PLOS ONE

RESEARCH ARTICLE

An Efficient Stepwise Statistical Test to Identify Multiple Linked Human Genetic Variants Associated with Specific Phenotypic Traits

Iksoo Huh[1], **Min-Seok Kwon**[2], **Taesung Park**[1,2]*

1 Department of Statistics, Seoul National University, Gwanak-gu, Seoul, Korea, **2** Interdisciplinary Program in Bioinformatics, Seoul National University, Gwanak-gu, Seoul, Korea

* tspark@stats.snu.ac.kr

Citation: Huh I, Kwon M-S, Park T (2015) An Efficient Stepwise Statistical Test to Identify Multiple Linked Human Genetic Variants Associated with Specific Phenotypic Traits. PLoS ONE 10(9): e0138700. doi:10.1371/journal.pone.0138700

Editor: Zhongxue Chen, Indiana University Bloomington, UNITED STATES

Received: May 4, 2015

Accepted: September 2, 2015

Published: September 25, 2015

Data Availability Statement: The Wellcome Trust Case Control Consortium (WTCCC) data is available by application to the Wellcome Trust Case Control Consortium Data Access Committee. The data request should be made on the website https://www.sanger.ac.uk/legal/DAA/MasterController. Any inquiries should be sent to cdac@sanger.ac.uk. The Korea Association Resource (KARE) project data will be publicly distributed by the Distribution Desk of Korea Biobank Network (https://koreabiobank.re.kr/). The data request should be made directly to Distribution Desk of Korea Biobank Network. Any

Abstract

Recent advances in genotyping methodologies have allowed genome-wide association studies (GWAS) to accurately identify genetic variants that associate with common or pathological complex traits. Although most GWAS have focused on associations with single genetic variants, joint identification of multiple genetic variants, and how they interact, is essential for understanding the genetic architecture of complex phenotypic traits. Here, we propose an efficient stepwise method based on the Cochran-Mantel-Haenszel test (for stratified categorical data) to identify causal joint multiple genetic variants in GWAS. This method combines the CMH statistic with a stepwise procedure to detect multiple genetic variants associated with specific categorical traits, using a series of associated I × J contingency tables and a null hypothesis of no phenotype association. Through a new stratification scheme based on the sum of minor allele count criteria, we make the method more feasible for GWAS data having sample sizes of several thousands. We also examine the properties of the proposed stepwise method via simulation studies, and show that the stepwise CMH test performs better than other existing methods (e.g., logistic regression and detection of associations by Markov blanket) for identifying multiple genetic variants. Finally, we apply the proposed approach to two genomic sequencing datasets to detect linked genetic variants associated with bipolar disorder and obesity, respectively.

Introduction

Many comprehensive genome-wide association studies (GWAS) have now been conducted to identify previously unknown single nucleotide polymorphisms (SNPs) associated with numerous normal and pathological phenotypes. These novel genetic markers are well tabulated in GWAS catalogs that are updated regularly [1]. However, the majority of such genetic markers are obtained via single-marker analysis, due to the constraints of commonly used statistical methods for judging genetic associations [2,3]. This limitation is problematic to the

inquiries should be sent to admin@koreabiobank.re.kr.

Funding: This work was supported by a grant funded by the National Research Foundation (NRF) of the Korea government (MSIP) (2012R1A3A2026438), and by the Bio-Synergy Research Project (2013M3A9C4078158) of the Ministry of Science, ICT, and Future Planning, also through the NRF.

Competing Interests: All authors declare no competing interests.

advancement of biomedicine, as it is now known that complex diseases which severely impact the health of the general population, are coordinately influenced by multiple genetic factors.

Because the biological and biochemical pathways related to genetic markers often interact to induce disease states, the importance of identifying specific pathway-associated genetic variants is increasingly being recognized [4,5]. Furthermore, joint identification may improve the predictive performance of specific types of statistical analyses [6]. However, traditional models, with logistic regression as one salient example, cannot easily handle multiple genetic markers [4,7,8] when there are sparse cell counts in contingency tables of a categorical trait and single nucleotide polymorphism (SNP) combinations, because the standard errors of parameter estimates tend to be inflated, and the p-value becomes close to one. In addition, extensive computing time is required for estimating parameters in the presence of numerous SNPs.

Consequently, several methods have been proposed to overcome these problems. One such method, penalized logistic regression, is efficient when a large set of SNPs is needed [6]. For example, when penalized logistic regression is fitted, multiple SNPs expected to jointly affect a disease phenotype are selected from the model. However, this approach is not able to simultaneously handle whole SNP datasets owing to its intensive computational burden. Therefore, a small number of SNPs (e.g., 1000 SNPs), having strong marginal effects, are filtered from a single SNP analysis prior to the application of penalized logistic regression [6]. Even with this limitation, in most cases, penalized logistic regression still provides a large number of SNPs, which complicates further biological interpretation.

A second approach, multifactor dimensionality reduction (MDR), is a nonparametric, model-free, and combinatorial approach for interaction analysis that identifies a multi-locus model for association in case-control studies [9]. The MDR method reduces multi-locus genotypes into high- vs. low-risk disease groups. If the ratio of cases and controls in a combination of genotypes is larger than a pre-assigned threshold T (e.g., T = 1), the cell of combination is labeled "high risk;" if smaller than the threshold, it is labeled "low risk." Based on the label of each cell in the contingency table, MDR runs 10-fold cross-validation to select an SNP set with the smallest prediction error and/or the most consistently large training accuracies. Thus, this method avoids the sparsity problem by assuming that sparse cells are undefined. However, since MDR selects k-way interactions purely by the prediction performance of an exhaustive search, it is impractical to detect high-order interactions. Additionally, although some MDR approaches were proposed to reduce computation time [10], or space searching [11], detection of $>$ 3-way interactions from GWAS data is not yet possible.

In addition to the above, a variety of other methods have been proposed to identify gene-gene interactions. For example, Detection of ASSOciations using Markov Blanket (DASSO-MB) was proposed to detect interactions via a Markov blanket used to shield a specific variable from all other variables [12]. This method employs a goodness-of-fit test combined with a stepwise procedure. In some simulation settings, this method outperforms MDR. However, the method has their own drawbacks, such as increased degrees of freedom upon the addition of SNPs.

In this report, we propose a stepwise method for the identification of SNPs that jointly associate with specific phenotypes. Our method uses Cochran-Mantel-Haenszel (CMH) statistics, commonly used in contingency table analysis [13,14] to sequentially test the conditional independence phenotypes from genetic factors. Although the use of CMH statistics for association tests was previously proposed [15], it implementation has proved impractical to handle GWAS datasets, due to the number of strata, derived by distinct genotype combinations, increasing exponentially with increased size of the selected SNP set. To resolve this limitation, we propose a new criterion for stratification categorized by the sum of minor allele counts (MACs). This categorization alleviates intensive computational burden, and therefore facilitates the joint

identification of high-order SNPs in large sample datasets. In addition, we also discuss application of ordinal phenotypes. In the results section, we use simulations to compare our method with stepwise logistic regression and DASSO-MB. Finally, we apply our modified CMH approach to two GWAS datasets to detect collective multiple genetic variants related to bipolar disorder and obesity, respectively.

Materials and Methods

Generalized CMH Method

The original CMH test proposed by Mantel and Haenszel is the method to tests conditional independence of $2\times2\times K$ contingency tables [16,17], meaning that the method is commonly used to test for conditional independence between two binary variables, after adjusting for the effect of confounding variables with K strata. Statistics of the test follow chi-square distributions with one degree of freedom and perform best when the associations of two binary variables have the same directions in each partial table. Situational application of this approach has been generalized by Birch [18], Landis [19], and Mantel [20] to an $I\times J\times K$ table in which the predictor variable and the response variable have I and J levels, respectively, that can be treated as not only nominal but also as ordinal. Therefore, the generalized CMH method consists of two more tests, in addition to the conditional independence test for two nominal variables. One test examines the mean score difference when one variable is ordinal, and the other test evaluates the correlation when both variables are ordinal [20]. The generalized CMH statistics is given as

$$L^2 = [\sum_k B_k(n_k - \mu_k)]'[\sum_k B_k V_k B_k']^{-1}[\sum_k B_k(n_k - \mu_k)]$$

In the above equation, $\boldsymbol{B_k}$ is the Kronecker product between the row score $\mathbf{u_k}$ and the column score $\mathbf{v_k}$, $\mathbf{n_k}$ and $\boldsymbol{\mu_k}$ are vectors of observed and expected counts of length of $I\times J$ in the k^{th} strata, respectively. $\boldsymbol{V_k}$ is an $(I\times J)\times(I\times J)$ variance matrix of $\mathbf{n_k}$, evaluated under an assumed hypergeometric distribution. Therefore, $\mathbf{n_k}$ and $\boldsymbol{\mu_k}$ is represented as $(n_{11k},n_{12k},\ldots,n_{IJk})$ and $(n_{1+k}\times n_{+1k},n_{1+k}\times n_{+2k},\ldots,n_{I+k}\times n_{+Jk})/n_{++k}$ respectively. Moreover, elements of $\boldsymbol{V_k}$ consist of covariance terms between n_{ijk} and $n_{i'j'k}$, and are represented as $n_{i+k}(\omega_{ii'}n_{++k} - n_{i'+k})n_{+jk}(\omega_{jj'}n_{++k} - n_{+j'k})/(n_{++k}^2(n_{++k} - 1))$ where $\omega_{ab} = 1$, when $a = b$ and $\omega_{ab} = 0$ otherwise.

Three types of tests can be derived by imposing ordinal or nominal weights on $\mathbf{u_k}$ and $\mathbf{v_k}$. When $\mathbf{u_k}$ is used as the nominal variable, it is described as a $(I-1)\times I$ matrix $(\boldsymbol{I}, \mathbf{-1})$, where $\boldsymbol{I}$ is an identity matrix of size $I-1$, and $\mathbf{1}$ denotes a column vector of $I-1$ ones. When $\mathbf{u_k}$ is used as the ordinal variable, it is given as $(u_1,u_2,\ldots,u_I)$, with an ordered score vector given to each level of predictor. $\mathbf{v_k}$ is constructed similarly with $\mathbf{u_k}$. Therefore, the general association test is conducted if both variables are nominal, the mean score test is conducted if only one variable is ordinal, and the correlation test is conducted if both variables are ordinal. The degrees of freedom are given as $(I-1)\times(J-1)$ for the general association test, $I-1$ or $J-1$ for the mean score test, and 1 for the correlation test [21].

Application of the CMH Method to SNP Data

We next applied the CMH test to identify genetic variants that mutually associate with a trait of interest. Let Y represent the trait status of a subject, for example, a specific disease. The number of values which Y can have is two for a binary trait, or >2 for ordinal or multinomial traits. Let I denote the number of values which Y can have. If there is one SNP associated with Y, the data can be summarized by an $I\times3$ contingency table. We further assume that the genetic model is codominant for generality, and the CMH test is performed without stratification.

Strata1	AA	Aa	Aa	Total
Trait1	n_{111}	n_{121}	n_{131}	n_{1+1}
Trait2	n_{211}	n_{221}	n_{231}	n_{2+1}
...	...	...	...	...
Trait J	n_{J11}	n_{J21}	n_{J31}	n_{J+1}
Total	n_{+11}	n_{+21}	n_{+31}	n_{++1}

Strata2	AA	Aa	aa	Total
Trait1	n_{112}	n_{122}	n_{132}	n_{1+2}
Trait2	n_{212}	n_{222}	n_{232}	n_{2+2}
...	...	...	...	...
Trait J	n_{J11}	n_{J21}	n_{J31}	n_{J+1}
Total	n_{+11}	n_{+21}	n_{+31}	n_{++1}

....

Strata K	AA	Aa	aa	Total
Trait1	n_{11k}	n_{12k}	n_{13k}	n_{1+k}
Trait2	n_{211}	n_{221}	n_{231}	n_{2+1}
...	...	...	...	...
Trait J	n_{J11}	n_{J21}	n_{J31}	n_{J+1}
Total	n_{+11}	n_{+21}	n_{+31}	n_{++1}

Fig 1. Contingency table for CMH test between trait and *p* SNPs. This test can be represented as H_0: $Y \perp$SNP|(number of *p*-1 SNPs). "A" and "a" represent major and minor alleles, respectively.

doi:10.1371/journal.pone.0138700.g001

When there are two SNPs, S_1 and S_2, the data can be summarized in a $I\times3\times3$ contingency table, and the CMH test evaluates the conditional independence between Y and S_1, given S_2. When there are three SNPs, S_1, S_2, and S_3, the data can be summarized in a $I\times3\times3^2$ contingency table and the CMH tests the association between Y and S_1, given (S_2, S_3). Consequently, the genetic frequency data of case-control study can be summarized in an $I\times J\times K$ contingency table, where $I = 2$ $J = 3$, $K = 3^{p-1}$, and p = the total number of SNPs (Fig 1). However, this stratification scheme has the potential problem that K would be too large in the case of many SNPs. Because this scheme may require excessive computation, it is impractical to apply this stratification scheme to GWAS data. For example, if there are 10 SNPs in hand, the number of strata could reach $3^9 = 19683$, indicating that this dataset is too divided to accurately reflect its properties. Therefore, we propose an alternative stratification scheme, based on the assumption that subjects with similar sums of minor allele counts (MACs) may have similar risks of disease traits [22,23]. In addition, we can fix the maximum number of strata through clustering subjects whose MACs exceed a predefined criterion as one stratum. This new scheme makes computation much faster than the former scheme, and we expect that several dozen SNPs can be easily identified in a reasonable time.

Stepwise CMH Test

The general CMH test has $(I-1)\times(J-1)$ degrees of freedom. Therefore, when we apply this test to the codominant genetic model, it follows a chi-squared distribution with $(I-1)\times2$ degrees of freedom at most. Next, the test for the additive, dominant or recessive genetic models can be similarly developed using the mean score of the CMH test [20].

To identify multiple causal SNPs among a large total number of SNPs, we propose the following stepwise CMH procedure, following determination of stratification criteria based on MACs. Firstly, in the forward step, the most significant SNP associated with the disease of interest is added to the previously selected SNPs via the CMH test. Therefore, if there are N SNPs in the total dataset, and p (= 0 for the first step) SNPs previously selected by the test, a SNP whose CMH statistic p-value conditioned on the p selected SNPs is the smallest and smaller than the threshold would be added. For the first forward step, the CMH test is applied without stratification.

Secondly, in the backward step, we implement the CMH test to remove the least significant SNPs among the previously selected SNP set. All SNPs in the set are tested in the presence of all of the other SNPs. If any SNP has a p-value that exceeds the removal threshold, the SNP whose p-value is largest is excluded from the SNP set; otherwise, this step does not remove any SNPs. This backward step can be optionally skipped by the researcher.

Our stepwise CMH method iterates between the forward and backward steps until no additional variable is added to the current model at the forward step. Once a set of significant SNPs is identified, these are removed from the whole SNP dataset, and the first step of the stepwise procedure is repeated. The whole stepwise procedure is repeated until no more significant SNPs are selected in the first forward step.

Results and Discussion

Simulation Studies

To investigate the utility of our stepwise CMH method, we conducted a simulation study to compare it with the logistic regression and DASSO-MB approaches. For simplicity, we only considered the main effects of two and three causal SNP models. The odds and penetrances for the three causal SNP model are provided in Tables 1 and 2 [24]. The corresponding true logistic model for binary traits is assumed as follows:

$$\mathbf{logit}(p(y=1)) = \beta_0 + \sum_{p=1}^{P} \beta_p \times SNP_p$$

To generate datasets according to the true model, we first determined the total penetrance, which defines the proportion of cases in whole samples. Then, we set the values for the baseline effect α and genetically additive effect θ. In case of the codominant model and binary traits, θ has two values: θ_1 and θ_2, where θ_1 is the marginal effect between major homogeneous and heterogeneous genotypes, given by 0.7, and θ_2 is the marginal effect between major and minor homogeneous genotypes, given by 0.5.

In this simulation study, we assumed linkage equilibrium between causal SNPs, with minor allele frequencies (MAFs) set to 0.03, 0.05, 0.1, and 0.2. We then generated 1000 datasets, each consisting of 1000 cases, 1000 controls. We set the number of SNPS to be 100, 300, 500, and 1,000 (including non-causal SNPs). Two accuracy measures were used to compare the stepwise CMH method to others. First, the detection probability (Dprob) was estimated by dividing the number of correctly captured SNPs by the total number of true SNPs. Second, the proportion of datasets out of all 1000 datasets that detected all of the causal SNPs was evaluated (power). In addition, two threshold values were used to evaluate significance: Bonferroni correction criterion 5×10^{-4} and a looser criterion 5×10^{-3}.

The simulation results are shown in Figs 2 and 3. Fig 2 shows the results for the model with two causal SNPs, and Fig 3 for the model with three causal SNPs. For the codominant model with binary traits, both accuracy measures for the stepwise CMH method, when the MAF was relatively low (0.03 and 0.05), were clearly greater than those of the stepwise logistic method. However, with moderate MAFs (0.10 and 0.20), the two approaches provided comparable results. This is because when the MAF value is small enough to induce sparse minor allele counts of some strata, logistic regression produces very large standard errors, and some p-values of the coefficients are greatly inflated. However, the CMH statistic has a more robust variance estimate that is not substantially affected by sparse cells, due to the fact that the CMH

Table 1. Odds table for simulation studies and binary trait in three causal SNP model ($\theta_0 = 0$).

((#of c) = *k*)	*AA*	*Aa*	*Aa*
BB	$\alpha(1+\theta_k)$	$\alpha(1+\theta_1)(1+\theta_k)$	$\alpha(1+\theta_2)(1+\theta_k)$
Bb	$\alpha(1+\theta_1)(1+\theta_k)$	$\alpha(1+\theta_1)^2(1+\theta_k)$	$\alpha(1+\theta_1)(1+\theta_2)(1+\theta_k)$
Bb	$\alpha(1+\theta_2)(1+\theta_k)$	$\alpha(1+\theta_1)(1+\theta_2)(1+\theta_k)$	$\alpha(1+\theta_2)^2(1+\theta_k)$

doi:10.1371/journal.pone.0138700.t001

Table 2. Penetrance table for Table 1 in three causal SNP model ($\theta_o = 0$).

(#of $c = k$)	*AA*	*Aa*	*Aa*
BB	$\frac{\alpha(1+\theta_k)}{1+\alpha(1+\theta_k)}$	$\frac{\alpha(1+\theta_1)(1+\theta_k)}{1+\alpha(1+\theta_1)(1+\theta_k)}$	$\frac{\alpha(1+\theta_2)(1+\theta_k)}{1+\alpha(1+\theta_2)(1+\theta_k)}$
Bb	$\frac{\alpha(1+\theta_1)(1+\theta_k)}{1+\alpha(1+\theta_1)(1+\theta_k)}$	$\frac{\alpha(1+\theta_1)^2(1+\theta_k)}{1+\alpha(1+\theta_1)^2(1+\theta_k)}$	$\frac{\alpha(1+\theta_1)(1+\theta_2)(1+\theta_k)}{1+\alpha(1+\theta_1)(1+\theta_2)(1+\theta_k)}$
bb	$\frac{\alpha(1+\theta_2)(1+\theta_k)}{1+\alpha(1+\theta_2)(1+\theta_k)}$	$\frac{\alpha(1+\theta_1)(1+\theta_2)(1+\theta_k)}{1+\alpha(1+\theta_1)(1+\theta_2)(1+\theta_k)}$	$\frac{\alpha(1+\theta_2)^2(1+\theta_k)}{1+\alpha(1+\theta_2)^2(1+\theta_k)}$

doi:10.1371/journal.pone.0138700.t002

variance is only calculated using marginal counts. Therefore, the performance of the stepwise CMH method is better than that of the stepwise logistic method. This pattern is consistent regardless of the number of SNPs.

By contrast, DASSO-MB generally showed low power and Dprob values, with a few exceptions. DASSO-MB is unfavorable in situations when a model contains only main effects, due to increased degrees of freedom of tests with increased numbers of variables included in the selected set. For example, if the second SNP in the forward step was tested once a SNP was already selected, the degrees of freedom could reach six (Fig 2), while including a third variable could result in 18 degrees of freedom (Fig 3). Such increases in degrees of freedom result in decreased power.

Comparison of CMH with Two Other Methods via a Toy Example

To demonstrate the superiority of the stepwise CMH method more clearly, we provided an artificially generated toy example consisting of samples of 50 cases and 50 controls, and genotypes of two informative SNPs denoted S_1 and S_2, respectively. The structure of the dataset is

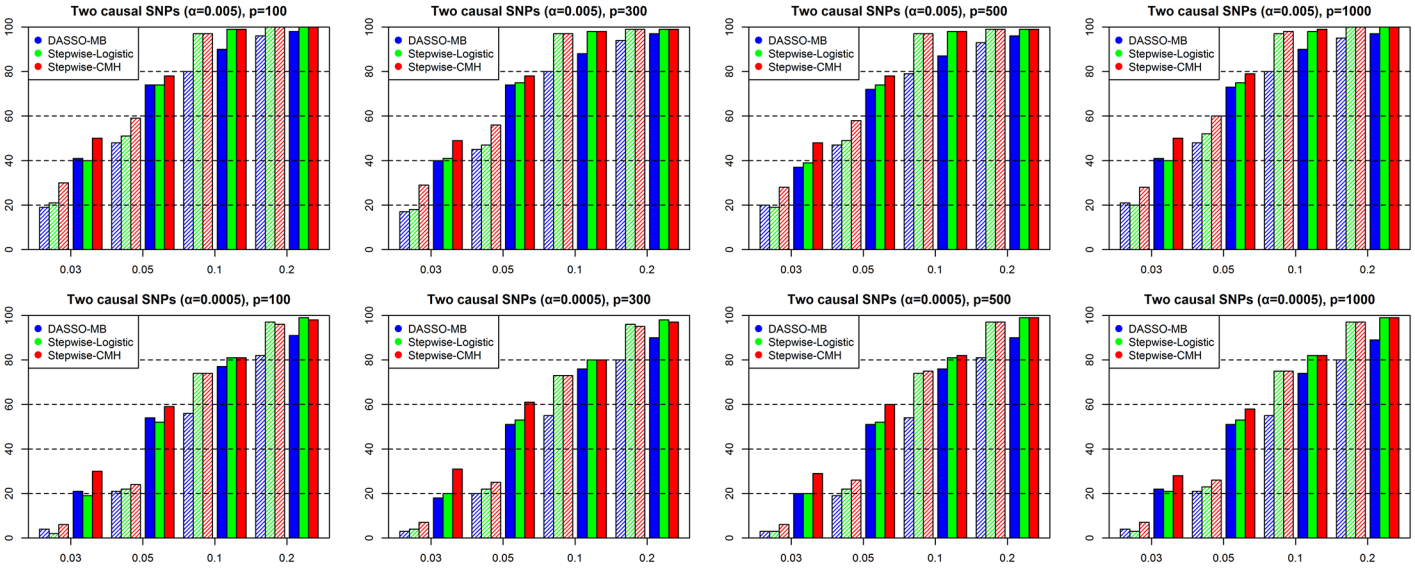

Fig 2. Performance comparison of stepwise CMH method, stepwise logistic and DASSO-MB methods for the codominant model with two causal SNPs. Blue bars represent the result of the DASSO-MB method, green bars represent the result of the stepwise logistic method, and red bars represent the result of the stepwise CMH method. Bars with diagonal lines are the results of power and solid bars are those of Dprob. The x-axis represents the MAF of true causal SNPs and the y-axis represents the value of two accuracy measures based on three approaches.

doi:10.1371/journal.pone.0138700.g002

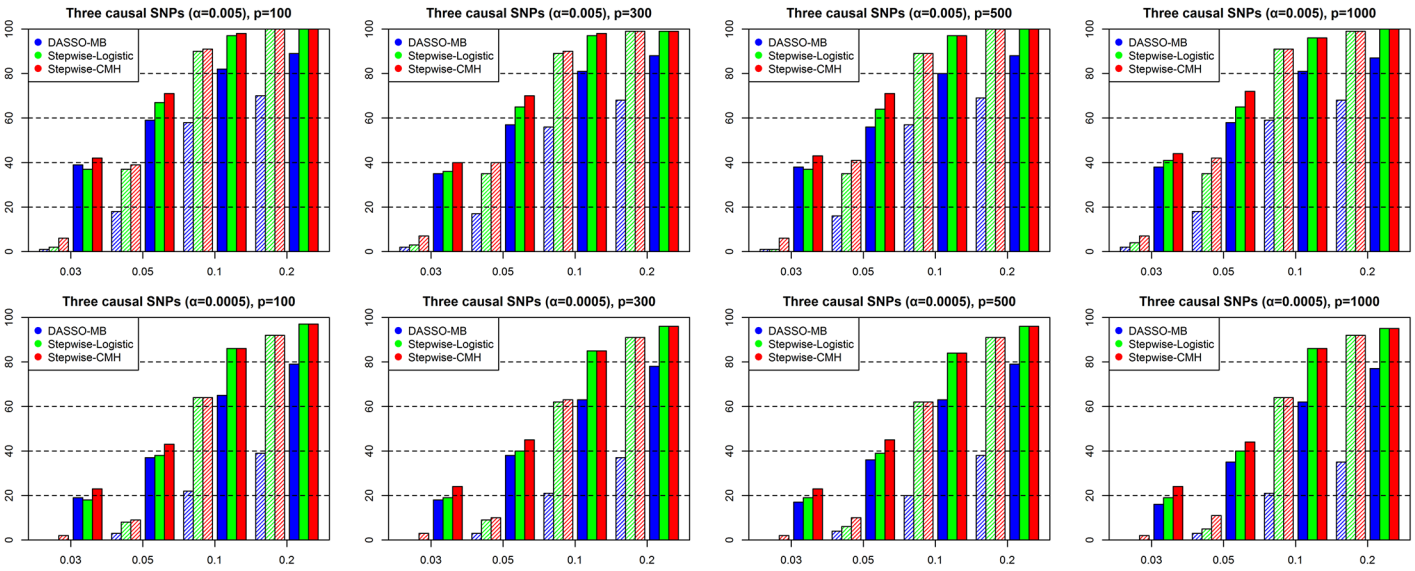

Fig 3. Performance comparison of stepwise CMH method, stepwise logistic and DASSO-MB methods for the codominant model with three causal SNPs.

doi:10.1371/journal.pone.0138700.g003

shown in Fig 4. If we set an entrance cutoff of 0.05, S_1 would be selected in the first forward step by all three methods, because all three p-values would be < 0.05. However, in the second forward step, S_2 was selected only by the CMH method, while some sparse cells in the genotype table resulted in p-value inflation in logistic regression (Fig 4, Table 3). DASSO-MB also could not detect the second SNP, because this method allows larger degrees of freedom than the other two methods (Table 3).

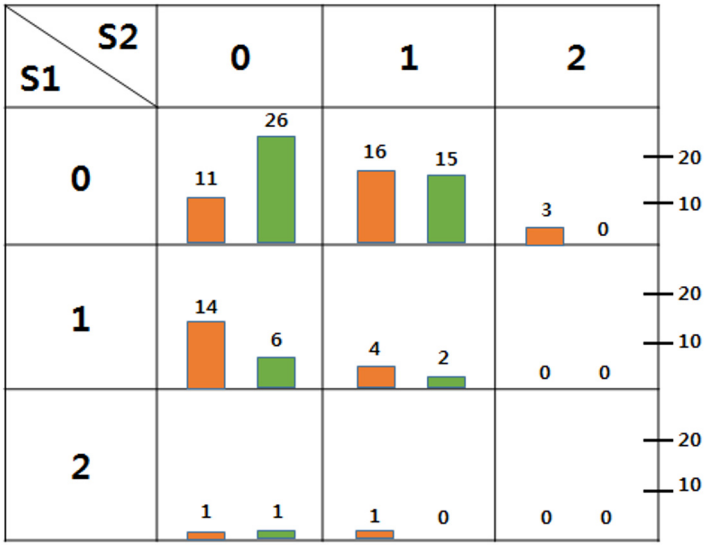

Fig 4. Toy example dataset structure to show superiority of the stepwise CMH method. We counted the numbers of cases and controls for each genotype combination and expressed them as vertical bars to visualize. Orange bars represent the counts of cases and the green bars do counts of controls.

doi:10.1371/journal.pone.0138700.g004

Table 3. Toy example dataset application result.

Methods	p-value (1st forward step)	DF	p-value (2nd forward step)	DF
Stepwise CMH	0.0146	2	0.0304	2
Stepwise Logistic	0.0203	2	0.2450	2
DASSO-MB	0.0498	2	0.1330	6

doi:10.1371/journal.pone.0138700.t003

Application of the CMH Approach for Analyzing Wellcome Trust Case Control Consortium (WTCCC) Bipolar Disorder Data

We next applied our proposed CMH method to a WTCCC bipolar disorder (BD) dataset. BD is well known to be highly heritable and polygenic [25,26], and the dataset consisted of 2,938 controls and 1,868 cases, and a total of 354,022 genome-wide SNP markers. We filtered out SNP markers with MAFs $< 1\%$, or having call rates $< 95\%$. We also selected tagged SNPs, based on the criteria of $r^2 > 0.5$ between 50 adjacent SNPs. These filtering steps resulted in 134,254 SNPs usable for our analysis. We then applied the stepwise CMH method to the dataset, based on a codominant genetic model. We set the entry and removal threshold value of significance as $\alpha = 5\times10^{-5}$. After all the stepwise procedures were completed, two SNP sets were selected (Table 4).

In the first SNP set, *rs11112069*, located in the gene *CHST11*, encoding carbohydrate sulfotransferase-11, which catalyzes sulfate transfer to position 4 of the N-acetylgalactosamine (GalNAc) residue of chondroitin. Chondroitin sulfate is known to facilitate axonal patterning and cell migration during the early growth and development of the mammalian central nervous system [27]. *CHST11* is also related to neuronal function, suggesting a possible (but yet unknown) relationship to BD. Another SNP, *rs420259*, is located in the partner and localizer of *BRCA2* (*PALB2*), which facilitates DNA repair by recruiting *BRCA2* and *RAD51* to double-stranded DNA breaks, and *PALB2* and *BRCA2* have been associated with both BD and schizophrenia in a Scandinavian study [28]. *BRCA2*, expressed in the mouse brain, was shown to be important for normal neurogenesis, particularly in the cerebellum, a region involved in emotional processing that is often dysfunctional in BD [29]. *rs17484671* is located in the gene NR3C2, encoding nuclear receptor subfamily 3, group C, and member 2 [30], the drug target for dipolar disorder [31]. The SNP *rs12537100* is located in an intronic region of the gene *THSD7A*, thrombospondin, type I, domain containing 7A. Thrombospondins are key regulators of synaptogenesis in the central nervous system [32]. *rs7260296*, located 10KB downstream of the gene *NTE*, also known as *PNPLA6*, patatin-like phospholipase domain containing 6. NTE is a lysophospholipase that maintains intracellular phospholipid homeostasis by converting lysophosphatidylcholine to glycerophosphocholine [33]. *PNPLA6* is directly related to

Table 4. WTCCC bipolar disorder data analysis result (Entrance cutoff = 5×10^{-5}, Removal cutoff = 5×10^{-5}).

Set	SNP name	GENE
1st Set	*rs11112069*,*rs420259*,*rs17484671*, *rs7260296*, *rs4918068*,*rs6705537*, *rs12594576*, *rs6908950*, *rs11984645*, *rs8021692*, *rs4027132*, *rs4276227*, *rs7152966*, *rs17561681*, *rs9510385*, *rs4567706*	*CHST11*, *PALB2*, *NR3C2*, *PNPLA6*, *OBFC1*, *USP34*, *FAH*, *GLTSCR1L*, *OPRK1*, *TDRD9*, *LPIN1*, *CMTM8*, *TSHR*, *GABRA5*, *unknown*, *unknown*,
2nd Set	*rs7184080*, *rs12537100*, *rs2609653*	*LOC101928392*, *THSD7A*, *unknown*,

doi:10.1371/journal.pone.0138700.t004

Table 5. KARE BMI data analysis result (Entrance cutoff = 5×10^{-5}, Removal cutoff = 5×10^{-5}).

Set	SNP name	GENE
1st Set	*rs6893893,rs2196534,rs6079272, rs1736913, rs4639483,rs6462517, rs4518599,rs10878690, rs1012780, rs7107562,rs11682163*	*ATP10B, Unknown, MACROD2, HLA-F-AS1, RSPO2, AK025321, HIP1, AK055974, LOC100289473, Unknown, ALLC*
2nd Set	*rs6656287*	*ZCCHC17*

doi:10.1371/journal.pone.0138700.t005

neuronal function [34], and its dysfunction may associate with the onset of BD. Moreover, the majority of other SNPs in our set were reported previously [35–38].

In summary, many selected SNPs directly or indirectly related to neuronal function. Therefore, joint identification of the putative causal SNPs could provide more biologically meaningful interpretation and motivation of further investigation, such as pathway analysis.

Application to Korea Association Resource (KARE) project

We also applied our stepwise CMH method to a GWA dataset from the Korean Association REsource (KARE) project, initiated in 2007 to undertake a large-scale GWAS of 260 traits among 10,038 participants (aged between 40 and 69) of Ansung (n = 5,018) and Ansan (n = 5,020) population-based cohorts. Among the 260 traits, we selected body mass index (BMI) to detect causal variants associated with obesity. Here, BMI was treated as an ordinal variable with four categories: normal ($18.5 \leq$ BMI < 25), overweight ($25 \leq$ BMI < 30), mildly obese ($30 \leq$ BMI < 35), and severely obese ($35 \leq$ BMI < 40), and the subjects were numbered from 1 (normal) to 4 (severe), respectively. The dataset consisted of 8842 individuals with a total of 352,228 genome-wide SNPs. We filtered out SNP markers with MAFs $< 1\%$ or call rates $< 95\%$. We also selected tagged SNPs, based on the criteria of $r^2 > 0.5$ between 50 adjacent SNPs. These criterions resulted in 137,400 SNPs usable for our analysis, with the same cutoff used in our analysis of the BD dataset. After all stepwise procedures were completed, the two SNP sets were selected (Table 5). Among the selected SNPs, we found no SNPs that were reported in previous studies. However, in the first SNP set, *ATP10B*, *MACROD2*, *and HIP2*, to which *rs6893893*, *rs6079272*, *and rs4518599* respectively annotated, were reported to associate with various BMI-related traits [39–41]. Moreover, in the second SNP set, *ZCCHC17*, the location of *rs6656287*, was previous associated with alcohol dependence, which may affect eating behavior [42,43].

Conclusions

Our stepwise CMH method has two large advantages over stepwise logistic regression. The first is that it addresses the sparsity problem, as variance inflation can only be induced in the presence of sparse cells of a genotype count table. Secondly, while logistic regression suffers from intensive computing time (necessary for its iterative optimization algorithm), the stepwise CMH test avoids this problem, as the CMH test statistic is calculated by a simple matrix operation, and the standard error is not affected by the sparsity of cells. In GWAS, as the number of SNPs increases, the chance of including rare SNPs in the stepwise procedure also increases, making it difficult for logistic regression to identify high-order joint identification. Therefore, the stepwise CMH approach is a more appropriate approach than stepwise logistic regression for identification of rare variants in GWAS.

Even though the CMH statistic was originally proposed for detecting conditional independence, a specific SNP set identified via the stepwise CMH approach is informative for

identifying joint genetic variants, as the forward and backward steps guarantee that all of the components in the SNP set are significant in the presence of the other SNPs.

Recently, many variable selection methods were developed which use penalization such as LASSO or SCAD [44,45]. We have not directly compared the proposed CMH method to these penalized approaches, due to the fact that our stepwise CMH method tends to select a small number of SNPs with joint effects, while the penalized approaches tend to select a large number of SNPs, if an optimal value of tuning parameter is selected via cross validation. In future comparative studies, we will compare our stepwise CMH to the penalization approaches, while also controlling the number of variables selected. In the presence of ordinal or multinomial traits [46,47], we expect the usefulness of our approach to increase.

Our method focuses on statistical analysis of common variants from GWAS. The traditional GWAS are usually based on the assumption of common disease and common variant (CD-CV). A next generation sequencing (NGS) technique adopts the assumption of common disease and rare variant (CD-RV). Recently, several gene-based aggregation methods for the analysis of rare variants have been proposed [48, 49]. A more complete review of aggregation methods, please refer to [50, 51]. Those aggregation methods are powerful in detecting causal rare variants which are expected to explain missing heritability. However, they may have low power when only a small portion of variants are causal in a region [52]. We are working on developing the stepwise CMH type of statistics for the rare variant analysis.

R code for the stepwise CMH test is provided at a dedicated website (http://bibs.snu.ac.kr/software/stepCMH).

Acknowledgments

The authors thank Joon Yoon for helpful comments. The authors also thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

Author Contributions

Conceived and designed the experiments: IH MK TP. Performed the experiments: IH. Analyzed the data: IH MK. Contributed reagents/materials/analysis tools: IH. Wrote the paper: IH MK TP.

References

1. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014; 42: 1001–1006.
2. Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. Theor Popul Biol, 2001; 60: 155–166. PMID: 11855950
3. Liu L, Zhang D, Liu H, Arendt C. Robust methods for population stratification in genome wide association studies. BMC Bioinformatics, 2013; 14:132. doi: 10.1186/1471-2105-14-132 PMID: 23601181
4. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 2005; 37: 413–417. PMID: 15793588
5. Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012; 44: 369–375. doi: 10.1038/ng.2213 PMID: 22426310
6. Cho S, Kim K, Kim YJ, Lee JK, Cho YS, Lee JY, et al. Joint Identification of Multiple Genetic Variants via Elastic-Net Variable Selection in a Genome-Wide Association Analysis. Ann Hum Genet. 2010; 74: 416–428. doi: 10.1111/j.1469-1809.2010.00597.x PMID: 20642809
7. He H, Oetting WS, Brott MJ, Basu S. Power of multifactor dimensionality reduction and penalized logistic regression for detecting gene-gene Interaction in a case-control study. BMC Med Genet. 2009; 10:127. doi: 10.1186/1471-2350-10-127 PMID: 19961594

8. Park MY, Hastie T. Penalized logistic regression for detecting gene interactions. Biostatistics. 2008; 9:1.
9. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl Fritz F, et al Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. Am J Hum Genet 2001; 69:1.
10. Kwon MS, Kim K, Lee S, Chung W, Yi SG, Namkung J, et al. GWAS-GMDR: A program package for genome-wide scan of gene-gene interactions with covariate adjustment based on multifactor dimensionality reduction. 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops. 703–707. 2011 Nov 12. doi: 10.1109/ BIBMW.2011.6112456.
11. Oh S, Lee J, Kwon MS, Weir B, Ha K, Park T. A novel method to identify high order gene-gene interactions in genome-wide association studies: Gene-based MDR. BMC Bioinformatics. 2012; 13:9.
12. Han B, Park M, Chen XW. A Markov blanket-based method for detecting causal SNPs in GWAS. BMC Bioinformatics. 2010; 11:3.
13. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet. 2008; 40: 616–622. doi: 10.1038/ng.109 PMID: 18385676
14. Rosenstock J, Dailey G, Massi-Benedetti M, Fritsche A, Lin Z, Salzman A. Reduced hypoglycemia risk with insulin glargine: a meta-analysis comparing insulin glargine with human NPH insulin in type 2 diabetes. Diabetes Care. 2005; 28: 950–955. PMID: 15793205
15. Huh IS, Oh S, Park T. A Chi-square test for detecting multiple joint genetic variants in Genome-wide association studies. 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops 708–713. 2011 Nov 12. doi: 10.1109/BIBMW. 2011.6112457
16. Mantel N, Haenszel W. Statistical aspect of the analysis of data from retrospective studies of disease. J Natl Cancer Inst. 1959; 22: 719–748. PMID: 13655060
17. Cochran WG. Some methods of Strengthening the common χ^2 tests. Biometrics 1954; 10: 417–451.
18. Birch MW. The detection of partial association II:The general case. J R Stat Soc Series B 1965; 27:111–124.
19. Landis JR, Heyman ER, Koch GG. Average partial association in three-way contingency tables: A review and discussion of alternative tests. Int Stat Rev. 1978; 46: 237–254.
20. Mantel N. Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel Procedure. J Am Stat Assoc. 1963; 58: 690–700.
21. Alan A. An Introduction to Categorical Data Analysis 2nd edition. John Wiley & Sons.
22. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH.Multiple rare alleles contribute to low plasma levels of hdl cholesterol. Science. 2005; 305: 869–872.
23. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; 83: 311–321. doi: 10.1016/j.ajhg.2008.06.024 PMID: 18691683
24. Li J, Chen Y. Generating samples for association studies based on HapMap Data. BMC Bioinformatics. 2008; 9:44. doi: 10.1186/1471-2105-9-44 PMID: 18218094
25. McGuffin P, Rijsdijk F, Andrew M, Sham P, Katz R, Cardno A. The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. Arch Gen Psychiatry. 2003; 60: 497–502. PMID: 12742871
26. International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460:748–752. doi: 10.1038/nature08185 PMID: 19571811
27. Karumbaiah L, Saxena T, Betancur M, Bellamkonda RV. Chondroitin Sulfate Glycosaminoglycans for CNS Homeostasis—Implications for Material Design. Curr Med Chem. 2014; 21: 4257–4281. PMID: 25139544
28. Tesli M, Athanasiu L, Mattingsdal M, Kähler AK, Gustafsson O, Andreassen BK. Association analysis of PALB2 and BRCA2 in bipolar disorder and schizophrenia in a scandinavian case-control sample. Am J Med Genet B Neuropsychiatr Genet. 2010; 153B: 1276–1282. doi: 10.1002/ajmg.b.31098 PMID: 20872766
29. Bolbecker AR, Mehta C, Johannesen JK, Edwards CR, O'Donnell BF Shekhar A. Eyeblink conditioning anomalies in bipolar disorder suggest cerebellar dysfunction. Bipolar Disord. 2009; 11: 19–32. doi: 10.1111/j.1399-5618.2008.00642.x PMID: 19133963
30. Psychiatric GWAS Consortium Bipolar Disorder Working Group, Sklar P, Ripke S, Scott LJ, Andreassen OA, Cichon S, et al. Large-scale genome-wide association analysis of bipolar disorder identifies a

new susceptibility locus near ODZ4. Nat Genet. 2011; 43:977–983. doi: 10.1038/ng.943 PMID: 21926972

31. Juruena MF, Gama CS, Berk M, Belmonte-de-Abreu PS. Improved stress response in bipolar affective disorder with adjunctive spironolactone (mineralocorticoid receptor antagonist): case series. J Psychopharmacol. 2009; 23: 985–987. doi: 10.1177/0269881108092121 PMID: 18583441

32. Risher WC, Eroglu C. Thrombospondins as key regulators of synaptogenesis in the central nervous system. Matrix Biol. 2012; 31: 170–177. doi: 10.1016/j.matbio.2012.01.004 PMID: 22285841

33. Topaloglu AK, Lomniczi A, Kretzschmar D, Dissen GA, Kotan LD, McArdle CA. Loss of Function Mutations in PNPLA6 Encoding Neuropathy Target Esterase Underlie Pubertal Failure and Neurological Deficits in Gordon Holmes Syndrome. J Clin Endocrinol Metab. 2014; 99:E2067–2075. doi: 10.1210/jc.2014-1836 PMID: 25033069

34. Song Y, Wang M, Mao F, Shao M, Zhao B, Song Z. Knockdown of Pnpla6 protein results in motor neuron defects in zebrafish. Dis Model Mech. 2013; 6:404–413. doi: 10.1242/dmm.009688 PMID: 22996643

35. Wellcome Trust Case Control Consortium, Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 2007; 447: 661–678. PMID: 17554300

36. Zhang Q, Long Q, Ott J. AprioriGWAS, a New Pattern Mining Strategy for Detecting Genetic Variants Associated with Disease through Interaction Effects. PLoS Comput Biol. 2014; 10:6.

37. Ollila HM, Soronen P, Silander K, Palo OM, Kieseppä T, Kaunisto MA, et al. Findings from bipolar disorder genome-wide association studies replicate in a Finnish bipolar family-cohort. Mol Psychiatry. 2009; 14: 351–353. doi: 10.1038/mp.2008.122 PMID: 19308021

38. Craddock N, Jones L, Jones IR, Kirov G, Green EK, Grozeva D, et al.Strong genetic evidence for a selective influence of GABAA receptors on a component of the bipolar disorder phenotype. Mol Psychiatry. 2010; 15: 146–153. doi: 10.1038/mp.2008.66 PMID: 19078961

39. Nolan DK, Sutton B, Haynes C, Johnson J, Sebek J, Dowdy E, et al. Fine mapping of a linkage peak with integration of lipid traits identifies novel coronary artery disease genes on chromosome 5. BMC genet. 2012; 13:12. doi: 10.1186/1471-2156-13-12 PMID: 22369142

40. Slavin TP, Feng T, Schnell A, Zhu X, Elston RC. Two-marker association tests yield new disease associations for coronary artery disease and hypertension. Hum Genet, 2011; 130: 725–733. doi: 10.1007/s00439-011-1009-6 PMID: 21626137

41. Hasstedt SJ, Highland HM, Elbein SC, Hanis CL, Das SK, American Diabetes Association GENNID Study Group. et al. Five Linkage Regions Each Harbor Multiple Type 2 Diabetes Genes in the African American Subset of the GENNID Study. J Hum Genet. 2013; 58:378–383. doi: 10.1038/jhg.2013.21 PMID: 23552671

42. Zuo L, Wang K, Zhang XY, Krystal JH, Li CS, Zhang F, et al. NKAIN1–SERINC2 is a functional, replicable and genome-wide significant risk gene region specific for alcohol dependence in subjects of European descent. Drug Alcohol Depend. 2013; 129: 254–264. doi: 10.1016/j.drugalcdep.2013.02.006 PMID: 23455491

43. Bulik CM, Klump KL, Thornton L, Kaplan AS, Devlin B, Fichter MM et al. Alcohol use disorder comorbidity in eating disorders: a multicenter study. J Clin Psychiatry. 2012; 65: 1000–1006.

44. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol. 1996; 58: 267–288.

45. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001; 96:1348–1360.

46. Kim K, Kwon MS, Oh S, Park T. Identification of multiple gene-gene interactions for ordinal phenotypes. BMC Med Genomics. 2013; 6:S9. doi: 10.1186/1755-8794-6-S2-S9 PMID: 23819572

47. Morris AP, Lindgren CM, Zeggini E, Timpson NJ, Frayling TM, Hattersley AT, et al. A powerful approach to sub-phenotype analysis in population-based genetic association studies. Genet Epidemiol, 2010; 34: 335–343. doi: 10.1002/gepi.20486 PMID: 20039379

48. Li Bingshan, Leal Suzanne M. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. Am J Hum Genet. 2008; 83(3): 311–321 doi: 10.1016/j.ajhg.2008.06.024 PMID: 18691683

49. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test Am J Hum Genet. 2011; 89(1): 82–93 doi: 10.1016/j.ajhg.2011.05.029 PMID: 21737059

50. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. Am J Hum Genet. 2014; 95(1):5–23 doi: 10.1016/j.ajhg.2014.06.009 PMID: 24995866

51. Derkach A, Lawless JF, Sun L. Pooled Association Tests for Rare Genetic Variants: A Review and Some New Results. Stat Sci. 2014; 29(2): 302–321

52. Ament SA, Szelinger S, Glusman G, Ashworth J, Hou L, Akula N, et al. Rare variants in neuronal excitability genes influence risk for bipolar disorder. Proc Natl Acad Sci U S A. 2015; 112(11):3576–81 doi: 10.1073/pnas.1424958112 PMID: 25730879