A Robust and Powerful Set-Valued Approach to Rare Variant Association Analyses of Secondary Traits in Case-Control Sequencing Studies

Guolian Kang,[*,1,2] Wenjian Bi,[*,1] Hang Zhang,[†,‡] Stanley Pounds,[*] Cheng Cheng,[*] Sanjay Shete,[§] Fei Zou,[**] Yanlong Zhao,[†] Ji-Feng Zhang,[†,‡] and Weihua Yue[††,2]

[*]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, [†]Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, People's Republic of China, [‡]School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China, [§]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, [**]Department of Biostatistics, The University of North Carolina at Chapel Hill, North Carolina 27599, and [††]Institute of Mental Health, Key Laboratory of Mental Health, Ministry of Health & National Clinical Research Center for Mental Disorders, Sixth Hospital, Peking University, Beijing 100191, People's Republic of China

ABSTRACT In many case-control designs of genome-wide association (GWAS) or next generation sequencing (NGS) studies, extensive data on secondary traits that may correlate and share the common genetic variants with the primary disease are available. Investigating these secondary traits can provide critical insights into the disease etiology or pathology, and enhance the GWAS or NGS results. Methods based on logistic regression (LG) were developed for this purpose. However, for the identification of rare variants (RVs), certain inadequacies in the LG models and algorithmic instability can cause severely inflated type I error, and significant loss of power, when the two traits are correlated and the RV is associated with the disease, especially at stringent significance levels. To address this issue, we propose a novel set-valued (SV) method that models a binary trait by dichotomization of an underlying continuous variable, and incorporate this into the genetic association model as a critical component. Extensive simulations and an analysis of seven secondary traits in a GWAS of benign ethnic neutropenia show that the SV method consistently controls type I error well at stringent significance levels, has larger power than the LG-based methods, and is robust in performance to effect pattern of the genetic variant (risk or protective), rare or common variants, rare or common diseases, and trait distributions. Because of the SV method's striking and profound advantage, we strongly recommend the SV method be employed instead of the LG-based methods for secondary traits analyses in case-control sequencing studies.

KEYWORDS secondary traits; rare variants association analyses; case-control sequencing study; set-valued model

To date, genome-wide association studies (GWAS) worldwide have detected several thousands of common single-nucleotide polymorphisms (SNPs) with small or moderate risk for over 200 diseases or traits. These GWAS usually originate in a study design focusing on a specific primary trait, such as a case-control design for Parkinson's disease (Simón-Sánchez *et al.* 2009), bipolar disorder, or coronary artery disease (Wellcome Trust Case Control Consortium 2007). Often, besides the primary case-control data, data on many secondary traits that may share the same associated genetic variants with the primary disease are also measured, and are readily available. Nowadays, analyses of secondary traits beyond GWAS have gained prominence because assessing the genetic association of secondary traits may provide critical insights into disease etiology or pathology. For example, a common variant (CV) in the FTO (fat mass and obesity associated) gene predisposes individuals to diabetes through the effect on body mass index (BMI)/obesity (Flayling *et al.* 2007).

doi: 10.1534/genetics.116.192377

Manuscript received June 6, 2016; accepted for publication December 29, 2016; published Early Online December 30, 2016.

Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.192377/-/DC1.

[1] These authors contributed equally to this work.

[2] Corresponding authors: Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105. E-mail: Guolian.kang@stjude.org; and Institute of Mental Health, Key Laboratory of Mental Health, Ministry of Health & National Clinical Research Center for Mental Disorders, Sixth Hospital, Peking University, Beijing 100191, People's Republic of China. E-mail: dryue@bjmu.edu.cn.

Other examples include Grundy *et al.* (2004), Kammerer *et al.* (2004), Kathiresan *et al.* (2008), Loos *et al.* (2008), Willer *et al.* (2008), Teslovich *et al.* (2010), and Edmondson *et al.* 2011. Similar to GWAS, many ongoing whole-genome, or whole-exome, next generation sequencing (NGS) studies, such as the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (Do *et al.* 2015), and the Cohorts for Heart and Aging Research in Genomic Epidemiology resequencing project (Lin *et al.* 2014), also collect various secondary traits data besides the primary trait data.

For genetic association analyses of secondary traits in a case-control study, generalized regression methods are commonly used. The naïve application of regression modeling to the secondary trait ignores the crucial fact that, in a case-control study, the data are not a random, and representative sample of the secondary trait if it is correlated with the primary trait (Lee *et al.* 1997; Lin and Zeng 2009; Monsees *et al.* 2009; Wang and Shete, 2011a; He *et al.* 2012). Several statistical methods have been developed for the secondary traits analysis, taking into consideration the statistical design issue. An inverse-probability-of-sampling-weighted regression method (IPW) was used by incorporating sampling selection probability into likelihood calculation, but the power was generally low because of the method's relatively large variance of parameter estimates (Monsees *et al.* 2009; Tapsoba *et al.* 2014). The logistic regression (LG) method based on retrospective likelihood conditional on disease status has been applied to improve the power compared to IPW (Lin and Zeng 2009). However, it has a severely inflated type I error rate at stringent significance levels when there is correlation between the two traits because of its assumptions of the distributions of two traits and/or the instable algorithm (Wang and Shete 2011b). Gaussian Copulas method (He *et al.* 2012) was used to model the joint distribution of two traits, and had power similar to that of LG method, and maintained the type I error rate well at a significance level of 0.01. The Wang and Shete's (2011a) bias-correction method to correct odds ratios controls the type I error rate well in secondary binary trait analysis (Wang and Shete 2011a), but cannot be applied to secondary continuous traits. A weighted estimating equation method has also been proposed to handle correlations between two traits, but it needs a bootstrap procedure for hypothesis testing (Song *et al.* 2016). A reparameterized approximate profile likelihood has been developed to correctly estimate the parameter when the interaction between secondary trait and genotype on the primary trait is present (Ghosh *et al.* 2013). However, the performances of all of the methods above are not investigated for the identification of rare variants (RVs), and/or at stringent significance levels. A common theme in all methods above is that the two binary traits are modeled by logistic regression. This type of model, however, frequently has a relatively large variance of the parameter estimators, especially in extreme situations, such as small sample sizes or RVs (Kang *et al.* 2014; Bi *et al.* 2015). Therefore, in developing an efficient method that takes into consideration the statistical design issues for a secondary trait analysis, it is critical to use a more refined statistical model, and a more robust computational algorithm.

Binary traits are often derived from their underlying continuous variables by splitting the range at some thresholds and categorizing individuals above and below that threshold into two separate groups of "affected" and "unaffected." For example, obesity is defined based on BMI, "High" or "Normal" high density lipoprotein cholesterol (HDL-C) is defined based on HDL, and "lower" or "Normal" low density lipoprotein cholesterol (LDL-C) is defined based on LDL (Grundy *et al.* 2004; Kathiresan *et al.* 2008; Loos *et al.* 2008; Willer *et al.* 2008; Teslovich *et al.* 2010). LG-based models above are unable to capture the threshold effect, whereas the set-valued (SV) model is an approach to capturing such effect in the modeling process (Kang *et al.* 2014). The SV model is to model the relationship between independent variables and a set-valued dependent variable that can be generated by a quantization process of the corresponding continuous latent or unknown variable. And the SV model has been employed in genetic association studies for the identification of genetic markers for binary or ordered categorical primary traits, and has better performance than LG-based methods (Kang *et al.* 2014; Bi *et al.* 2015).

Here, we extend the SV approach to secondary traits analyses, in which we use SV model to model the dichotomizing process of the continuous variable for the primary binary trait. The similar dichotomization process will also be used to model the secondary binary trait. An advantage of this method is that it can employ two underlying latent continuous traits from which the binary outcomes are conceptually generated, making it more efficient and robust in estimating the model parameters than the existing methods. We also demonstrate the effectiveness of our method by extensive simulation studies, and an analysis of six continuous and one binary secondary traits in a GWAS of benign ethnic neutropenia/leukopenia (Nalls *et al.* 2008).

Methods

The general idea of our method is to jointly model the primary binary trait and the secondary traits by using SV model. This is based on a dichotomization process of a continuous variable for an observed binary variable that we previously developed for the primary binary traits (Kang *et al.* 2014). Below, we let D denote the case-control disease status ($1 =$ disease/case, $0 =$ no disease/control), Y denote the secondary trait, and X denote the genotype values, coded as 0, 1, or 2 based on the number of minor alleles of a SNP.

SV model for a secondary continuous trait

We propose a novel SV model, in which the primary binary trait (D) can be regarded as the SV observation of a continuous latent variable (D_{lv}):

$$\begin{cases} Y = \beta_0 + \beta_1 X + e, \\ D_{lv} = \gamma_0 + \gamma_1 X + \gamma_2 Y + e_{lv}, \\ D = I_{[D_{lv} > 0]}, \end{cases} \tag{1}$$

where random variables $e(e_{lv})$ follow a normal distribution, with a mean of 0 and variance $\sigma^2(\sigma_{lv}^2)$, and they are independent. $I_{[.]}$ is an indicator function: if $D_{lv} > 0$, then the subject is declared as a case $(D = 1)$, otherwise, it is a control $(D = 0)$. The null hypothesis of H_0: $\beta_1 = 0$ corresponds to no genetic effect of the SNP on the secondary trait. Following model (1), the conditional probability density function of the secondary trait, and the conditional probability of disease status, are

$$\begin{aligned} g(Y|X) &= f_\sigma(Y - \beta_0 - \beta_1 X), \\ P(D = 1|X, Y) &= F_{\sigma_{lv}}(\gamma_0 + \gamma_1 X + \gamma_2 Y), \end{aligned}$$

where $f_\sigma(.)$ and $F_\sigma(.)$ are probability density function and cumulative distribution function of normal distribution with a mean of 0 and a variance of σ^2.

SV model for a secondary binary trait

We propose a novel SV model in which both the primary and secondary binary traits can be regarded as the SV observations of two continuous latent variables:

$$\begin{cases} Y_{lv} = \beta_0 + \beta_1 X + e, \\ Y = I_{[Y_{lv} > 0]}, \\ D_{lv} = \gamma_0 + \gamma_1 X + \gamma_2 Y + e_{lv}, \\ D = I_{[D_{lv} > 0]}, \end{cases} \tag{2}$$

where random variables $e(e_{lv})$ follow a normal distribution, with a mean of 0 and variance $\sigma^2(\sigma_{lv}^2)$, and they are independent, $I_{[.]}$ is an indicator function: Secondary trait $Y = 1$ or 0 depends on whether or not Y_{lv} is >0, and disease status $D = 1$ or 0 depends on whether or not D_{lv} is >0. The null hypothesis of H_0: $\beta_1 = 0$ corresponds to no genetic effect of the SNP on the secondary trait. Following model (2), conditional probabilities of secondary trait and disease status are as follows:

$$\begin{aligned} P(Y = 1|X) &= F_\sigma(\beta_0 + \beta_1 X), \\ P(D = 1|X, Y) &= F_{\sigma_{lv}}(\gamma_0 + \gamma_1 X + \gamma_2 Y). \end{aligned}$$

Parameter estimate and test statistics

Suppose we observe $(D_i, Y_i, X_i),\ i = 1, \ldots, n$, where n is the total sample size. For a case-control study, the sampling process is conditional on the case-control status; hence, we use a retrospective likelihood function as follows:

$$\prod_{i=1}^{N} P(Y_i, X_i|D_i) = \prod_{i=1}^{N} \frac{P(D_i|Y_i, X_i) \cdot P(Y_i|X_i) \cdot P(X_i)}{P(D_i)},$$

where $P(D_i) = \sum_x \sum_y \left[P(D_i|y, x) \cdot P(y|x) \cdot P(x)\right]$.

Here, we assume Hardy-Weinberg equilibrium (HWE), and parameterize the genotype distribution with one parameter, the minor allele frequency (MAF), instead of two parameters $P(X = 0)$, and $P(X = 1)$ as that in Lin and Zeng (2009) (see Supplemental Material, File S1). We maximize this function to estimate model parameter, denoted as $\widehat{\beta_1}$, and compute the derived closed-form Fisher information matrix based on estimated parameters. Then, we can obtain the estimated variance of $(\widehat{\beta_1})$ from the Fisher information matrix. Next, the Wald statistic is constructed to test H_0: $\beta_1 = 0$. Asymptotically, Wald statistic approximately follows a central χ^2 distribution with 1 degree of freedom under the H_0. The detailed implementation of the parameter estimate algorithm, and testing inference procedure, of the SV method is in File S1.

Simulation study

We performed extensive simulation studies to assess the performance of the SV method against the competing alternatives LG-based methods of Lin and Zeng (2009), denoted by LG, and of Ghosh *et al.* (2013), denoted by $\mathrm{LG_{Zou}}$. For $\mathrm{LG_{Zou}}$, if its program gives an error message and stops running, then we will take NA as its output. MAF as an input parameter in $\mathrm{LG_{Zou}}$ will be given by the true MAF for simulations.

Data simulation: Given the MAF of the tested SNP p, we first generated genotypes for a population of 50,000,000 individuals based on the genotype frequencies, $P(X)$, calculated according to HWE, *i.e.*, $P(X = 0) = (1-p)^2$, $P(X = 1) = 2p(1 - p),\ P(X = 2) = p^2$.

Next, we generated traits including disease status and secondary traits for the population. We considered two data simulation models: SV.simu and LG.simu. SV.simu was based on models (1) and (2) for the secondary continuous and binary traits. LG.simu was based on the LG-based model proposed by Lin and Zeng (2009); that is, when the secondary trait was continuous, then the LG.simu model was

$$\begin{cases} Y = \beta_0 + \beta_1 X + e, \\ P(D = 1) = \dfrac{\exp(\gamma_0 + \gamma_1 X + \gamma_2 Y)}{1 + \exp(\gamma_0 + \gamma_1 X + \gamma_2 Y)}. \end{cases} \tag{3}$$

When the secondary trait was binary, then the LG.simu model was

$$\begin{cases} P(Y = 1) = \dfrac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}, \\ P(D = 1) = \dfrac{\exp(\gamma_0 + \gamma_1 X + \gamma_2 Y)}{1 + \exp(\gamma_0 + \gamma_1 X + \gamma_2 Y)}. \end{cases} \tag{4}$$

In both SV.simu and LG.simu, $\beta_0 = 1$ and standard deviation (SD) of e of 1 were used to simulate the secondary continuous trait, which is the same as that used by Lin and Zeng (2009). For the secondary binary trait, we selected $\beta_0 = -1.28(-2.2)$ in LG.simu (SV.simu) so that $P(Y = 1|X = 0) \approx 0.1$.

To simplify our notations, we use OR.D.X to characterize correlation between disease and genotype, and OR.D.Y to characterize correlation between disease and secondary traits. They both are the conditional odds ratios. Their definitions and the one-to-one mapping functions to describe relationships between them and their corresponding parameters (γ_1 and γ_2) in models (3) and (4) can be found in File S2. Hence, given OR.D.X and OR.D.Y, their corresponding parameters γ_1 and γ_2 can be calculated based on their one-to-one mapping functions to simulate the traits data.

After the genotype and traits data were simulated for the population, $n/2$ cases and $n/2$ controls were randomly selected as the sample data for the further secondary trait association analysis.

Type I error simulations: Three values for MAFs of SNPs were considered: 0.005, 0.05, and 0.3. First, we fixed both OR. D.X and OR.D.Y at 1.2, and varied the prevalence of disease from common at 10%, to rare at 5×10^{-5}; when the secondary trait was continuous, the total numbers of cases and controls were $n = 2000$ and $n = 4000$ for CVs with MAF $\geq$ 0.05, and for RVs with MAF $=$ 0.005, respectively. When the secondary trait was binary, the total numbers of cases and controls were $n = 4000$ and $n = 8000$ for CVs with MAF $\geq$ 0.05 and for RVs with MAF = 0.005, respectively. Data were simulated only by LG.simu, since this is enough to show the robustness of the SV method as well as the unstable algorithms implemented in LG and $\mathrm{LG_{Zou}}$. Second, given a disease prevalence of 0.01 and OR.D.X = 1.2, we considered OR.D.Y of 1, 1.2, and 1.5, respectively, representing no correlation, small correlation, and large correlation between disease status D and the secondary trait Y, respectively. The same sample sizes were used as above.

We considered liberal significance levels $\alpha = 0.05, 0.01$, and 0.001, and stringent significance level $\alpha = 5 \times 10^{-4}$, 10^{-4}, and 10^{-5} under $H_0: \ \beta_1 = 0$. 5,000,000 replicated datasets were simulated, and the type I error rate was estimated to be the proportion of replicates with P-values $<\alpha$. For ease of readability, we reported the ratio of empirical estimate of type I error $\hat{\alpha}$ over the expected level of significance, *i.e.*, $R = \hat{\alpha}/\alpha$, for all Tables and Figures reporting type I error results, so that, for a well-controlled test, the ratio should be close to 1.

Power simulations

Extensive simulations were designed to test the power of the three methods. We considered the same parameter settings described in the simulation section for type I error rate, including MAFs, OR.D.X, and the prevalence. We first fixed β_1 at 0.2, 0.5, and 1, and varied the disease prevalence same as above. The other parameters were the same as those for type I error estimation above. Then, given a disease prevalence of 0.01, we considered three situations as (1) we first fixed the same setting of sample size, and increased parameter β_1 from 0.1 to 2 at an increment of 0.1; (2) we fixed $\beta_1 = 1$ and increased the sample size; (3) we fixed the sample size and β_1 but varied the correlations between two traits (OR.D.Y). Data sets were generated 10,000 times for each configuration, and power was estimated to be the proportion of replicates with P-values $<\alpha = 10^{-5}$.

Data availability

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article.

Results

Simulation results: effect of prevalence of diseases

As many GWAS or NGS have been conducted/conducting for both common and rare diseases, it would be of great practical interests to first evaluate the relative performance of the proposed SV method to detect CVs and RVs given different prevalence of diseases, compared to two LG-based methods.

Empirical type I error rate: Figure 1 and Table S1 display the empirical type I errors for three methods given OR.D. X = OR.D.Y = 1.2. Remarkably, SV method robustly and consistently controls type I error rate at any given simulated significance levels to identify both CVs and RVs for binary as well as continuous secondary traits, regardless of whether the disease is as common as 10%, or as rare as 5×10^{-5}. The average values and SD of R at all given significance levels are 0.956 and 0.118, the median is 0.999, and the range is from 0.653 to 1.321.

By contrast, the prevalence of disease has a significant effect on the performance of both LG and $\mathrm{LG_{Zou}}$, especially for identifying RVs at a more stringent significant level. Given a common primary disease with a prevalence of 1% or higher, the more stringent the significance level, the larger the estimated type I error. For example, to identify a SNP with MAF $=$ 0.005 associated with a secondary continuous trait given a disease prevalence of 1%, the type I error rate of LG could be eight times higher than the given level of 0.01, or 5548 times higher than the given level of 10^{-5} due to its unstable estimates of parameters and its variances (Table 2 and Table S2, also see *Variance of the genetic association parameter estimate*). $\mathrm{LG_{Zou}}$ performs better than LG, but still cannot control type I error at stringent significance levels. For identifying CVs, $\mathrm{LG_{Zou}}$ controls type I error rate well for both binary and continuous secondary traits, but LG controlled type I error for identifying a SNP with MAF of 0.3 only when secondary trait is binary (Figure 1, A and B). If the disease is as rare as ≤0.5%, the type I error rates of both methods are generally controlled well, but not controlled for RVs with MAF of 0.005 when secondary trait is binary, especially $\mathrm{LG_{Zou}}$. For example, to identify a RV with MAF $=$ 0.005 associated with a secondary binary trait, the type I error rate of $\mathrm{LG_{Zou}}$ could be 239 times higher than the given level of 10^{-5} due to its unstable estimates of parameters and its variances (Table 2 and Table S2, also see *Variance of the genetic association parameter estimate*).

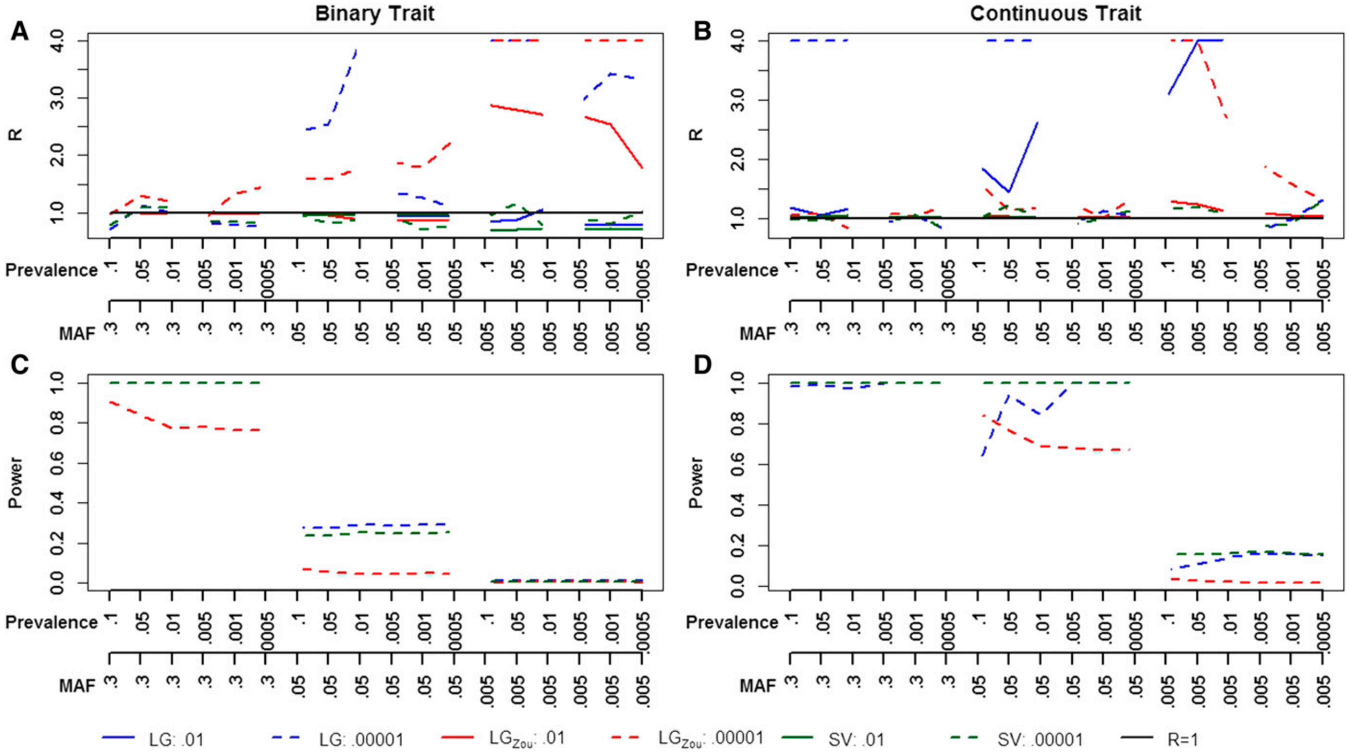

Figure 1 The ratio ($R = \hat{\alpha}/\alpha$) and empirical power for LG, LG_{Zou}, and SV methods. (A) and (B) shows the type I error rates results of three methods for secondary binary and continuous traits, respectively. (C) and (D) shows the empirical power results of three methods for secondary binary and continuous traits, respectively. OR.D.X = OR.D.Y = 1.2 and β_1 = 0.5 for power estimations. The data were generated with LG.simu. The solid and dotted lines correspond to α = 0.01 and 10^{-5}, respectively. The solid black line is for R = 1. If $R \geq 4$, then we set it to be 4. The blue, red, and dark green lines correspond to LG, LG_{Zou}, and SV methods, respectively. The power was estimated at a level of 10^{-5}. When secondary trait is continuous, the sample size of either case or control is $n/2$ = 1000 ($n/2$ = 2000) for CVs (RVs); when the secondary trait is binary, the sample size of either case or control is $n/2$ = 2000 ($n/2$ = 4000) for CVs (RVs).

Empirical power: Strikingly, a similar conclusion holds for power estimation (Figure 1, C and D); that is, there is no obvious effect of the disease prevalence on the power of the SV method because of its quite stable estimates of the parameters and its variance (Table 2 and Table S2, also see section *Variance of the genetic association parameter estimate*).

In contrast, the power of both LG and LG_{Zou} will be significantly affected by the disease prevalence, and is smaller than, or identical to, that of the SV method, especially for identifying a SNP with MAF = 0.05, and a RV with MAF = 0.005 if the disease prevalence is >0.5%. However, we need interpret their power with caution since their type I error rates cannot be controlled in these situations. If the disease is as rare as ≤0.5%, the power of both methods to identify SNPs associated with secondary continuous traits were interpretable, since their type I error rates are controlled, and were smaller than, or identical to, that of SV method; that is, the SV method has highest power followed by LG and LG_{Zou} (Figure 1D). Due to the similar performance of LG and LG_{Zou} above, in Figure 2, Figure 3, Figure 4, Figure S1, Figure S2, Figure S3, Figure S4, Figure S5, Figure S6, Table 1, Table 2, Table S3, Table S4, and Table S5 below, only results corresponding to LG are included. The following sections show results given prevalence of disease of 0.01.

Effect of correlations between two traits

The empirical type I error rate: Table 1 and Table S3 show the empirical type I error rates of the SV and LG methods. The SV method consistently controls type I error rate well at any given significance levels for both CVs and RVs, regardless of whether the secondary trait is continuous or binary, whether the primary and secondary traits are correlated or not, and which simulation model was used to simulate the trait data. As expected, at a liberal significance level of α = 0.05 or 0.01, the LG method correctly maintains type I error control for both CVs and RVs, regardless of which simulation method was used when the secondary trait is binary. Interestingly, when the secondary trait is continuous, the LG method could maintain the type I error at level of 0.05 and 0.01 when MAF is 0.3, but could not when MAF ≤0.05 because of its small estimate of the variance of the estimated parameter (Table 2, also see *Variance of the genetic association parameter estimate*). At stringent significance levels of α = 10^{-4} or 10^{-5}, the LG method could not control the type I error rate in many of the simulated situations, except when the variant is common (MAF = 0.3), and the secondary trait is binary. In addition, as decrease in MAF of the tested SNP, the estimated type I error rate of the LG method increases sharply.

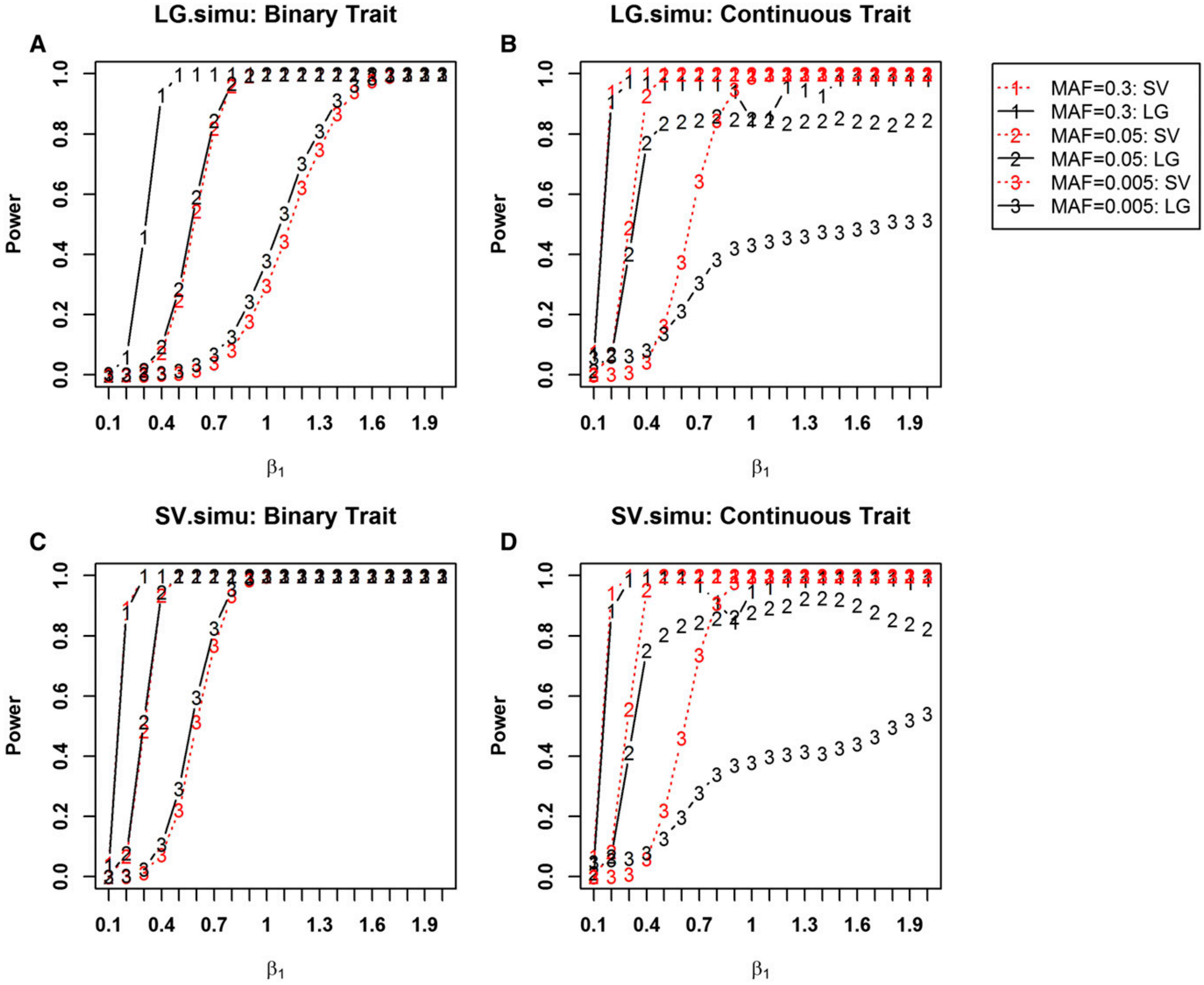

Figure 2 Power of the SV and LG methods as a function of effect size β_1 under an additive genetic model. (A) and (B) show results of a binary and continuous secondary trait with the LG simulation model. (C) and (D) show results of a binary and continuous secondary trait with the SV-based simulation model. Sample sizes in all cases are consistent with those used for simulations to estimate the type I error rates. The solid and dotted lines correspond to the LG and SV methods, respectively. The numbers 1–3 correspond to the tested SNPs with MAFs of 0.3, 0.05, and 0.005, respectively. OR.D.X = OR.D.Y = 1.2, and prevalence is 0.01. When secondary trait is continuous, the sample size of either case or control is $n/2 = 1000$ ($n/2 = 2000$) for CVs (RVs); when the secondary trait is binary, the sample size of either case or control is $n/2 = 2000$ ($n/2 = 4000$) for CVs (RVs).

Empirical power: Figure 2 shows the empirical power of two methods as a function of effect size (β_1) at a significance level $\alpha = 10^{-5}$ for an additive genetic model. As expected, the power of both methods generally increases with increase in effect size, regardless of the type of secondary trait, simulation model or MAFs. When the secondary trait is binary, the SV method has similar power to the LG method for CVs, but has slightly less power for RVs (Figure 2, A and C). This result is consistent with the simulation results of type I error rate, in which the LG method could not control the type I error rate for RVs at a significance level of 10^{-5}. When the secondary trait is continuous, if the effect size is very small, then the LG method has slightly greater power than the SV method, but both have power <0.1. However, as increase in effect size, the power of the SV method sharply increases to 1 with $\beta_1 = 0.5$ for detecting a SNP with a MAF of 0.05, and 0.99 with $\beta_1 = 1$ for detecting a SNP with a MAF of 0.005. However, the power of the LG method gradually increases with corresponding power estimates of 0.84 and 0.43 when the data are simulated by using LG.simu, although the LG method does not correctly control type I error rate at a significance level of 10^{-5}. These facts mean that the respective probabilities of identifying the two variants with their respective MAFs of 0.05 and 0.005 by using the SV method are, respectively, 1.25 times, and more than twice, compared with those of using the LG method. The SV method does maintain a correct type I error rate, but the LG method does not (Figure 2B). As MAF of the tested SNP decreases, the power difference between both methods increases. Furthermore, some interesting cases happen for secondary continuous

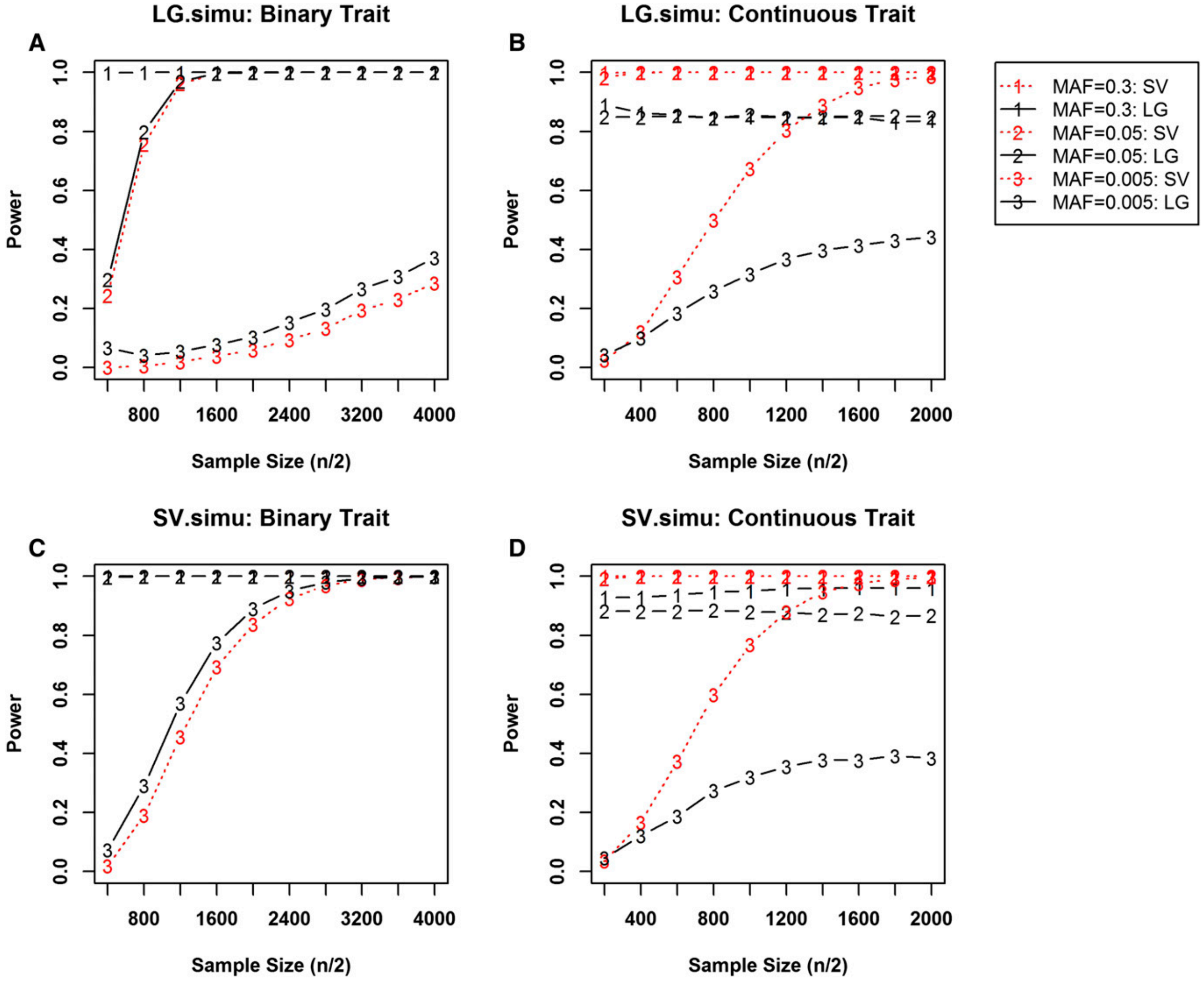

Figure 3 Powers of the SV and LG methods as a function of sample size under an additive genetic model. (A) and (B) show results of a binary and continuous secondary trait with the LG simulation model. (C) and (D) show results of a binary and continuous secondary trait with the SV-based simulation model. The sample size on the *x*-axis is the number of individuals in either cases or controls. The solid and dotted lines correspond to the LG and SV methods, respectively. The numbers 1–3 correspond to the tested SNPs with MAFs of 0.3, 0.05, and 0.005, respectively. OR.D.X = OR.D. Y = 1.2, prevalence is 0.01, and effect size $\beta_1$is fixed at 1.

trait, where the power of the LG method suddenly decreases and then increases as increase in β_1 when MAF is 0.05 and 0.3. We closely examined the results, and found ~20% of simulations in which the LG method wrongly estimated parameter β_1 as being close to 0, decreasing its statistical power (Figure S1). The conclusion is the same when the data are simulated by using SV.simu (Figure 2D).

Figure 3 shows the power of two methods as a function of sample size at a significance level $\alpha = 10^{-5}$ for an additive genetic model. Here, we fixed effect size $\beta_1 = 1$. Regardless of CVs or RVs, data simulation methods and the type of secondary traits, the power of the SV method increases with increase in sample size unless its power is already 1. When the secondary trait is binary, the power of the LG method also increases regardless of the type of variants and data simulation methods: it is almost identical to and slightly greater than that of the SV method when a tested SNP had a MAF ≥0.05 or 0.005, respectively. However, the LG method's type I error could not be controlled when a tested SNP had a MAF of 0.005.

When a secondary trait is continuous, the power of the SV method CV was nearly one under any simulated sample size for CVs, and increased rapidly with increase in sample size for RVs. But the power of the LG method increased slowly. For example, the power of the SV method increases from 0.12 to 0.95 when the total sample size increases from $n = 800$ to $n = 3200$, but that of the LG method increases only from 0.10 to 0.41. These facts mean that, with the supplement of 1200 cases and 1200 controls, SV method can absolutely identify this rare variant with a MAF of 0.005 at a significance level of 10^{-5} when $\beta_1 = 1$, but the probability of identifying this variant by using the LG method is only 0.38. Additionally,

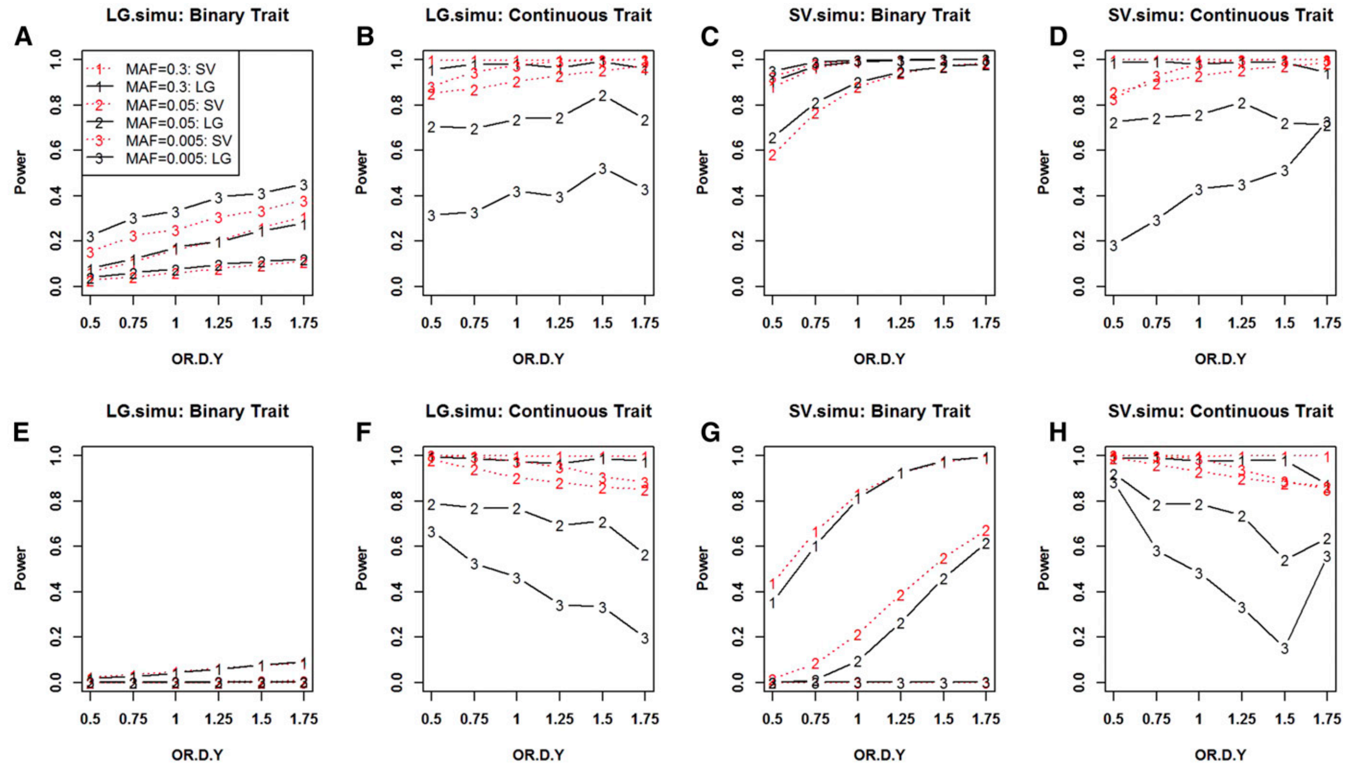

Figure 4 Powers of the SV and LG methods as a function of correlations between two traits under an additive genetic model. (A)/(E) and (B)/(F) show results of a binary and continuous secondary trait with the LG simulation model. (C)/(G) and (D)/(H) show results of a binary and continuous secondary trait with the SV-based simulation model. The solid and dotted lines correspond to the LG and SV methods, respectively. The numbers 1–3 correspond to the tested SNPs with MAFs of 0.3, 0.05, and 0.005, respectively. The prevalence is 0.01, and OR.D.X = 1.2. The effect sizes β_1 are fixed at 1, 0.4 and 0.25 for (A)–(D) and at −1, −0.4, and −0.25 for (E)–(H) for the tested SNPs with MAFs of 0.005, 0.05, and 0.3, respectively. When the secondary trait is continuous, the sample size of either case or control is *n*/2 = 1000 (*n*/2 = 2000) for CVs (RVs); when the secondary trait is binary, the sample size of either case or control is *n*/2 = 2000 (*n*/2 = 4000) for CVs (RVs).

to investigate whether the failure shown in Figure 2 is related to sample size, we used LG.simu to simulate one study with an extremely large sample size of total $n = 20{,}000$ subjects. With a significance level of 10^{-5}, the power of the LG method was still 0.88 for MAF = 0.3, and 0.41 for MAF = 0.05, respectively. And the power of the SV method was one, regardless of rare or CVs. Increasing sample size did not increase power of the LG method, indicating that, under specific parameter settings, the likelihood function of the LG method may have a local maximum point.

Figure 4 and Figure S4 display the power of both methods as a function of the correlations between two traits when OR.D.X = 1.2 or OR.D.X = 0.8. For the secondary continuous trait, if the tested SNP is CV with a MAF of 0.3, then the power of both methods is close to 1 and remains similar regardless of the correlations between two traits or the sign of the effect sizes (*i.e.*, genetic variant is risk or protective for the secondary binary trait). Interestingly, as decrease in MAF of the tested SNP, the power of the SV method to identify a risk (protective) allele for the secondary trait increases (decreases) as the actual value of correlation between two traits increases due to its decreasing (increasing) $\widehat{\mathrm{SE}}(\widehat{\beta_1})$, regardless of the trait simulation methods (Table S5). In detail, as correlation increases, the numbers of patients with heterozygous and homozygous of minor allele increase and decrease when the genetic variant is risk and protective allele for the secondary trait, respectively, which leads to the smaller and larger $\widehat{\mathrm{SE}}(\widehat{\beta_1})$, although $\widehat{\beta_1}$ is stable. By contrast, the power of the LG method is not a clear function of correlations because of its instable estimate of parameter $(\widehat{\beta_1})$ and variance $[\widehat{\mathrm{SE}}(\widehat{\beta_1})]$ (Table S5). The power changes can be very large. For example, as OR.D.Y increases from 0.5, to 1.5, to 1.75, the power of the LG method to identify a protective allele with a MAF of 0.005 are 0.88, to 0.15, to 0.56 (Figure 4H).

For the secondary binary trait, it is obvious, as increase in the actual values of correlations between two traits, the power of both methods increases regardless of CVs or RVs, trait simulation models, and the sign of the effect sizes (*i.e.*, genetic variant is risk or protective for the secondary binary trait) (Figure 4, A, C, E, and G). In detail, as the actual correlation values between two binary traits increase, the distribution of secondary binary trait conditional on the fixed primary trait becomes more balanced, so that $\widehat{\mathrm{SE}}(\widehat{\beta_1})$ becomes smaller, though $\widehat{\beta_1}$ stays very similar, which leads to greater power of both methods (Table S5, see also *Variance of the genetic association parameter estimate* below). The increase in power can be very large when the increase in the actual value of the correlation is very large. For example, as

Table 1 A comparison of the ratios ($R = \hat{\alpha}/\alpha$) for the SV and LG methods

		α = 0.01						α = 1e−5					
		OR.D.Y = 1		OR.D.Y = 1.2		OR.D.Y = 1.5		OR.D.Y = 1		OR.D.Y = 1.2		OR.D.Y = 1.5	
Simulation Model	MAF	SV	LG	SV	LG	SV	LG	SV	LG	SV	LG	SV	LG
Continuous													
LG.simu	0.3	1.00	1.22	1.02	1.18	1.11	1.13	1.12	**139.0**	1.20	**132.76**	1.32	**81.33**
	0.05	1.00	**3.28**	1.02	**2.75**	1.06	**2.99**	1.12	**1604.2**	1.26	**1267.04**	1.02	**1429.1**
	0.005	1.00	**7.25**	1.01	**8.13**	1.03	**9.27**	1.29	**4977.1**	1.02	**5591.08**	1.13	**6486.1**
SV.simu	0.3	1.00	1.26	1.00	1.40	1.01	1.53	1.10	**157.96**	1.14	**225.99**	1.20	**63.92**
	0.05	1.01	**3.25**	1.00	**2.75**	1.01	**3.66**	1.00	**1573.3**	1.30	**1342.06**	0.88	**1978.9**
	0.005	1.00	**7.49**	1.00	**7.34**	1.00	**8.83**	0.91	**5158.2**	0.89	**5015.27**	0.98	**6150.4**
Binary													
LG.simu	0.3	1.00	0.99	1.00	0.99	1.01	0.98	0.94	1.02	0.82	0.88	1.02	0.82
	0.05	0.95	0.95	0.98	0.97	0.98	0.96	0.70	**12.75**	1.02	**9.23**	0.74	**6.53**
	0.005	0.65	1.11	0.71	1.08	0.78	1.08	0.82	**327.1**	1.00	**316.26**	0.90	**311.3**
SV.simu	0.3	0.99	0.98	0.99	1.00	0.99	1.13	1.06	0.92	0.88	0.86	0.82	1.08
	0.05	0.96	0.95	0.96	0.92	0.97	0.92	0.92	**1.68**	0.80	1.34	0.84	0.98
	0.005	0.70	0.94	0.76	0.86	0.82	0.87	0.88	**156.8**	0.58	**144.01**	0.58	**150.1**

$\beta_1 = 0$, OR.D.X was fixed at 1.2; prevalence was fixed at 0.01; when the secondary trait was continuous, $\beta_0 = 1$, and the sample size of either case or control is 1000 (2000) for CVs (RVs); when the secondary trait is binary, then $\beta_0 = -2.2$ (-1.28), if the simulation model is the LG-based (SV) model, and sample size of either case or control is 2000 (4000) for CVs (RVs). Bold type indicates that the type I error rate could not be controlled.

OR.D.Y increases from 0.5 to 1.75, the power of the SV method to identify a protective allele with MAF 0.05 at a significance level of 10^{-5} increases from 0.01 to 0.67 (Figure 4G), and that to identify a protective allele with MAF 0.005 at a significance level of 0.01 increases from 0 to 0.99 (Figure S3C). The conclusions of the power as a function of correlations between two traits when the OR.D.X = 0.8 are similar (Figure S4).

All the conclusions above hold for dominant and recessive genetic disease models (Figure S5 and Figure S6).

Variance of the genetic association parameter estimate

Effect of the prevalence of disease: No matter the disease is as common as a prevalence of 10%, or as rare as a prevalence of 5×10^{-4}, if OR.D.Y = 1.2, the estimate of the genetic association parameter by the SV method was robust, and very close to the true parameters if the trait is continuous, and was nearly a fixed times smaller than the true parameters if the trait is binary; the SD of the estimate parameter [$\mathrm{SD}(\widehat{\beta_1})$] was very close to the mean of the estimated standard error $\overline{\widehat{\mathrm{SE}}(\widehat{\beta_1})}$ if the data were simulated by LG.simu. This leads to the profound and remarkable properties of the SV method in terms of controlled type I error rate and improved power compared to the LG and $\mathrm{LG_{Zou}}$ methods (Table S2).

Similarly, if the disease is rare, such as a prevalence varying from 0.005 to 5×10^{-4}, to identify SNPs associated with a continuous secondary trait, the conclusions for LG and $\mathrm{LG_{Zou}}$ methods were similar to that of SV, but $\mathrm{LG_{Zou}}$ had a greater $\mathrm{SD}(\widehat{\beta_1})$ than SV and LG, which led to the conservative properties of $\mathrm{LG_{Zou}}$, regardless of whether the SNP is associated with secondary trait or not (Figure 1B and Table S2). In sharp contrast, to identify SNPs associated with a binary secondary trait, $\mathrm{LG_{Zou}}$ had a bias estimate of the parameter, and smaller estimate of the variance of parameter estimate, for a SNP with MAF of 0.005, which leads to the uncontrolled type I error rate for rare SNPs (Figure 1A and Table S2).

By remarkable contrast, if the disease is common, such as a prevalence varying from 1 to 10%, to identify common or rare SNPs associated with a continuous secondary trait, LG had smaller estimate of the variance of the parameter estimate [$\mathrm{SD}(\widehat{\beta_1}) > \overline{\widehat{\mathrm{SE}}(\widehat{\beta_1})}$] so that its type I error rate could not be controlled at stringent significant levels, but its power was comparable to that of SV method (Figure 1B and Table S2). However, $\mathrm{LG_{Zou}}$ performed better than LG for common and rare SNPs, but still had an inflated type I error for rare SNPs in some situations. Furthermore, to identify SNPs associated with secondary binary traits, both methods performed similarly, that is, for rare SNPs, they both had smaller estimate of the variance of the parameter estimate, which leads to uncontrolled type I error rates (Figure 1A and Table S2).

Effect of correlations between two traits: Given a disease prevalence of 0.01, Table 2 gives a mean of $\widehat{\beta_1}$, a mean of the estimated SEs of $\widehat{\beta_1}$ [denoted as $\overline{\widehat{\mathrm{SE}}(\widehat{\beta_1})}$], and SD of $\widehat{\beta_1}$ [denoted as $\mathrm{SD}(\widehat{\beta_1})$] for the LG and SV methods based on 10,000 simulation repetitions under the null hypothesis $\beta_1 = 0$. As expected, as MAF decreased, both $\overline{\widehat{\mathrm{SE}}(\widehat{\beta_1})}$ and $\mathrm{SD}(\widehat{\beta_1})$ increased for both LG and SV methods. And $\overline{\widehat{\mathrm{SE}}(\widehat{\beta_1})}$ appeared to be close to $\mathrm{SD}(\widehat{\beta_1})$ for the SV method in all simulation setups, leading to its correct control of the type I error rate (Table 2). However, the SD of estimated parameter [$\mathrm{SD}(\widehat{\beta_1})$] for the LG method was obviously larger than the mean of the estimated standard error $\overline{\widehat{\mathrm{SE}}(\widehat{\beta_1})}$, especially for RVs and secondary continuous traits, leading to its severely

Table 2 The mean of $\widehat{\beta_1}$, mean of estimated SE of $\widehat{\beta_1}$, and SD of $\widehat{\beta_1}$ for the SV and LG methods under the null hypothesis ($\beta_1 = 0$) based on 10,000 simulations

	OR.D.Y = 1										OR.D.Y = 1.5									
	SV					LG					SV					LG				
MAF[a]	$\lvert\overline{\widehat{\beta_1}}\rvert$	$sd(\widehat{\beta_1})$	$\overline{\widehat{se}(\widehat{\beta_1})}$	$\frac{\lvert\overline{\widehat{\beta_1}}\rvert}{sd(\widehat{\beta_1})}$	$\frac{\lvert\overline{\widehat{\beta_1}}\rvert}{\overline{\widehat{se}(\widehat{\beta_1})}}$	$\lvert\overline{\widehat{\beta_1}}\rvert$	$sd(\widehat{\beta_1})$	$\overline{\widehat{se}(\widehat{\beta_1})}$	$\frac{\lvert\overline{\widehat{\beta_1}}\rvert}{sd(\widehat{\beta_1})}$	$\frac{\lvert\overline{\widehat{\beta_1}}\rvert}{\overline{\widehat{se}(\widehat{\beta_1})}}$	$\lvert\overline{\widehat{\beta_1}}\rvert$	$sd(\widehat{\beta_1})$	$\overline{\widehat{se}(\widehat{\beta_1})}$	$\frac{\lvert\overline{\widehat{\beta_1}}\rvert}{sd(\widehat{\beta_1})}$	$\frac{\lvert\overline{\widehat{\beta_1}}\rvert}{\overline{\widehat{se}(\widehat{\beta_1})}}$	$\lvert\overline{\widehat{\beta_1}}\rvert$	$sd(\widehat{\beta_1})$	$\overline{\widehat{se}(\widehat{\beta_1})}$	$\frac{\lvert\overline{\widehat{\beta_1}}\rvert}{sd(\widehat{\beta_1})}$	$\frac{\lvert\overline{\widehat{\beta_1}}\rvert}{\overline{\widehat{se}(\widehat{\beta_1})}}$
Continuous secondary trait simulated from LG.simu																				
0.3	0.000	0.034	0.034	0.005	0.005	0.000	0.042	0.037	0.008	0.009	0.005	0.034	0.034	0.135	0.136	0.000	0.060	0.034	0.007	0.012
0.05	0.002	0.071	0.070	0.026	0.027	0.004	0.363	0.110	0.010	0.033	0.005	0.070	0.069	0.074	0.074	0.012	0.320	0.129	0.037	0.092
0.005	0.001	0.153	0.153	0.009	0.009	0.046	1.436	0.475	0.032	0.097	0.007	0.153	0.152	0.047	0.047	0.087	1.414	0.669	0.062	0.130
Continuous secondary trait simulated from SV.simu																				
0.3	0.000	0.033	0.034	0.014	0.014	0.000	0.041	0.036	0.004	0.005	0.000	0.033	0.033	0.005	0.005	0.010	0.109	0.037	0.088	0.257
0.05	0.000	0.068	0.068	0.004	0.004	0.003	0.315	0.106	0.010	0.028	0.001	0.067	0.067	0.015	0.015	0.012	0.342	0.084	0.036	0.147
0.005	0.001	0.148	0.148	0.004	0.004	0.002	1.464	0.448	0.001	0.004	0.001	0.147	0.146	0.010	0.010	0.031	1.124	0.325	0.027	0.095
Binary secondary trait simulated from LG.simu																				
0.3	0.001	0.041	0.041	0.018	0.018	0.002	0.080	0.080	0.021	0.021	0.003	0.038	0.038	0.068	0.067	0.000	0.074	0.074	0.003	0.003
0.05	0.005	0.086	0.085	0.064	0.064	0.012	0.168	0.166	0.074	0.075	0.002	0.077	0.078	0.030	0.030	0.011	0.152	0.152	0.075	0.075
0.005	0.025	0.197	0.189	0.125	0.131	0.062	0.451	2.195[b]	0.137	0.028	0.013	0.180	0.172	0.073	0.076	0.040	0.359	0.338	0.112	0.119
Binary secondary trait simulated from SV.simu																				
0.3	0.000	0.041	0.041	0.006	0.006	0.001	0.080	0.079	0.009	0.010	0.000	0.035	0.035	0.003	0.003	0.015	0.069	0.069	0.219	0.218
0.05	0.003	0.083	0.082	0.042	0.042	0.008	0.162	0.161	0.052	0.053	0.003	0.072	0.071	0.037	0.037	0.022	0.141	0.139	0.156	0.158
0.005	0.017	0.189	0.182	0.090	0.094	0.044	0.378	0.359	0.116	0.122	0.011	0.157	0.156	0.069	0.070	0.045	0.312	0.307	0.143	0.145

[a] OR.D.X was fixed at 1.2; prevalence was fixed at 0.01; when the secondary trait is continuous, then $\beta_0 = 1$, and the sample size of either case or control is $n/2 = 1000$ ($n/2 = 2000$) for CVs (RVs); when the secondary trait is binary, then $\beta_0 = -2.2$ (-1.28) if the simulation model is the LG-based (SV) model, and the sample size of either case or control is $n/2 = 2000$ ($n/2 = 4000$) for CVs (RVs).

[b] Out of 10,000 simulation replicates, one replicate had an estimated SE of $\widehat{\beta_1}$ 1.8×10^4, which makes $\overline{\widehat{SE}(\widehat{\beta_1})}$ large.

inflated type I error especially at stringent significance level. Additionally, the mean of $\widehat{\beta_1}$ by the SV method is much closer to the true value $\beta_1 = 0$, compared with that by the LG method. For example, when the secondary trait was continuous, and LG.simu was employed to simulate RVs (MAF = 0.005), the SV method gave almost the same value of 0.153 for both $\overline{\widehat{\mathrm{SE}}(\widehat{\beta_1})}$ and $\mathrm{SD}(\widehat{\beta_1})$, while the LG method gave $\overline{\widehat{\mathrm{SE}}(\widehat{\beta_1})}$ of 0.475 and $\mathrm{SD}(\widehat{\beta_1})$ of 1.436. The mean of $\widehat{\beta_1}$ by the SV method was 0.001, which is <<0.046 by the LG method. The conclusions are also validated in (A)–(C) of Figure S1 and Figure S2, the scatter plot for both $\widehat{\beta_1}$ and $\widehat{\mathrm{SE}}(\widehat{\beta_1})$.

We also recorded summary results for parameter estimations under the alternative hypothesis in Table S4. Similar to those for the null hypothesis, the SV method's estimate of the parameter, and its estimated variance under the alternative hypothesis, were very stable. The mean of the estimated parameter was close to the true value, and the SD of the estimated parameter $[\mathrm{SD}(\widehat{\beta_1})]$ was very close to the mean of the estimated SEs of the estimated parameter $\overline{\widehat{\mathrm{SE}}(\widehat{\beta_1})}$) regardless of the type of variants (common or rare variant) and trait simulation method, showing that the power of the SV method stay stable and optimal.

By contrast, under the LG method's own trait simulation model, the mean of the estimate of parameter by the LG method for the CVs with MAFs of 0.3 was close to the true value for both secondary binary and continuous traits, except $\beta_1 = 1$. Similarly, the SD of the estimated parameter $[\mathrm{SD}(\widehat{\beta_1})]$ was close to the mean of the estimated SEs of the estimated parameter $\overline{\widehat{\mathrm{SE}}(\widehat{\beta_1})}$. Both findings demonstrate that the power of the LG method for CVs will be comparable to that of the SV method. However, the power of the LG method at a stringent significance level should be interpreted with caution because of the method's inflated type I error rate. Interestingly, when the secondary trait was continuous, MAF was 0.3, and true $\beta_1 = 1$, the mean value of $\widehat{\beta_1}$ by LG method was 0.851 (<<1), which indicates a number of inaccurate smaller estimates of parameter β_1. (D1) in Figure S1 shows that, in this particular parameter setup, there were a lot of repetitions whose $\widehat{\beta_1}$ close to 0, further explaining the decrease of power in Figure 2B. As the MAF of the tested SNP decreased, its mean of the estimate of parameter ($\left|\overline{\widehat{\beta_1}}\right|$) was much smaller than the true value, especially for secondary continuous traits, and the SD of the estimated parameter $[\mathrm{SD}(\widehat{\beta_1})]$ was larger than the mean of the estimated SEs of the estimated parameter $\overline{\widehat{\mathrm{SE}}(\widehat{\beta_1})}$). This situation leads to the method's power being less than that of the SV method. For example, when the trait is simulated by using LG.simu, with $\beta_1 = 0.8$, a MAF of 0.005, and 1000 cases and 1000 controls, then the respective estimate of β_1 by the SV and LG methods were 0.804 and 0.371, for a continuous trait, with their

$\mathrm{SD}(\widehat{\beta_1})$ of 0.148 and 1.372, which clearly shows the profound power advantage of the SV method compared to the LG method. Additionally, for the secondary binary trait, we found that the LG and SV methods estimated parameters more accurately when data were simulated with a model of its own, and that the ratio of parameter estimation was similar to that of the primary binary trait analysis by logistic and probit regression.

Interestingly, for secondary continuous traits, regardless of the trait simulation models, as the correlations between two traits increase, the estimate of the parameter by the SV method to identify a risk (protective) genetic variant remains very stable, but $\mathrm{SD}(\widehat{\beta_1})$ and $\overline{\widehat{\mathrm{SE}}(\widehat{\beta_1})}$ decrease (increase), because the number of individuals with heterozygous and/or homozygous genotypes of the risk allele increase (decrease), especially for the RVs (Table S5), which proved the power increase (decrease) as increase in correlations between two traits (Figure 4, B, D, F, and H). For the LG method, the trend of the power as a function of the correlations is not very clear because of its unstable estimate of the parameter, and its estimated SE. In contrast, for secondary binary traits, as the actual value of the correlations between two trait increases, the estimate of the parameter by both methods remains very stable, but $\mathrm{SD}(\widehat{\beta_1})$ and $\overline{\widehat{\mathrm{SE}}(\widehat{\beta_1})}$ decrease, regardless of whether the genetic variant is a risk or protective factor, in that the number of individuals with $D = 1$ and $Y = 1$ increases, so that $\widehat{\mathrm{SE}}(\widehat{\beta_1})$ decreases, leading to the increasing power of both methods.

Application to a GWAS of benign ethnic neutropenia

Benign ethnic neutropenia (BEN) is a clinical condition more commonly observed in African-Americans, with a prevalence of ~4.4% (Hsieh *et al.* 2007). The NHLBI and National Institute of Diabetes and Digestive and Kidney Disease conducted a study to identify genetic determinants of BEN. The study used DNA samples and phenotypes available from The REasons for Geographic and Racial Differences in Stroke study. Subjects with leukocyte counts in the lowest 1–7th percentile were considered as cases, and subjects with leukocyte counts in the highest 85th–95th percentile were considered as controls. After removing 10 patients without genotype data, we analyzed 984 genetically independent subjects (489 cases and 495 controls) with the genotype of 677,755 SNPs (dbGaP Study Accession: phg000307.v1) provided by the BEN study that passed preimputation filters (IMPUTE2 SNP types 2 and 3). Quality control of the genotypic and phenotypic data was performed by the study team of this study. We further removed SNPs whose MAFs were <0.005, to enable us to investigate the RV associations. In our analysis, we dealt with leukocyte counts case-control status as the primary binary trait, and analyzed seven secondary traits, including one binary trait of stroke, and six continuous traits of the total cholesterol (TC), HDL, LDL, platelet count (PC), triglycerides, and C-reactive protein (CRP), using the three methods above. Depending on the normality of the data, either a two-sample *t*-test or a Mann-Whitney-Wilcoxon test was used to test for correlations between the primary and secondary traits. HDL ($P_{\mathrm{cor}} = 7.2 \times 10^{-10}$), PC ($P_{\mathrm{cor}} = 3.6 \times 10^{-26}$), triglycerides ($P_{\mathrm{cor}} = 1.5 \times 10^{-24}$), and CRP ($P_{\mathrm{cor}} = 3.0 \times 10^{-32}$) were significantly correlated with the continuous white blood cell count, and the binary case-control status (Figure S7). The distribution of age and gender in cases and controls were comparable. Before analysis, we followed the same normalization process as He *et al.* (2012), and standardized each continuous trait by subtracting its mean estimated from controls, and then dividing by its SD estimated by using all samples.

Table 3 summarizes the number of SNPs with no *P*-value outputs, *P*-value of 0, or *P*-value $<10^{-7}$ or 10^{-5} for each secondary trait. For the six continuous traits, the number of SNPs identified by the LG method was much larger than that identified by the SV method with $\mathrm{LG_{Zou}}$ in-between. Because of too many extremely small *P*-values, the QQ-plot of the LG method deviated clearly from the expected line, with respective genomic inflation factors (computed by function estlambda in R package GenABEL with default arguments; 10^{-100} are used to appropriate *P* values of 0 in case of the errors of the regression algorithm) of 1.80, 4.91, 2.32, 1.73, 1.80, and 2.15 for TC, HDL, LDL, PC, triglycerides, and CRP, respectively (Figure S8). By sharp contrast, the SV method is more stable when analyzing these six secondary continuous traits, yielding a QQ-plot very close to the diagonal line, with corresponding genomic inflation factors of 0.97, 1.05, 0.97, 0.93, 0.94, and 0.92 (Figure S8). Similarly, $\mathrm{LG_{Zou}}$ has genomic inflation factors greater than, but close to, 1 (1.16, 1, 1.16, 1.27, and 1.52 for TC, HDL, LDL, PC, and triglycerides, respectively), except 2.65 for analyzing CRP. All six genomic inflation factors are much >1 by the LG method, and very close to 1 by the SV method (Figure S8), showing that there is no population stratification problem in this data. Therefore, they indicate that the LG method might have generated too many false positive results based on our simulation results above for common diseases (please also see below for justification). Additionally, the number of SNPs without a *P*-value generated by the LG method is also much larger than that generated by the SV and $\mathrm{LG_{Zou}}$ methods. For instance, when analyzing the secondary trait of platelet count, there are 147,358 SNPs with no *P*-value output by the LG method, which is 10,526 and 5,668 times more than that found by the SV and $\mathrm{LG_{Zou}}$ methods, respectively. Furthermore, SV did not generate any SNP with a *P*-value of 0 for seven traits, and $\mathrm{LG_{Zou}}$ generated 13 SNPs with a *P*-value of 0 when analyzing CRP; in striking comparison, LG generated hundreds of SNPs with a *P*-value of 0. We would definitely suggest that the LG method not be used for mapping RVs when the secondary trait is continuous. For the binary trait of stroke, the three methods perform similarly. Their genomic inflation factors are 0.978, 0.970, and 1.12 for the LG, $\mathrm{LG_{Zou}}$, and SV methods, respectively. Under a

Table 3 Summary information of seven secondary traits genetic association analyses in GWAS of BEN with a prevalence of 0.044

	P = NA					P = 0					$P = 10^{-7}$					$P = 10^{-5}$				
Traits[a]	SV	LG	LG_{Zou}	Both	Three	SV	LG	LG_{Zou}	Both	Three	SV	LG	LG_{Zou}	Both	Three	SV	LG	LG_{Zou}	Both	Three
PC ($P_{cor} = 3.6 \times 10^{-26}$)[b]	14	147,358	26	24	9	0	146	0	0	0	0	278	6	0	0	5	368	98	5	3
TL ($P_{cor} = 1.5 \times 10^{-24}$)	732	39,255	10	6	4	0	215	0	0	0	0	418	2	0	0	6	564	48	1	1
CRP ($P_{cor} = 3.0 \times 10^{-32}$)	9	4,994	45	0	0	0	322	13	0	0	3	570	1465	12	3	31	793	5064	76	28
HDL ($P_{cor} = 7.2 \times 10^{-10}$)	86	21,675	17	1	0	0	1378	0	0	0	0	2716	2	0	0	11	3298	10	2	0
LDL ($P_{cor} = 0.24$)	66	12,742	8	1	1	0	402	0	0	0	1	763	5	1	1	11	986	48	2	1
TC ($P_{cor} = 0.48$)	21	6,486	28	9	4	0	209	0	0	0	0	477	54	0	0	9	690	451	8	1
Stroke ($P_{cor} = 0.42$)	4062	1,924	17,314	505	491	0	1	2	0	0	0	3	2	0	0	5	7	23	0	0
Total	4990	234,434	17,448	546	509	0	2673	15	0	0	4	5225	1536	13	4	78	6706	5742	94	34

[a] TC, total cholesterol; HDL, high-density lipoprotein; LDL, low-density lipoprotein; PC, platelet count; TL, triglycerides; CRP, C - reactive protein; Both, SNPs identified by the LG-based methods: both LG and LG_{Zou} methods. Three, SNPs identified by three methods.

[b] P_{cor} are the correlation/association test P-values based on two sample t-test or Mann–Whitney-Wilcoxon test for continuous secondary traits, and based on chi-square test for stroke.

significance level of 10^{-5}, SV, LG, and LG_{Zou} identified 5, 7, and 23 SNPs with P-value $<10^{-5}$, but no single SNP was identified by all of these three methods. Of the five SNP identified by SV, two, with MAFs of 0.35 and 0.06, were also identified by LG, and the P-values of the other three identified by LG are slightly higher than those identified by SV. One P-value by LG_{Zou} was NA, and P-values of the other four were close to 10^{-4} (Table S6).

At a significance level of 10^{-7}, the SV method identified three SNPs (*rs856046*, *rs12100053*, and *rs11466310*) associated with CRP, and one SNP (*rs7412*) associated with LDL. All four SNPs were also identified by the LG and LG_{Zou} method, which means that 100% of SNPs identified by SV can be replicated by LG and LG_{Zou}. Similarly, LG and LG_{Zou} identified, in total, 5225 and 1536 SNPs for seven traits, but only 0.08 and 0.03% were replicated by three methods. Even for both LG-based methods, among 5225 and 1536 SNPs identified by LG and LG_{Zou}, 0.25 and 0.8% were replicated by LG_{Zou} and LG, respectively. For four SNPs identified by the three methods at a level of 10^{-7} above, SNP *rs7412* is a nonsynonymous SNP in *APOE* exon 4, whose corresponding protein is the principle cholesterol carrier in the brain, and the corresponding association with LDL has been widely reported (Thompson *et al.* 2005; Liu *et al.* 2013). SNP *rs856046*, whose association with CRP has been reported by Reiner *et al.* (2012), is located in gene *IFI16*. SNPs *rs11466310* and *rs12100053* are intron variants locating in genes *B9D2* and *TRIM13*, respectively. However, no associations between these two genes and CRP have been reported in the literature, and thus they need further validation in another independent study.

Similarly, at a significance level of 10^{-5}, the SV method identified, in total, 78 SNPs associated with seven secondary traits, among which 44% (34) were replicated by both LG and LG_{Zou} methods. However, LG and LG_{Zou} identified, in total, 6706 and 5742 SNPs with P-values $<10^{-5}$, among which 0.5 and 0.6% were replicated by SV and LG_{Zou}, and SV and LG, respectively. In addition, for the two LG-based methods, among 6706 SNPs identified by LG, and 5742 SNPs identified by LG_{Zou}, only 1.4 and 1.6% were replicated by the other LG-based method (Table 3 and detailed information in Table S6). These data indicate that LG-based methods generated more nonreplicable or nonreproducible results, and are sufficient to show that our SV method is a more robust, efficient, and reliable statistical method compared to the LG-based methods of LG and LG_{Zou}.

In addition, we also found that the assignment of different prevalence changes the results of the LG and LG_{Zou} methods greatly, but does not affect the SV method (data not shown), which is quite consistent with the simulation results above. The stronger the correlations between the primary and secondary traits, the more SNPs without a P-value, or with P-value of 0, are generated by the LG-based methods. However, there is no similar trend for the SV method. Certainly, some of the secondary trait association analyses results above may be affected by some confounding factors, such as gender, but our focus in this study is to conduct secondary trait association analysis without considering the covariates.

Discussion

We have proposed a novel statistical SV approach to identify CVs and RVs associated with secondary binary or continuous traits in a case-control study design. This method is much more reliable, robust, and efficient than the LG-based methods for many critical factors, including the MAF of markers (RVs or CVs), different link functions, prevalence of the primary disease (common or rare diseases), correlations between primary and secondary traits, type of secondary traits, type of associations between genetic variants and the secondary trait (risk or protective), and type of associations between genetic variants and primary trait (risk or protective). This method also has greater power in different genetic disease models while maintaining the type I error rate than do the LG-based methods, especially for evaluating the secondary continuous trait and RVs. Of the markers identified by the SV method at significance levels of 10^{-7} and 10^{-5}, 100% and ∼50% can be reproducible by the other LG-based methods. In remarkable contrast, <1% of markers identified by LG-based methods can be reproduced by the SV method, and <1% of markers identified by one LG-based method can be reproduced by the other LG-based method. Because of its striking and profound advantages, we strongly recommend the proposed new SV method be employed instead of the LG-based methods for secondary binary and continuous traits analyses in case-control sequencing studies.

We have focused on secondary binary and continuous traits for case-control studies. In some studies such as NHLBI ESP, the early-onset myocardial infarction case-control study (Do *et al.* 2015), investigators can be interested in the secondary traits of ordered categorical traits, such as categorized BMI (under-normal, normal, overweight, and obesity), and that of longitudinal traits, such as diastolic and systolic blood pressure, and LDL. We are extending our SV method to such traits in a case-control study. An ordered categorical trait can be analyzed as a binary trait, but doing so significantly affects the statistical power (Bi *et al.* 2015). Furthermore, in the current SV method, we do not adjust for some confounding covariates, such as genetic ancestry scores, which are commonly adjusted for in genetic association studies. Similarly, we also do not adjust for interactions between secondary traits and genetic marker on primary disease risk (Ghosh *et al.* 2013). We are extending our SV method to incorporate covariates, and/or interactions between genetic marker and secondary traits on the primary disease risk, into the model. As noted, this paper only models SNP-trait associations; however, the model can be easily extended to any biologically meaningful mutants, such as multi-allelic locus, copy number variants, and haplotype.

Besides the case-control study design, there are some other trait-dependent sampling designs, such as extreme phenotype sampling of a continuous primary trait, that are commonly used in NHLBI ESP, in which individuals with values of a continuous trait larger than a threshold c_1, and smaller than another threshold c_2, are selected for sequencing/genotyping. For example, in the NHLBI ESP BMI study, 267 individuals with BMI $> c_1 = 40$, and 178 individuals with BMI $< c_2 = 25$ were selected for sequencing out of 11,468 individuals from the Women's Health Initiative. The method proposed in this study can be readily applied to analyze the secondary binary and continuous traits by using a common threshold such as $(c_1+c_2)/2$; however, because doing so does not account for extremes appropriately, the results might not be valid. There is one existing statistical method available to adjust for the extreme phenotype sample design, but the method is targeted for combining the primary phenotype in one study, and the same secondary phenotypes under some sampling-dependent study designs to conduct SNP- or gene-based association tests, but it can be applied only to secondary quantitative traits but not qualitative traits (Lin *et al.* 2013). Thus, a valid uniform statistical method for RV association test in extreme phenotype sequencing design is urgently needed.

We have implemented the proposed new SV method in an R package, $\mathrm{SV}_{2\mathrm{bc}}$, which is available for free download from http://www.stjuderesearch.org/site/depts/biostats/software. The method can be easily applied to analyze secondary binary and continuous traits in a case-control study of candidate gene association analysis, GWAS, or NGS studies.

Acknowledgments

We greatly appreciate the editor and two anonymous reviewers for their insightful comments, which have improved our manuscript significantly. We thank Dr. Ying Yuan for discussions during the initiation stage of this project, and Cherise M. Guess for editing the manuscript. We thank Dr. Arpita Ghosh for kindly sharing the R program. We acknowledge dbGAP for approval of our use of benign ethnic neutropenia data. The data were obtained from Matthew Hsieh's ancillary proposal to the Reasons of Geographic and Racial Differences in Stroke study; Matthew Hsieh is supported by the intramural research program of NHLBI and NIDDK at the National Institutes of Health (NIH). Genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268200782096C and HHSN268201100011I. We acknowledge the High Performance Computing Facility (HPCF) at St. Jude Children's Research Hospital (SJCRH) for providing shared high performance computing resources that have contributed to the research results reported within this paper. This research is supported by the American Lebanese and Syrian Associated Charities, and grants from National Natural Science Foundation of China (11171333, 61134013, 81571313, and 81221002).

Literature Cited

Bi, W., G. Kang, Y. Zhao, Y. Cui, S. Yan *et al.*, 2015 SVSI: fast and powerful set-valued system identification approach to identifying rare variants in sequencing studies for ordered categorical traits. Ann. Hum. Genet. 79(4): 294–309.

Do, R., N. O. Stitziel, H. H. Won, A. B. Jørgensen, S. Duga *et al.*, 2015 Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. Nature 518 (7537): 102–106.

Edmondson, A. C., P. S. Braund, I. M. Stylianou, A. V. Khera, C. P. Nelson *et al.*, 2011 Dense genotyping of candidate gene loci identifies variants associated with high-density lipoprotein cholesterol. Circ. Cardiovasc. Genet. 4: 145–155.

Frayling, T. M., N. J. Timpson, M. N. Weedon, E. Zeggini, R. M. Freathy *et al.*, 2007 A common variant in the *fto* gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316: 889–894.

Ghosh, A., F. A. Wright, and F. Zou, 2013 Unified analysis of secondary traits in case-control association studies. J. Am. Stat. Assoc. 108(502): 566–576.

Grundy, S. M., H. B. Brewer, Jr., J. I. Cleeman, S. C. Smith, Jr., C. Lenfant *et al.*, 2004 Definition of metabolic syndrome: report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition. Circulation 109: 433–438.

He, J., H. Li, A. C. Edmondson, D. J. Rader, and M. Li, 2012 A Gaussian copula approach for the analysis of secondary traits in case-control genetic association studies. Biostatistics 13(3): 497–508.

Hsieh, M. H., J. E. Everhart, D. D. Byrd-Holt, J. F. Tisdale, and G. P. Rodgers, 2007 Prevalence of neutropenia in the U.S. population: age, sex, smoking status, and ethnic differences. Ann. Intern. Med. 146(7): 486–492.

Kammerer, C. M., N. Gouin, P. B. Samollow, J. F. VandeBerg, J. E. Hixson *et al.*, 2004 Two quantitative trait loci affect ACE activities in Mexican-Americans. Hypertension 43: 466–470.

Kang, G., W. Bi, Y. Zhao, J. F. Zhang, J. J. Yang *et al.*, 2014 A new system identification approach to identifying genetic variants in sequencing studies for a binary trait. Hum. Hered. 78: 104–116.

Kathiresan, S., O. Melander, C. Guiducci, A. Surti, N. P. Burtt *et al.*, 2008 Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat. Genet. 40: 189–197.

Lee, A. J., L. McMurchy, and A. J. Scott, 1997 Re-using data from case-control studies. Stat. Med. 16: 1377–1389.

Lin, D. Y., and D. Zeng, 2009 Proper analysis of secondary trait data in case-control association studies. Genet. Epidemiol. 33: 256–265.

Lin, D. Y., D. Zeng, and Z. Z. Tang, 2013 Quantitative trait analysis in sequencing studies under trait-dependent sampling. Proc. Natl. Acad. Sci. USA 110(30): 12247–12252.

Lin, H., M. Wang, J. A. Brody, J. C. Bis, J. Dupuis *et al.*, 2014 Strategies to design and analyze targeted sequencing data: cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium Targeted Sequencing Study. Circ. Cardiovasc. Genet. 7(3): 335–343.

Liu, C. C., T. Kanekiyo, H. Xu, and G. Bu, 2013 Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. Nat. Rev. Neurol. 9(2): 106–118.

Loos, R. J., C. M. Lindgren, S. Li, E. Wheeler, J. H. Zhao *et al.*, 2008 Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat. Genet. 40: 768–775.

Monsees, G. M., R. M. Tamimi, and P. Kraft, 2009 Genome-wide association scans for secondary traits using case-control samples. Genet. Epidemiol. 33: 717–728.

Nalls, M. A., J. G. Wilson, N. J. Patterson, A. Tandon, J. M. Zmuda *et al.*, 2008 Admixture mapping of white cell count: genetic locus responsible for lower white blood cell count in the Health ABC and Jackson Heart studies. Am. J. Hum. Genet. 82(1): 81–87.

Reiner, A. P., S. Beleza, N. Franceschini, P. L. Auer, J. G. Robinson *et al.*, 2012 Genome-wide association and population genetic analysis of C-reactive protein in African American and Hispanic American women. Am. J. Hum. Genet. 91(3): 502–512.

Simón-Sánchez, J., C. Schulte, J. M. Bras, M. Sharma, J. R. Gibbs *et al.*, 2009 Genome-wide association study reveals genetic risk underlying Parkinson's disease. Nat. Genet. 41(12): 1308–1312.

Song, X., I. Ionita-Laza, M. Liu, J. Reibman, and Y. We, 2016 A general and robust framework for secondary traits analysis. Genetics 202: 1329–1343 10.1534/genetics.115.181073.

Tapsoba Jde, D., C. Kooperberg, A. Reiner, C. Y. Wang, and J. Y. Dai, 2014 Robust estimation for secondary trait association in case-control genetic studies. Am. J. Epidemiol. 179(10): 1264–1272.

Thompson, J. F., M. Man, K. J. Johnson, L. S. Wood, M. E. Lira *et al.*, 2005 An association study of 43 SNPs in 16 candidate genes with atorvastatin response. Pharmacogenomics J. 5: 352–358.

Teslovich, T. M., K. Musunuru, A. V. Smith, A. C. Edmondson, I. M. Stylianou *et al.*, 2010 Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466: 707–713.

Wang, J., and S. Shete, 2011a Estimation of odds ratio of genetic variants for the secondary phenotypes associated with primary diseases. Genet. Epidemiol. 35(3): 190–200.

Wang, J., and S. Shete, 2011b Power and type I error results for a bias-correction approach recently shown to provide accurate odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. Genet. Epidemiol. 35(7): 739–743.

Wellcome Trust Case Control Consortium. 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447 (7145): 661–678.

Willer, C. J., E. K. Speliotes, R. J. Loos, S. Li, C. M. Lindgren *et al.*, 2008 Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat. Genet. 41: 25–34.

Communicating editor: N. Yi