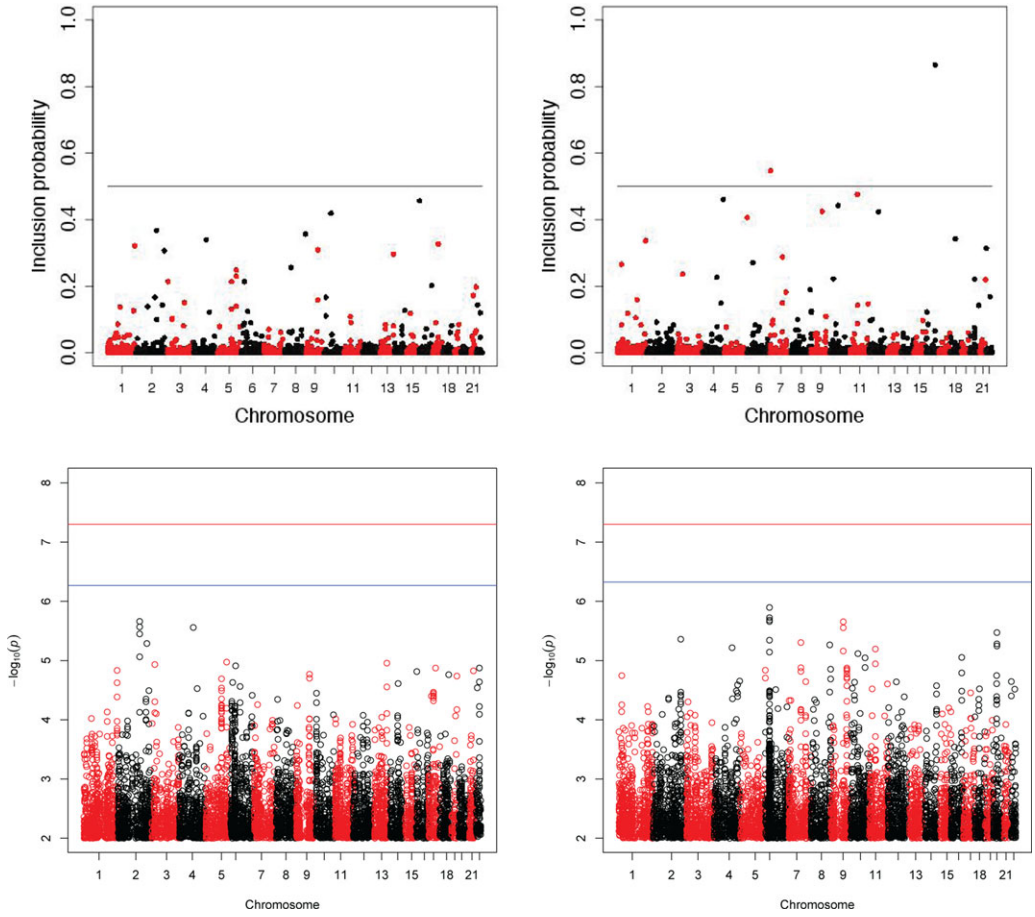

Figure 4. Manhattan plots of inclusion probabilities of all SNP-sets in all autosomes for "sw2-sw4" and "sw5-sw6" are shown in the first row, respectively. Manhattan plots of p-values of SNPs in all autosomes for "sw2-sw4" and "sw5-sw6" are shown in the second row, respectively. The two lines are the thresholds corresponding to 0.05 divided by the number of SNPs (upper) and SNP-sets (lower), respectively.

0.2 to provide more information for future research in case larger samples become available. In each SNP set, the most significant SNP from the single-SNP analysis is also shown in Table 1. We used the GENCODE resource [Harrow et al., 2012] to generate the list of genes with which we overlapped the SNP sets. For each gene, we searched its functionality and product in HGNC [Gray et al., 2013], which are shown in supplementary Tables S1 and S2.

In addition, we compared our results with existing literature. The SNP sets with SNPs recorded in the National Human Genome Research Institute (NHGRI) catalog of published GWAS [Welter et al., 2014] are shown in supplementary Table S3. SNP sets that overlap with genes recorded in the Online Mendelian Inheritance in Man (OMIM) [Hamosh et al., 2005] are listed in supplementary Table S4. We also searched results from previous genome-wide linkage studies and studies of copy-number variations (CNVs). SNP sets containing linkage and CNV regions are shown in supplementary Tables S5 and S6, respectively. Finally, we compared our SNP sets to a CNV morbidity map of developmental delay [Cooper et al., 2011] and listed the SNP sets intersecting the CNV regions in supplementary Table S7.

Two SNP sets in sw5–sw6 result in posterior inclusion probability greater than 0.5. One of them overlaps with the gene *MAD1L1*. This SNP set demonstrates that the BLVS method can identify genetic variants reported in previous studies with much bigger sample sizes that used marginal single-SNP association. *MAD1L1* was reported in the schizophrenia meta-analysis (5.93×10^{-13}) with over 21,000 cases and 38,000 controls where sw2–3 and sw5–sw6 are included as subsamples [Ripke et al., 2013]. In a more recent large-scale meta-analysis of schizophrenia [Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014], *MAD1L1* is the seventh most significant gene among 108 identified genes (8.2×10^{-15}). *MAD1L1* also appeared as a schizophrenia locus (rs10226475, $P = 5.06 \times 10^{-8}$) [Consortium, The Schizophrenia Psychiatric Genome-Wide Association Study (GWAS), 2011]. Biologically, *MAD1L1* is in a human accelerated region that is very different between humans and chimpanzees [Pollard et al., 2006], suggesting it plays an important role in human-specific traits.

In addition, the BLVS method could identify SNP sets with potential interest. The gene *VAT1L* is near the other SNP set with posterior inclusion probability greater than 0.5. *VAT1L*

is seldom reported in GWAS of schizophrenia. However, a region including *VAT1L* was reported in a genome-wide linkage study regarding attention deficit hyperactivity disorder (ADHD) [Zhou et al., 2008]. In addition, according to the Mouse Genome Informatics [Blake et al., 2014], *VAT1L* is related to behavior and neurological phenotypes; as such it may represent an interesting candidate gene for future studies.

The number of overlapping genes discovered in the two independent studies is small, which is likely caused by the low detection power, large number of potential causal SNPs, and discrepancy of the genotyping arrays. The detection power of the two studies is limited by their sample sizes. In addition, the number of causal SNPs/SNP sets is potentially large for Schizophrenia [Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014]. The overlap is less likely to happen when the number of significant genes is small whereas the number of causal genes is large. Moreover, we compared the genotype coverage of the two platforms, i.e., Affymetrix 6.0 (sw2–sw4) and Illumina OmniExpress arrays (sw5–sw6). About 26% SNPs in sw2–sw4 and 22% SNPs in sw5–sw6 are common across the two platforms. Consequently, the SNP sets in sw2–sw4 and sw5–sw6 are quite different.

Discussion

We have developed a linear mixed-effects model and an efficient BLVS procedure for studying the simultaneous association between tens of thousands of SNP sets and a complex trait. Our simulation studies have demonstrated that the multiple SNP set association is more powerful than other regularization methods for high-dimensional data and the marginal kernel method in many cases considered. Using subjects from the ADNI project, we have studied the association between volumes of different ROIs of human brains and SNP sets based on LD blocks and genes. In addition, we analyzed the Swedish schizophrenia dataset with our SNP-set association approach. Causal LD blocks and genes for different traits are identified. Some blocks are well studied in the literature, whereas others reflect new regions associated with AD and schizophrenia.

The current model can be further extended to characterize certain complex features of complex traits and genetic data. First, we analyzed the volumes of ROIs separately. However, volumes of ROIs are usually correlated. Joint modeling the volumes of ROIs may be helpful to achieve better mapping power through borrowing information among correlated traits. Second, some identified SNP sets in the ADNI analysis are associated with other diseases. Incorporating diagnostic status of AD as a secondary phenotype may improve the performance of the association studies. Third, family information may be available for highly heritable diseases such as schizophrenia. Compared with independent samples, family data may have better controlled environment factors, and thus increased power, e.g., for detecting causal SNP/SNP sets. The familial correlations can be modeled by additional random effects with appropriate correlation structures. Finally, the current model does not directly incorporate nonlinear SNP sets' effects or SNPs/SNP sets interaction, which may help to account for the unexplained variation among subjects. The current model uses the linear kernel to form the covariance matrix of the latent effects of SNP sets. Other kernel functions may be considered to address these problems.

The estimate of phenotype variance explained (PVE) is useful for revealing the missing heritability. Guan and Stephens [2011] studied estimation of PVE extensively with BVS. In our model, the PVE can be approximated with the MCMC samples of $\mathbf{a}$, $\boldsymbol{\gamma}$, and $\mathbf{B}$ in model (1) similar to Guan and Stephens [2011]. The estimation of PVE depends on several aspects, e.g., correctly identifying positive and negative genetic variables, estimation of coefficients, and estimation of latent variables. We focus our study on increasing detection power for important SNP sets in the presence of highly correlated SNPs and small effects. Studying PVE with BLVS may be investigated in future studies.

SNP-set methods have been developed in one way to improve the performance of GWAS studies for detecting ungenotyped causal SNPs. In contrast, single-SNP analysis may utilize genotype imputation to boost power for detecting ungenotyped causal SNPs [Marchini and Howie, 2010]. However, SNP-set methods may offer some advantages. Even with the most advanced reference panel [Barrett et al., 2005; Consortium, 2012], some causal SNPs may not be able to be successfully imputed. In addition, SNP-set methods can identify multiple SNPs with mild marginal effects, epistatic effects, nonlinear SNP-effects that can be easily missed by single-SNP methods even with imputed genotypes.

Acknowledgments

This material was based upon work partially supported by the NSF grant DMS-1127914 to the Statistical and Applied Mathematical Science Institute. The research of Dr. Zhu was supported by NSF grants SES-1357666 and DMS-1407655 and NIH grants MH086633, T32MH106440, and 1UL1TR001111. The research of Dr. Knickmeyer was supported by NIMH grant 1R01MH092335. The research of Dr. Zou was supported by NIH grant GM074175. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH and NSF.

The authors have no conflict of interests to declare. Data collection and sharing for this project was funded by the ADNI (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging and National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association, Alzheimer's Drug Discovery Foundation, Araclon Biotech, BioClinica, Biogen Idec, Bristol-Myers Squibb Company, Eisai, Elan Pharmaceuticals, Eli Lilly and Company, EuroImmun, F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Fujirebio, GE Healthcare, IXICO, Janssen Alzheimer Immunotherapy Research & Development, Johnson & Johnson Pharmaceutical Research & Development, Medpace, Merck & Co., Meso Scale Diagnostics, NeuroRx Research, Neurotrack Technologies, Novartis Pharmaceuticals Corporation, Pfizer, Piramal Imaging, Servier, Synarc, and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research provides funds to support ADNI clinical sites in Canada. Private-sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego, California, USA. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of Southern California.

Appendix

The MCMC Algorithm

Let $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$. $\boldsymbol{\gamma}_j^-$ and $\mathbf{B}_j^-$ be the subvector and submatrix of $\boldsymbol{\gamma}$ and $\mathbf{B}$ excluding the j th element and column, respectively. Denote $\mathbf{y}_j^- = \mathbf{y} - \mathbf{X}\mathbf{a} - \mathbf{B}_j^-\boldsymbol{\gamma}_j^-$, $\mathbf{y}_A^- = \mathbf{y} - \mathbf{B}\boldsymbol{\gamma}$, and $\mathbf{y}^- = \mathbf{y} - \mathbf{X}\mathbf{a} - \mathbf{B}\boldsymbol{\gamma}$.

1. Sample $\mathbf{a}$ from $N((\mathbf{X}^T\mathbf{X}/\psi + {\boldsymbol{\Sigma}_{a0}}^{-1})^{-1}\mathbf{X}^T\mathbf{y}_A^-, (\mathbf{X}^T\mathbf{X}/\psi + {\boldsymbol{\Sigma}_{a0}}^{-1})^{-1})$.
2. Sample ψ from IG($a_{01} + n/2$, $a_{02} + \mathbf{y}^{-T}\mathbf{y}^-/2$).
3. Sample $(\boldsymbol{\gamma}, \boldsymbol{\delta})$ from $P(\boldsymbol{\gamma}, \boldsymbol{\delta}|\mathbf{y}, \mathbf{a}, \mathbf{B}, \psi, \boldsymbol{\sigma}^2)$.
4. Sample σ_j^2 from IG($a_{\sigma 01} + 1/2$, $a_{\sigma 02} + \gamma_j^2/2$) if $\delta_j = 1$; or IG($a_{\sigma 01}$, $a_{\sigma 02}$) if $\delta_j = 0$.
5. Sample $\mathbf{b}_j$ from $N_n(\mathbf{b}_j^*, \boldsymbol{\Sigma}_j^*)$ for $j = 1, \ldots, q$ sequentially, where $\mathbf{b}_j^* = (\gamma_j/\psi)\boldsymbol{\Sigma}_j^*\mathbf{y}_j^-$, and $\boldsymbol{\Sigma}_j^*$ is derived in the next section.

Efficient Sampling of $\mathbf{b}_j$

Given that $\boldsymbol{\Sigma}_j$ is fixed and known, and $rank(\boldsymbol{\Sigma}_j) << n$, generating $\mathbf{b}_j$ can be very efficient because of the save of matrix factorization of $\boldsymbol{\Sigma}_j^*$ in each iteration. More specifically, let $r_j = rank(\boldsymbol{\Sigma}_j)$ and $\boldsymbol{\Sigma}_j = \mathbf{Q}_j\mathbf{V}_j\mathbf{Q}_j^T$, where $\mathbf{Q}_j$ is a $n \times r_j$ orthonormal matrix, $\mathbf{V}_j = diag(v_{j1}, \ldots, v_{jr_j})$ is a diagonal matrix , and $v_{jk} > 0$ for $k = 1, \ldots, r_j$. Consequently, $\boldsymbol{\Sigma}_j^* = \mathbf{Q}_j\mathbf{V}_j^*\mathbf{Q}_j^T$, where $\mathbf{V}_j^* = diag(v_1^*, \ldots, v_{r_j}^*)$, and $v_k^* = 1/((\gamma_j^2/\psi) + 1/v_{jk})$, for $k = 1, \ldots, r_j$. In each MCMC iteration, $\mathbf{b}_j$ can be generated as follows:

1. Generate $c_{jk} \sim N_1[0, v_k^*]$ for $k = 1, \ldots, r_j$.
2. Let $\mathbf{b}_j = \mathbf{Q}_j\mathbf{c}_j$, where $\mathbf{c}_j = (c_{j1}, \ldots, c_{jr_j})$.

The computation complexity of Step 5 is reduced from $O(n^3)$ to $O(nr_j)$. Since $r_j << n$, the computation is approximately O(n).

Efficient Sampling Algorithm for $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$

We sample $(\boldsymbol{\gamma}, \boldsymbol{\delta})$ from $P(\boldsymbol{\gamma}, \boldsymbol{\delta}|\mathbf{y}, \mathbf{a}, \mathbf{B}, \psi, \boldsymbol{\sigma}^2)$ through the following:

1. Sample $\boldsymbol{\delta}$ from $P(\boldsymbol{\delta}|\mathbf{y}, \mathbf{a}, \mathbf{B}, \psi, \boldsymbol{\sigma}^2)$.
2. Sample $\boldsymbol{\gamma}$ from $P(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{a}, \mathbf{B}, \psi, \boldsymbol{\delta}, \boldsymbol{\sigma}^2)$.

Let $\mathbf{B}_\delta$ and $\boldsymbol{\gamma}_\delta$ containing columns of $\mathbf{B}$ and elements of $\boldsymbol{\gamma}$ corresponding to $\delta_j = 1$, for $j = 1, \ldots, q$, respectively. Let $\boldsymbol{\Sigma}_{\delta 0}$ be the covariance matrix of prior distribution of $\boldsymbol{\gamma}_\delta$, $|\boldsymbol{\delta}|$ be the number of nonzero elements in $\boldsymbol{\delta}$, and $\boldsymbol{\gamma}_{-\delta}$ be the elements of $\boldsymbol{\gamma}$ corresponding to $\delta_j = 0$.

$$
\begin{aligned}
& P(\boldsymbol{\gamma}, \boldsymbol{\delta}|\mathbf{y}, \mathbf{a}, \mathbf{B}, \psi, \boldsymbol{\sigma}^2) \\
\propto\ & P(\mathbf{y}|\mathbf{a}, \boldsymbol{\gamma}, \psi, \mathbf{B})P(\boldsymbol{\gamma}, \boldsymbol{\delta}|\boldsymbol{\sigma}^2) \\
\propto\ & \exp\left\{-\frac{1}{2\psi}\left[{\boldsymbol{\gamma}_\delta}^T{\mathbf{B}_\delta}^T\mathbf{B}_\delta\boldsymbol{\gamma}_\delta - 2{\boldsymbol{\gamma}_\delta}^T{\mathbf{B}_\delta}^T\mathbf{y}_B^-\right]\right\}\sqrt{\frac{|\boldsymbol{\Sigma}_{\delta 0}^{-1}|}{(2\pi)^{|\boldsymbol{\delta}|}}} \\
& \exp\left\{-\frac{1}{2}{\boldsymbol{\gamma}_\delta}^T\boldsymbol{\Sigma}_{\delta 0}^{-1}\boldsymbol{\gamma}_\delta\right\}\prod_{j=1}^{q}\pi^{\delta_j}(1-\pi)^{(1-\delta_j)},
\end{aligned}
$$

where $\mathbf{y}_B^- = \mathbf{y} - \mathbf{X}\mathbf{a}$. Marginalizing $\boldsymbol{\gamma}_\delta$ out leads to

$$
P(\boldsymbol{\delta}|\mathbf{y}, \mathbf{a}, \mathbf{B}, \psi, \boldsymbol{\sigma}^2) \propto \frac{|\boldsymbol{\Sigma}_\delta|^{1/2}}{|\boldsymbol{\Sigma}_{\delta 0}|^{1/2}}\exp\left\{\frac{1}{2}\boldsymbol{\gamma}_\delta^{*T}\boldsymbol{\Sigma}_\delta^{-1}\boldsymbol{\gamma}_\delta^*\right\}
\prod_{j=1}^{q}\pi^{\delta_j}(1-\pi)^{(1-\delta_j)},
$$

where

$$
\boldsymbol{\Sigma}_\delta = \left(\psi^{-1}{\mathbf{B}_\delta}^T\mathbf{B}_\delta + \boldsymbol{\Sigma}_{\delta 0}^{-1}\right)^{-1}, \quad \boldsymbol{\gamma}_\delta^* = \psi^{-1}\boldsymbol{\Sigma}_\delta{\mathbf{B}_\delta}^T\mathbf{y}_B^-.
$$

Denote $\boldsymbol{\delta}_{-j}$ be the subvector of $\boldsymbol{\delta}$ excluding δ_j. We update $\boldsymbol{\delta}$ by conditionally sampling δ_j from $P(\delta_j|\mathbf{y}, \mathbf{a}, \psi, \boldsymbol{\delta}_{-j}, \mathbf{B}, \boldsymbol{\sigma}^2)$ for $j = 1, \ldots, q$. For a certain $k \in [1, q]$, without loss of generality, we assume that current state $\delta_k = 0$. Let $\eta_j = \delta_j$ for $j \neq k$, $\eta_k = 1$, and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_q)$. Let $\mathbf{B}_\eta$ and $\boldsymbol{\gamma}_\eta$ containing columns of $\mathbf{B}$ and elements of $\boldsymbol{\gamma}_\eta$ corresponding to $\eta_j = 1$, and $\boldsymbol{\Sigma}_{\eta 0}$ be the covariance matrix of prior distribution of $\boldsymbol{\gamma}_\eta$. R is used to calculate $P(\delta_k = 1|\mathbf{y}, \mathbf{a}, \psi, \boldsymbol{\delta}_{-k}, \mathbf{B}, \boldsymbol{\sigma}^2) = R/(1+R)$:

$$
\begin{aligned}
R &= \frac{P(\delta_k = 1|\mathbf{y}, \mathbf{a}, \psi, \boldsymbol{\delta}_{-j}, \mathbf{B}, \boldsymbol{\sigma}^2)}{P(\delta_k = 0|\mathbf{y}, \mathbf{a}, \psi, \boldsymbol{\delta}_{-j}, \mathbf{B}, \boldsymbol{\sigma}^2)} = \frac{P(\boldsymbol{\eta}|\mathbf{y}, \mathbf{a}, \psi, \mathbf{B}, \boldsymbol{\sigma}^2)}{P(\boldsymbol{\delta}|\mathbf{y}, \mathbf{a}, \psi, \mathbf{B}, \boldsymbol{\sigma}^2)} \\
&= \frac{|\boldsymbol{\Sigma}_\eta|^{1/2}|\boldsymbol{\Sigma}_{\delta 0}|^{1/2}\exp\left\{\frac{1}{2}\boldsymbol{\gamma}_\eta^*\boldsymbol{\Sigma}_\eta^{-1}\boldsymbol{\gamma}_\eta^{*T}\right\}}{|\boldsymbol{\Sigma}_\delta|^{1/2}|\boldsymbol{\Sigma}_{\eta 0}|^{1/2}\exp\left\{\frac{1}{2}\boldsymbol{\gamma}_\delta^*\boldsymbol{\Sigma}_\delta^{-1}\boldsymbol{\gamma}_\delta^{*T}\right\}}\frac{\pi}{(1-\pi)}, \qquad \text{(A1)}
\end{aligned}
$$

where

$$
\boldsymbol{\Sigma}_\eta = (\psi^{-1}{\mathbf{B}_\eta}^T\mathbf{B}_\eta + \boldsymbol{\Sigma}_{\eta 0}^{-1})^{-1}, \quad \boldsymbol{\gamma}_\eta^* = \psi^{-1}\boldsymbol{\Sigma}_\eta{\mathbf{B}_\eta}^T\mathbf{y}_B^-.
$$

Reorder the number of columns in $\mathbf{B}_\eta$ such that $\mathbf{B}_\eta = (\mathbf{B}_\delta, \tilde{\mathbf{b}}_k)$,

$$
\boldsymbol{\Sigma}_\eta^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_\delta^{-1} & \psi^{-1}{\mathbf{B}_\delta}^T\tilde{\mathbf{b}}_k \\ \psi^{-1}\tilde{\mathbf{b}}_k^T\mathbf{B}_\delta & \psi^{-1}\tilde{\mathbf{b}}_k^T\tilde{\mathbf{b}}_k + 1/\sigma_k^2 \end{pmatrix} \equiv \begin{pmatrix} \mathbf{M}_\delta & \mathbf{v}_\delta \\ \mathbf{v}_\delta^T & m_k \end{pmatrix}, \text{ and}
$$

$$
\boldsymbol{\Sigma}_\eta = \begin{pmatrix} \mathbf{M}_\delta & \mathbf{v}_\delta \\ \mathbf{v}_\delta^T & m_k \end{pmatrix}^{-1} = \begin{pmatrix} \tilde{\mathbf{M}}_\delta^{-1} & -\tilde{\mathbf{M}}_\delta^{-1}\mathbf{v}_\delta m_k^{-1} \\ -m_k^{-1}\mathbf{v}_\delta^T\tilde{\mathbf{M}}_\delta^{-1} & m_k^{-1} + m_k^{-2}\mathbf{v}_\delta^T\tilde{\mathbf{M}}_\delta^{-1}\mathbf{v}_\delta \end{pmatrix},
$$

where $\tilde{\mathbf{M}}_\delta = \mathbf{M}_\delta - \mathbf{v}_\delta m_k^{-1}\mathbf{v}_\delta^T$.

It can be shown that $|\boldsymbol{\Sigma}_\eta| = (|\tilde{\mathbf{M}}_\delta||m_k|)^{-1}$. From matrix determinant lemma, $|\tilde{\mathbf{M}}_\delta| = |\mathbf{M}_\delta|(1 - \mathbf{v}_\delta^T\boldsymbol{\Sigma}_\delta\mathbf{v}_\delta m_k^{-1})$. Thus, $|\boldsymbol{\Sigma}_\eta|/|\boldsymbol{\Sigma}_\delta| = (m_k - \mathbf{v}_\delta^T\boldsymbol{\Sigma}_\delta\mathbf{v}_\delta)^{-1}$. Denote $\mathbf{u}_\delta = \psi^{-1}{\mathbf{B}_\delta}^T\mathbf{y}_B^-$ and $s_k = \psi^{-1}\tilde{\mathbf{b}}_k^T\mathbf{y}_B^-$.

$$
\begin{aligned}
\boldsymbol{\gamma}_\eta^{*T}\boldsymbol{\Sigma}_\eta^{-1}\boldsymbol{\gamma}_\eta^* &= \mathbf{u}_\delta^T\tilde{\mathbf{M}}_\delta^{-1}\mathbf{u}_\delta - 2\mathbf{u}_\delta^T\tilde{\mathbf{M}}_\delta^{-1}\mathbf{v}_\delta m_k^{-1}s_k \\
&\quad + s_k^2\left(m_k^{-1} + m_k^{-2}\mathbf{v}_\delta^T\tilde{\mathbf{M}}_\delta^{-1}\mathbf{v}_\delta\right) \\
&= \boldsymbol{\gamma}_\delta^{*T}\boldsymbol{\Sigma}_\delta^{-1}\boldsymbol{\gamma}_\delta^* + \frac{\mathbf{u}_\delta^T\boldsymbol{\Sigma}_\delta\mathbf{v}_\delta\mathbf{v}_\delta^T\boldsymbol{\Sigma}_\delta\mathbf{u}_\delta}{m_k - \mathbf{v}_\delta^T\boldsymbol{\Sigma}_\delta\mathbf{v}_\delta} \\
&\quad - 2\left(\mathbf{u}_\delta^T\boldsymbol{\Sigma}_\delta\mathbf{v}_\delta^T + \frac{\mathbf{u}_\delta^T\boldsymbol{\Sigma}_\delta\mathbf{v}_\delta\mathbf{v}_\delta^T\boldsymbol{\Sigma}_\delta\mathbf{v}_\delta^T}{m_k - \mathbf{v}_\delta^T\boldsymbol{\Sigma}_\delta\mathbf{v}_\delta}\right)m_k^{-1}s_k \\
&\quad + m_k^{-1}s_k^2 + m_k^{-2}s_k^2\left(\mathbf{v}_\delta^T\boldsymbol{\Sigma}_\delta\mathbf{v}_\delta + \frac{\mathbf{v}_\delta^T\boldsymbol{\Sigma}_\delta\mathbf{v}_\delta\mathbf{v}_\delta^T\boldsymbol{\Sigma}_\delta\mathbf{v}_\delta}{m_k - \mathbf{v}_\delta^T\boldsymbol{\Sigma}_\delta\mathbf{v}_\delta}\right),
\end{aligned}
$$

where the second equation follows the binomial inverse theorem

$$
\tilde{\mathbf{M}}_\delta^{-1} = \mathbf{M}_\delta^{-1} + \mathbf{M}_\delta^{-1}\mathbf{v}_\delta\mathbf{v}_\delta^T\mathbf{M}_\delta^{-1}(m_k - \mathbf{v}_\delta^T\mathbf{M}_\delta^{-1}\mathbf{v}_\delta)^{-1}.
$$

Let $c_1 = \mathbf{u}_\delta^T \boldsymbol{\Sigma}_\delta \mathbf{v}_\delta$, $c_2 = \mathbf{v}_\delta^T \boldsymbol{\Sigma}_\delta \mathbf{v}_\delta$, and $c_3 = m_k - \mathbf{v}_\delta^T \boldsymbol{\Sigma}_\delta \mathbf{v}_\delta$. Equation (A1) can be simplified as

$$R = \frac{1}{(\sigma_k^2 c_3)^{1/2}} \exp\left\{\frac{1}{2}\left(\frac{c_1^2}{c_3} - 2\left(\frac{c_1 s_k}{m_k} - \frac{c_1 c_2 s_k}{m_k c_3}\right) + \frac{s_k^2}{m_k} + \frac{s_k^2(c_2 + c_2^2/c_3)}{m_k^2}\right)\right\} \frac{\pi}{(1-\pi)}. \quad \text{(A2)}$$

As $\boldsymbol{\Sigma}_\delta$ is known given that $\boldsymbol{\delta}$ is the current state, the computation complexity is reduced from $O(|\boldsymbol{\delta}|^3)$ in (A1) to $O(|\boldsymbol{\delta}|^2)$ in Equation (A2). After we get a new sample of $\boldsymbol{\delta}$, we sample $\boldsymbol{\gamma}_\delta \sim N(\boldsymbol{\gamma}_\delta^*, \boldsymbol{\Sigma}_\delta)$, and set $\boldsymbol{\gamma}_{-\delta} = \mathbf{0}$.

When $\delta_k = 1$, $\boldsymbol{\eta}$ is the current state and $\boldsymbol{\Sigma}_\eta$ is known, $\boldsymbol{\Sigma}_\delta$ can be calculated from the binomial inverse theorem,

$$\begin{aligned}\boldsymbol{\Sigma}_\delta &= \left(\tilde{\mathbf{M}}_\delta + \mathbf{v}_\delta m_k^{-1} \mathbf{v}_\delta^T\right)^{-1} \\ &= \tilde{\mathbf{M}}_\delta^{-1} - \left(-\tilde{\mathbf{M}}_\delta^{-1}\mathbf{v}_\delta m_k^{-1}\right)(m_k^{-1} + m_k^{-2}\mathbf{v}_\delta^T \tilde{\mathbf{M}}_\delta^{-1}\mathbf{v}_\delta)^{-1} \\ &\quad \left(-m_k^{-1}\mathbf{v}_\delta^T \tilde{\mathbf{M}}_\delta^{-1}\right),\end{aligned}$$

where $\tilde{\mathbf{M}}_\delta^{-1}$, $(-\tilde{\mathbf{M}}_\delta^{-1}\mathbf{v}_\delta m_k^{-1})$, and $(m_k^{-1} + m_k^{-2}\mathbf{v}_\delta^T \tilde{\mathbf{M}}_\delta^{-1}\mathbf{v}_\delta)$ are submatrices of $\boldsymbol{\Sigma}_\eta$. And Equation (A2) can be computed accordingly.

References

Abi-Dargham A. 2007. *Alterations of Serotonin Transmission in Schizophrenia.* In: AbiDargham A, Guillin O, editors. *Integrating the Neurobiology of Schizophrenia.* Volume 78, International Review of Neurobiology. Academic Press, San Diego, California, pp. 133–164.

Albert JH, Chib S. 1993. Bayesian analysis of binary and polychotomous response data. *J Am Stat Assoc* 88(422):669–679.

Allen MP, Zeng C, Schneider K, Xiong X, Meintzer MK, Bellosta P, Basilico C, Varnum B, Heidenreich KA, Wierman ME. 1999. Growth arrest-specific gene 6 (Gas6)/adhesion related kinase (Ark) signaling promotes gonadotropin-releasing hormone neuronal survival via extracellular signal-regulated kinase (ERK) and Akt. *Mol Endocrinol* 13(2):191–201.

Ballard DH, Cho J, Zhao H. 2010. Comparisons of multi-marker association methods to detect association between a candidate region and disease. *Genet Epidemiol* 34(3):201–212.

Barbieri MM, Berger JO. 2004. Optimal predictive model selection. *Ann Stat* 32(3):870–897.

Barrachina M, Maes T, Buesa C, Ferrer I. 2006. Lysosome-associated membrane protein 1 (LAMP-1) in Alzheimer's disease. *Neuropathol Appl Neurobiol* 32(5): 505–516.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21(2):263–265.

Bhadra A, Mallick BK. 2013. Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics* 69(2):447–457.

Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE; The Mouse Genome Database Group. 2014. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucl Acids Res* 42(D1):D810–D817.

Bu G. 2009. Apolipoprotein E and its receptors in Alzheimer's disease: pathways, pathogenesis and therapy. *Nat Rev Neurosci* 10(5):333–344.

Cannon TD, Keller M. 2006. Endophenotypes in the genetic analyses of mental disorders. *Annu Rev Clin Psychol* 40: 267–290.

Chapman J, Whittaker J. 2008. Analysis of multiple SNPs in a candidate gene or region. *Genet Epidemiol* 32(6):560–566.

Chen Z, Dunson DB. 2003. Random effects selection in linear mixed models. *Biometrics* 59(4):762–769.

Chiang MC, Barysheva M, Toga AW, Medland SE, Hansell NK, James MR, McMahon KL, de Zubicaray GI, Martin NG, Wright MJ and others,. 2011. BDNF gene effects on brain circuitry replicated in 455 twins. *NeuroImage* 55: 448–454.

Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V and others. 2011. A copy number variation morbidity map of developmental delay. *Nat Genet* 43(9):838–846.

Cruchaga C, Chakraverty S, Mayo K, Vallania FLM, Mitra RD, Faber K, Williamson J, Bird T, Diaz-Arrastia R, Foroud TM and others; NIA-LOAD/NCRAD Family Study Consortium. 2012. Rare variants in *APP*, *PSEN1* and *PSEN2* increase risk for AD in late-onset Alzheimer's disease families. *PLoS One* 7(2):e31039.

Egan MF, Straub RE, Goldberg TE, Yakub I, Callicott JH, Hariri AR, Mattay VS, Bertolino A, Hyde TM, Shannon-Weickert C and others. 2004. Variation in GRM3 affects cognition, prefrontal glutamate, and risk for schizophrenia. *Proc Natl Acad Sci USA* 101(34):12604–12609.

Fridley BL. 2009. Bayesian variable and model selection methods for genetic association studies. *Genet Epidemiol* 33(1):27–37.

Fridley BL, Biernacka JM. 2011. Gene set analysis of SNP data: benefits, challenges, and future directions. *Eur J Hum Genet* 19(8):837–843.

Fridley BL, Jenkins GD, Biernacka JM. 2010. Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. *PLoS One* 5(9):e12693.

Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 33(1):1–22.

Gabriel SB. 2002. The structure of haplotype blocks in the human genome. *Science* 296(5576):2225–2229.

Gauderman WJ, Murcray C, Gilliland F, Conti DV. 2007. Testing association between disease and multiple SNPs in a candidate gene. *Genet Epidemiol* 31(5):383–395.

Gelfand AE, Smith AF. 1990. Sampling-based approaches to calculating marginal densities. *J Am Stat Assoc* 85(410):398–409.

George EI, McCulloch RE. 1993. Variable selection via Gibbs sampling. *J Am Stat Assoc* 88(423):881–889.

Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, Bruford EA. 2013. Genenames.org: the HGNC resources in 2013. *Nucl Acids Res* 41(D1):D545–D552.

Guan Y, Stephens M. 2011. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann Appl Stat* 5(3):1780–1815.

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledge base of human genes and genetic disorders. *Nucl Acids Res* 33(suppl 1):D514–D517.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S and others. 2012. GENCODE: the reference human genome annotation for the ENCODE Project. *Genome Res* 22(9):1760–1774.

Håvik B, Le Hellard S, Rietschel M, Lybæk H, Djurovic S, Mattheisen M, Mühleisen TW, Degenhardt F, Priebe L, Maier W and others. 2011. The complement control-related genes CSMD1 and CSMD2 associate to schizophrenia. *Biol Psychiatry*, 70(1):35–42.

He Q, Lin D-Y. 2011. A variable selection method for genome-wide association studies. *Bioinformatics* 27(1):1–8.

Hoggart CJ, Whittaker JC, de Iorio M, Balding DJ. 2008. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet* 4(7):e1000130.

Jonsson T, Atwal JK, Steinberg S, Snaedal J, Jonsson PV, Bjornsson S, Stefansson H, Sulem P, Gudbjartsson D, Maloney J and others. 2012. A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature* 488(7409):96–99.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ , Eskin E. 2008. Efficient control of population structure in model organism association mapping. *Genetics* 178(3):1709–1723.

Kim J, Basak JM, Holtzman DM. 2009. The role of apolipoprotein E in Alzheimer's disease. *Neuron* 63(3):287–303.

Liang F, Song Q, Yu K. 2013. Bayesian subset modeling for high dimensional generalized linear models. *J Am Stat Assoc* 108(502):589–606.

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. 2011. FaST linear mixed models for genome-wide association studies. *Nat Methods* 8(10): 833–835.

Logsdon BA, Carty CL, Reiner AP, Dai JY, Kooperberg C. 2012. A novel variational Bayes multiple locus Z-statistic for genome-wide association studies with Bayesian model averaging. *Bioinformatics* 28(13):1738–1744.

Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11(7):499–511.

Medland SE, Jahanshad N, Neale BM, Thompson PM. 2014. Whole-genome analyses of whole-brain data: working within an expanded search space. *Nat Neurosci* 17(6):791–800.

Miyashita A, Arai H, Asada T, Imagawa M, Matsubara E, Shoji M, Higuchi S, Urakami K, Kakita A, Takahashi H and others. 2007. Genetic association of CTNNA3 with late-onset Alzheimer's disease in females. *Hum Mol Genet* 16(23):2854–2869.

Mukhopadhyay I, Feingold E, Weeks DE, Thalamuthu A. 2010. Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genet Epidemiol* 34(3):213–221.

O'Brien NL, Way MJ, Kandaswamy R, Fiorentino A, Sharp SI, Quadri G, Alex J, Anjorin A, Ball D, Cherian R and others. 2014. The functional GRM3 Kozak sequence variant rs148754219 affects the risk of schizophrenia and alcohol dependence as well as bipolar disorder. *Psychiatr Genet* 24(6):277–278.

O'Hara RB, Sillanpää MJ. 2009. A review of Bayesian variable selection methods: what, how and which. *Bayesian Anal* 4(1):85–117.

Otto EA, Hurd TW, Airik R, Chaki M, Zhou W, Stoetzel C, Patil S. B., Levy S, Ghosh AK, Murga-Zamalloa CA and others. 2010. Candidate exome capture identifies mutation of SDCCAG8 as the cause of a retinal-renal ciliopathy. *Nat Genet* 42(10):840–850.

Pan W. 2010. A unified framework for detecting genetic association with multiple SNPs in a candidate gene or region: contrasting genotype scores and LD patterns between cases and controls. *Hum Hered* 69(1):1–13.

Pan W. 2011. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genet Epidemiol* 35(4):211–216.

Park T, Casella G. 2008. The Bayesian lasso. *J Am Stat Assoc* 103(482):681–686.

Patil H, Tserentsoodol N, Saha A, Hao Y, Webb M, Ferreira PA. 2012. Selective loss of RPGRIP1-dependent ciliary targeting of NPHP4, RPGR and SDCCAG8 underlies the degeneration of photoreceptor neurons. *Cell Death Dis* 3(7):e355.

Poduri A, Evrony GD, Cai X, Elhosary PC, Beroukhim R, Lehtinen MK, Hills LB, Heinzen EL, Hill A, Hill RS and others. 2012. Somatic activation of AKT3 causes hemispheric developmental brain malformations. *Neuron* 74(1):41–48.

Pollard KS, Salama SR, Lambert N, Lambot M-A, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A and others. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443(7108): 167–172.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38(8):904–909.

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei L-J, Sunyaev SR. 2010. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 86(6):832–838.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC . 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559–575.

Ripke S, O'Dushlaine C, Chambert K, Moran JL, Khler AK, Akterin S, Bergen SE, Collins AL, Crowley JJ, Fromer M and others. 2013. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* 45(10):1150–1159.

Rivière J-B, Mirzaa GM, O'Roak BJ, Beddaoui M, Alcantara D, Conway RL, St-Onge J, Schwartzentruber JA, Gripp KW, Nikkel SM and others; Finding of Rare Disease Genes (FORGE) Canada Consortium. 2012. De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. *Nat Genet* 44(8):934–940.

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am JHum Genet* 70(2):425–434.

Scharinger C, Rabl U, Sitte HH, Pezawas L. 2010. Imaging genetics of mood disorders. *NeuroImage* 53: 810–821.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511(7510):421–427.

Shen DG, Davatzikos C. 2004. Measuring temporal morphological changes robustly in brain MR images via 4-dimensional template warping. *NeuroImage* 21: 1508–1517.

Skarman A, Shariati M, Jans L, Jiang L, Sørensen P. 2012. A Bayesian variable selection procedure to rank overlapping gene sets. *BMC Bioinformatics* 13(1):73.

Soler-López M, Zanzoni A, Lluís R, Stelzl U, Aloy P. 2011. Interactome mapping suggests new mechanistic details underlying Alzheimer's disease. *Genome Res* 21(3):364–376.

Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. 2006. BioGRID: a general repository for interaction datasets. *Nucl Acids Res* 34(suppl 1):D535–D539.

Stelzer G, Dalah I, Stein TI, Satanower Y, Rosen N, Nativ N, Oz-Levi D, Olender T, Belinky F, Bahir I and others. 2011. In-silico human genomics with GeneCards. *Hum Genomics* 5(6):709–717.

TGP Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56–65.

The Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. 2011. Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* 43(10):969–976.

Tzeng J-Y, Zhang D. 2007. Haplotype-based association analysis via variance-components score test. *Am J Hum Genet* 81(5):927–938.

Tzeng J-Y, Devlin B, Wasserman L, Roeder K. 2003. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am J Hum Genet* 72(4):891–902.

Tzeng J-Y, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale M M, Worrall BB, Hsu F-C, Thomas DC, Sullivan PF. 2011. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *Am J Hum Genet* 89(2):277–288.

Vardarajan BR. 2013. *Identification of Gene-Gene Interactions for Alzheimer's Disease Using Co-operative Game Theory*. PhD thesis, Boston University, Boston, MA.

Visscher PM, Goddard ME, Derks EM, Wray NR. 2012. Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. *Mol Psychiatry* 17(5):474–485.

Walther DJ, Peter J-U, Bashammakh S, Hrtnagl H, Voits M, Fink H., Bader M. 2003. Synthesis of serotonin by a second tryptophan hydroxylase isoform. *Science* 299(5603):76.

Wang T, Elston RC. 2007. Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am J Hum Genet* 80(2):353–360.

Wang K, Li M, Hakonarson H. 2010. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet* 11(12):843–854.

Wang Y, Nie J, Yap P-T, Shi F, Guo L, Shen D. 2011. Robust Deformable-surface-based Skull-stripping for Large-scale Studies. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2011*. Springer, Berlin Heidelberg, pp. 635–642.

Wei Z, Li M, Rebbeck T, Li H. 2008. U-statistics-based tests for multiple genes in genetic association studies. *Ann Hum Genet* 72(6):821–833.

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L and others. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucl Acids Res* 42(D1):D1001–D1006.

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. 2010. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet* 86(6):929–942.

Yagami T, Ueda K, Asakura K, Sakaeda T, Nakazato H, Kuroda T, Hata S, Sakaguchi G, Itoh N, Nakano T and others. 2002. Gas6 rescues cortical neurons from amyloid beta protein-induced apoptosis. *Neuropharmacology* 43(8):1289–1296.

Yang Y, Zou H. 2015. A fast unified algorithm for solving group-lasso penalized learning problems, *Statistics and Computing* 25:1129–1141.

Zhang Y, Pan W. 2015. Principal component regression and linear mixed model in association analysis of structured samples: competitors or complements? *Genet Epidemiol* 39(3):149–155.

Zhou K, Dempfle A, Arcos-Burgos M, Bakker SC, Banaschewski T, Biederman J, Buitelaar J, Castellanos FX, Doyle A, Ebstein RP and others. 2008. Meta-analysis of genome-wide linkage scans of attention deficit hyperactivity disorder. *Am J Med Genet B Neuropsychiatr Genet* 147B(8):1392–1398.

Zhou X, Carbonetto P, Stephens M. 2013. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet* 9(2):e1003264.