Bayesian Detection of Causal Rare Variants under Posterior Consistency

Faming Liang[1]*, **Momiao Xiong**[2]

1 Department of Statistics, Texas A&M University, College Station, Texas, United States of America, **2** Division of Biostatistics, University of Texas School of Public Health, Houston, Texas, United States of America

Abstract

Identification of causal rare variants that are associated with complex traits poses a central challenge on genome-wide association studies. However, most current research focuses only on testing the global association whether the rare variants in a given genomic region are collectively associated with the trait. Although some recent work, e.g., the Bayesian risk index method, have tried to address this problem, it is unclear whether the causal rare variants can be consistently identified by them in the small-n-large-P situation. We develop a new Bayesian method, the so-called Bayesian Rare Variant Detector (BRVD), to tackle this problem. The new method simultaneously addresses two issues: (i) (Global association test) Are there any of the variants associated with the disease, and (ii) (Causal variant detection) Which variants, if any, are driving the association. The BRVD ensures the causal rare variants to be consistently identified in the small-n-large-P situation by imposing some appropriate prior distributions on the model and model specific parameters. The numerical results indicate that the BRVD is more powerful for testing the global association than the existing methods, such as the combined multivariate and collapsing test, weighted sum statistic test, RARECOVER, sequence kernel association test, and Bayesian risk index, and also more powerful for identification of causal rare variants than the Bayesian risk index method. The BRVD has also been successfully applied to the Early-Onset Myocardial Infarction (EOMI) Exome Sequence Data. It identified a few causal rare variants that have been verified in the literature.

Citation: Liang F, Xiong M (2013) Bayesian Detection of Causal Rare Variants under Posterior Consistency. PLoS ONE 8(7): e69633. doi:10.1371/journal.pone.0069633

Editor: Kai Wang, University of Southern California, United States of America

Received March 15, 2013; **Accepted** June 12, 2013; **Published** July 26, 2013

Funding: FL's research was partially supported by grants from the National Science Foundation (DMS-1007457 and DMS-1106494) and an award (KUS-C1-016-04) made by King Abdullah University of Science and Technology (KAUST). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: MX is a PLOS ONE editorial board member. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: fliang@stat.tamu.edu

Introduction

Testing the phenotypic association of millions of individual SNPs across the genome has been one of the major goals of the genome-wide association study (GWAS). To date, hundreds of putative disease gene loci have been detected based on the common disease common variant assumption. However, the detected genetic variants typically account for only a small fraction of disease heritability. Nowadays, it has been widely acknowledged that the missing disease heritability may be due to rare variants. Many studies show that the rare variants tend to have larger effects than common variants. As pointed out in [1], most rare variants can have much greater odds ratio than common variants, and many non-synonymous rare mutations from exon sequencing are functional variants for some common diseases. The rare variant effects have been investigated in some studies. For example, [2] found that the rare variants in the IFIH1 gene are strongly associated with Type I diabetes, and [3] found that multiple rare variants in NPC1L1 are associated with reduced sterol absorption and plasma low density lipoprotein levels. Therefore, development of statistical methods that are powerful enough to detect causal rare variants has become essential for the GWAS.

The statistical power of genetic variant detection depends on the sample size, the variant effect and the minor allele frequency (MAF). Since the MAF of the rare variant is low, the single variant testing-based methods, such as the χ^2-test and Fisher's exact test, that are traditionally used in common variant association studies, tend to have a low power. To address this issue, methods that test the collective effect of rare variants for a given genomic region have been developed, see e.g., the combined multivariate and collapsing (CMC) test [4], weighted sum statistic (WSS) test [5], and sequence kernel association test (SKAT) [6]. The CMC and WSS tests are variant pooling methods, in which the rare variants are collapsed or summed into a super-variant and then the disease association is tested with this super-variant. Their power can depend on the weighting scheme they employed, which often emphasizes low frequency alleles in controls. Numerous alternative methods [7,8] are largely their variations. The SKAT test is developed based on random effect models, which assumes a common distribution for the genetic effects of variants at different sites and tests for the null hypothesis that the distribution has zero variation.

Although testing the collective effects of rare variants is challenging, identifications of the rare variants which, if any, are driving the association (i.e., the so-called causal rare variants) is even more challenging and scientifically more interesting. Along this research direction, some methods have been developed, e.g., the RARECOVER method [9], variable threshold (VT) method

[10], evolutionary mixed model for pooled association testing (EMMPAT) method [11], hierarchical generalized linear model (HGLM) method [12,13], and Bayesian risk index (BRI) method [14]. The RARECOVER method uses a greedy search algorithm to determine an association set of variants. The VT method selects all variants with the MAF lower than a varying threshold to be included in the association set. The RARECOVER and VT focus mainly on the global association test and lack a formal test to determine the marginal effect of each variant, and thus are unable to formally determine which variants are most likely driving the association. The EMMPAT simultaneously evaluates the effects of all variants under the framework of mixed effect models. This is similar to HGLM, where the regression coefficients are simultaneously estimated for all variants. As a consequence of the simultaneous parameter estimation, when the number of variants is greater than the number of subjects, the variant effects evaluated by EMMPAT and HGLM might not be very reliable due to the multicollinearity of variants. The BRI is a Bayesian method, which can evaluate the marginal effect of each variant by allowing for uncertainty into which variants are included in the association set.

While BRI has made a solid step toward detection of causal rare variants, it is unclear whether it can identify causal rare variants consistently for small-n-large-P problems, in which the number of variants can be much greater than the number of subjects. In addition, BRI assumes the effect of each causal variant to be the same. Since this is not true for real problems, the performance of BRI may be sub-optimal. In this paper, we propose a new Bayesian method, the so-called Bayesian Rare Variant Detector (BRVD), for identification of causal rare variants. The new method simultaneously answers two questions:

- (Global association test) Are there any of the variants associated with the disease?
- (Causal variant detection) Which variants, if any, are driving the association?

The BRVD ensures the causal rare variants to be consistently identified in the small-n-large-P situation by imposing some appropriate prior distributions on the model and model specific parameters. In addition, to enhance detection of causal rare variants, the BRVD specifies for each variant a different prior selection probability (or weight) which is adversely proportional to its MAF. To accelerate the computation, we also propose a parallel version of BRVD based on the strategy of divide-and-conquer. The parallel BRVD has an embarrassingly parallel structure and can be conveniently applied to the problems for which the number of variants is extremely large. Our numerical results indicate that the BRVD can be more powerful for testing the global association than the existing methods, such as CMC, WSS, SKAT, C-alpha, RARECOVER, VT, and BRI, and more powerful than BRI for identification of causal rare variants. The BRVD has also been successfully applied to the early-onset myocardial infarction (EOMI) data: It identified a few causal rare variants that have been verified in the literature.

Materials and Methods

The global association test and Bayesian factor

Assume that n subjects are sequenced in a genomic region with P SNPs. Let X be a $n\times P$ genotype matrix coded as $X_{ij}=0,1,2$ for the number of copies of the minor allele measured for individual i at SNP j, let Z be a $n\times q$ matrix of covariates, e.g., age and race, and let Y be a n-dimensional binary vector indicating the disease status of the n subjects. The BRVD uses a logistic regression model to relate the covariates and a subset of variants to the disease status variable. Let ξ denote a subset of variants, and let $|\xi|$ denote the number of variants included in ξ. Let M_ξ denote the logistic regression model corresponding to the subset ξ, which can be expressed as

$$\operatorname{logit} P(\boldsymbol{Y}=1|M_\xi)=\alpha_0+\boldsymbol{Z}\,\boldsymbol{\alpha}+\boldsymbol{X}_\xi\,\boldsymbol{\beta}_\xi, \tag{1}$$

where $\boldsymbol{X}_\xi$ denotes the genotype matrix corresponding to the subset ξ, and α_0, $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_q)$ and $\beta_\xi=(\beta_1^\xi,\ldots,\beta_{|\xi|}^\xi)$ are the regression coefficients. For this model, the global association test is to test the hypotheses

$$H_0:|\xi|=0 \quad \mathit{versus} \quad H_1:|\xi|>0. \tag{2}$$

Let Ω_0 denote the parameter space of the null model M_0, i.e., the domain of the parameters α_0 and $\boldsymbol{\alpha}$. Let Ω_1 denote the parameter space of the alternative models, which can be expressed as $\Omega_1=\Omega_0\times\bigcup_{M_\xi\in\mathcal{M}}\Omega_\xi$, where $\mathcal{M}$ denotes the set of all possible models with $|\xi|>0$ and Ω_ξ is the domain of $\boldsymbol{\beta}_\xi$.

Let $\pi(\alpha_0,\boldsymbol{\alpha})$ denote the prior distribution of $(\alpha_0,\boldsymbol{\alpha})$, let $\pi(M_\xi|H_1)$ denote the prior probability imposed on the model M_ξ under the hypothesis H_1, and let $\pi(\beta_\xi|M_\xi,H_1)$ denote the prior distribution of β_ξ. Then the Bayesian factor for the test (2) can be expressed as

$$\begin{aligned} BF(H_1:H_0)&=\frac{\sum_{M_\xi\in\mathcal{M}}\pi(M_\xi|H_1)\int f_1(\boldsymbol{Y}|\alpha_0,\boldsymbol{\alpha},M_\xi,{}_1\beta_\xi)\pi(\alpha_0,\boldsymbol{\alpha})\pi(\boldsymbol{\beta}_\xi|M_\xi,H_1)d\alpha_0 d\,\boldsymbol{\alpha} d\,\boldsymbol{\beta}_\xi}{\int f_0(\,\boldsymbol{Y}|\alpha_0,\boldsymbol{\alpha})\pi(\alpha_0,\boldsymbol{\alpha})d\alpha_0 d\,\boldsymbol{\alpha}}\\ &\overset{\Delta}{=}\frac{\pi(\mathcal{D}|H_1)}{\pi(\mathcal{D}|H_0)}, \end{aligned} \tag{3}$$

where $f_0(\cdot)$ and $f_1(\cdot)$ denote the likelihood functions of the null and alternative models, respectively; $\mathcal{D}$ denotes the data; and $\pi(\mathcal{D}|H_1)$ and $\pi(\mathcal{D}|H_0)$ are the Bayesian evidence corresponding to the hypotheses H_1 and H_0, respectively. As in [14,15], (3) can also be expressed as the weighted average of the individual Bayes factors for comparing each model in H_1 to the null model M_0 with the weights given by the prior probability $\pi(M_\xi|H_1)$; that is,

$$BF(H_1:H_0)=\sum_{M_\xi\in\mathcal{M}}\pi(M_\xi|H_1)BF(M_\xi:M_0), \tag{4}$$

where $BF(M_\xi:M_0)$ is defined as the ratio of $\int f_1(\boldsymbol{Y}|\alpha_0,\boldsymbol{\alpha},M_\xi,\boldsymbol{\beta}_\xi)\pi(\alpha_0,\boldsymbol{\alpha})\pi(\boldsymbol{\beta}_\xi|M_\xi,H_1)d\alpha_0 d\,\boldsymbol{\alpha} d\,\boldsymbol{\beta}_\xi$ and $\int f_0(\boldsymbol{Y}|\alpha_0,\boldsymbol{\alpha})\pi(\alpha_0,\boldsymbol{\alpha})d\alpha_0 d\,\boldsymbol{\alpha}$. Let $\pi(H_0)$ denote the prior probability imposed on the null model, and let $\pi(H_1)=1-\pi(H_0)$ denote the total prior probabilities imposed on the alternative models. Then the respective posterior probabilities of H_0 and H_1 are given by

$$\pi(H_0|\mathcal{D})=\frac{\pi(H_0)}{\pi(H_0)+\pi(H_1)BF(H_1:H_0)},\quad \pi(H_1|D)=1-\pi(H_0|\mathcal{D}).$$

A value of $\mathrm{BF}(H_1:H_0)>1$ means that the alternative hypothesis is more strongly supported by the data under consideration than the null hypothesis. Harold Jeffreys [16] gave a scale, which is reproduced in Table 1, for interpretation of Bayes factors. Decisions about which hypothesis is more likely true can be made based on the scale of Bayes factors.

The Bayes factor (3) depends on the prior distributions, $\pi(\alpha_0,\boldsymbol{\alpha})$, $\pi(M_\xi|H_1)$, and $\pi(\boldsymbol{\beta}_\xi|M_\xi,H_1)$. In particular, the dependence on the model prior $\pi(M_\xi|H_1)$ can be substantial. This inevitably leads to

Table 1. Jeffrey's grades of evidence (Jeffreys, 1961).

Grade	BF(H_1:H_0)	$\pi(H_1\|D)$	Evidence against H_0
1	1~3	0.50~0.75	Barely worth mentioning
2	3~10	0.75~0.91	Substantial
3	10~30	0.91~0.97	Strong
4	30~100	0.97~0.99	Very strong
5	>100	>0.99	Decisive

The posterior probability $\pi(H_1|D)$ is calculated with the prior probabilities $\pi(H_0)=\pi(H_1)=1/2$.

doi:10.1371/journal.pone.0069633.t001

ambiguity in interpretation of Bayes factors. To minimize the ambiguity, we suggest to choose the priors $\pi(M_\xi|H_1)$ and $\pi(\boldsymbol{\beta}_\xi|M_\xi,H_1)$ such that the Bayesian evidence of H_1 is maximized. The resulting prior is the so-called type-II maximum likelihood prior [17]. Since maximizing the evidence over general priors is impossible, we further suggest to maximize the evidence over a specified class of priors. This will be detailed below. We note that a similar strategy has been suggested in [18] for testing a point null hypothesis. Since α_0 and $\boldsymbol{\alpha}$ are common parameters for all models, $\pi(\alpha_0, \boldsymbol{\alpha})$ is fixed to a Gaussian-truncated-inverse-gamma prior in all simulations of this paper.

The prior and posterior distributions

Let α_i, $i=0,1,\ldots,q$, be subject to the independent Gaussian prior:

$$\alpha_i \sim N(0,\sigma_\alpha^2), \quad i=0,1,\ldots,q, \tag{5}$$

where the variance σ_α^2 is subject to a truncated inverse-gamma prior

$$\sigma_\alpha^2 \sim IG(a,b;A,B), \tag{6}$$

defined on the interval $[A,B]$, where a and b are the shape and scale parameters, respectively. The density function of (6) is given by

$$f(\sigma_\alpha^2)=\frac{1}{Q(a,b/A)-Q(a,b/B)}\frac{b^a}{\Gamma(a)}\frac{e^{-b/\sigma_\alpha^2}}{\sigma_\alpha^{2(a+1)}},$$

where $Q(a,x)=\int_0^x e^{-t}t^{a-1}dt/\Gamma(a)$ is an incomplete gamma function and can be evaluated numerically. In the literature, σ_α^2 is usually assumed an inverse-gamma prior distribution. Here σ_α^2 is restricted to take values from the bounded interval $[A,B]$. As shown in Lemma 1 of File S1 (Section S1), this restriction plays an important role in establishing the posterior consistency [19,20] for the model (1). The posterior consistency means the true density of Y can be estimated consistently by the density of Y under the models sampled from the posterior distribution. For the same reason, we let $\beta_1^\xi,\ldots,\beta_{|\xi|}^\xi$ be subject to the independent Gaussian prior

$$\beta_i^\xi \sim N(0,\sigma_\beta^2), \quad i=1,2,\ldots,|\xi|, \tag{7}$$

with the variance σ_β^2 being subject to the truncated inverse-gamma prior $IG(a,b;A,B)$. For simplicity of computation, we further assume $\sigma_\alpha^2=\sigma_\beta^2$; that is, α_i and β_i^ξ have the same prior variance.

Let v_i denote the prior selection probability of variant i. Let $\delta_\xi(i)=1$ if variant i is included in the subset ξ and 0 otherwise. The prior probability of the model M_ξ under H_1 is given by

$$\pi(M_\xi|H_1)=\frac{\Pi_{i=1}^P v_i^{\delta_\xi(i)}(1-v_i)^{1-\delta_\xi(i)}}{1-\Pi_{i=1}^P(1-v_i)}. \tag{8}$$

To enhance selection of causal rare variants, we suggest to set v_i as a decreasing function of MAF. In this paper, we set

$$v_i=\frac{1}{1+P^{\gamma_i}}, \tag{9}$$

where γ_i is restricted to the interval $[\epsilon,1)$ for some constant $\epsilon>0$. In this paper, we set $\gamma_i=\gamma^L+(\gamma^R-\gamma^L)(\log(MAF_i)-\min_j\log(MAF_j))/(\max_j\log(MAF_j)-\min_j\log(MAF_j))$, where MAF_i denotes the minor allele frequency of variant i, and γ^L and γ^R are hyperparameters to be specified by the user. In addition, we fix $\gamma^R=0.99$ and choose $\gamma^L\in[\epsilon,\gamma^R]$ such that the Bayes factor $\mathrm{BF}(H_1:H_0)$ is maximized. Note that (9) is not necessarily optimal. In practice, one may try different settings for γ_i and γ^R.

As shown in File S1 (Section S1), the above prior setting, together with the identifiability condition of the true model, leads to the consistency of causal variant selection. Our priors are different from the conventional "Gaussian–inverse-gamma–beta" priors in two aspects. First, we let σ_α^2 and σ_β^2 be subject to the truncated inverse-gamma prior, which ensures the eigenvalues of the prior covariance matrix of $(\alpha_0,\alpha_1,\ldots,\alpha_q,\beta_1^\xi,\ldots,\beta_{|\xi|}^\xi)$ to be bounded. While the boundedness condition cannot be achieved with the inverse-gamma prior. Second, we define v_i in (9) as a decreasing function of P. As explained in [21], this is important for variant selection in the small-n-large-P scenario, because it controls for the multiplicity: If P grows large, then $v_i\to 0$. Under appropriate conditions, it can be shown that the resulting *a priori* model size $\sum_{i=1}^P v_i$ is bounded by a function (of n) of order $o(n^\zeta)$ for some $\zeta<1$. While this condition cannot be satisfied if v_i is subject to a beta prior for which both the shape and scale parameters are constants independent of n.

Let $\pi(H_0)$ and $\pi(H_1)$ denote the prior probabilities imposed on H_0 and H_1, respectively. Then the posterior distribution of the model (1) is given by

$$\pi(\alpha_0,\, \boldsymbol{\alpha},M_\xi,\, \boldsymbol{\beta}_\xi|\mathcal{D})=\frac{\pi(H_1)\pi(\alpha_0,\, \boldsymbol{\alpha},M_\xi,\, \boldsymbol{\beta}_\xi,\mathcal{D}|H_1)I(|\xi|\geq 1)+\pi(H_0)\pi(\alpha_0,\, \boldsymbol{\alpha},\mathcal{D}|H_0)I(|\xi|=0)}{\pi(H_1)\pi(\mathcal{D}|H_1)+\pi(H_0)\pi(\mathcal{D}|H_0)}, \tag{10}$$

where $I(\cdot)$ is the indicator function, and $\pi(\alpha_0,\, \boldsymbol{\alpha},M_\xi,\, \boldsymbol{\beta}_\xi,\mathcal{D}|H_1)$ and $\pi(\alpha_0,\, \alpha,\mathcal{D}|H_0)$ are given in File S1 (Section S0).

In all simulations of this paper, we fixed the hyperparameters $a=1$, $b=1$, $A=0.01$, $B=100.0$, and $\gamma^R=0.99$. The choice of a, b, A and B allows σ^2 to vary over the interval $[0.01,100]$ which is large enough for most rare variant selection problems. The only remaining hyperparameter is γ^L, which can be determined by maximizing the Bayes factor $\mathrm{BF}(H_1:H_0)$ over the interval $[\epsilon,\gamma^R]$. For most examples of this paper, we tried $\gamma^L=0.4$, 0.5, …, 0.9, 0.95, 0.99 or a subset of them.

Bayes factor estimation

For the global association test, the key step is Bayes factor estimation. As implied by (4), an exact evaluation of the global Bayes factor needs to sum over all models under H_1. When P is large, this is prohibitive. For this reason, [14,15] suggested to replace the sum over the entire model space M with the sum over the models sampled by a Markov chain Monte Carlo (MCMC) algorithm. However, the resulting estimator is shown to provide only a lower bound for the global Bayes factor. In this paper, we propose to estimate the global Bayes factor using the stochastic approximation Monte Carlo (SAMC) algorithm [22]. The resulting estimator is consistent.

To facilitate the description of the SAMC algorithm, we define the following notations. Let $\omega=(\alpha_0, \alpha, M_\xi, \beta_\xi, H_1)$ for a model simulated from the posterior distribution (10) under H_1, and let $\omega=(\alpha_0, \alpha, H_0)$ for a model simulated under H_0. Define

$$\psi(\omega)=\left\{\begin{array}{c}\pi(H_1)\pi(\alpha_0, \boldsymbol{\alpha}, M_\xi, \boldsymbol{\beta}_\xi, \mathcal{D}|H_1), \text{ under } H_1, \\ \pi(H_0)\pi(\alpha_0, \boldsymbol{\alpha}, \mathcal{D}|H_0), \text{ under } H_0,\end{array}\right\}$$

which is the unnormalized posterior distribution of the model (1). Let $U(\omega)=-\log\psi(\omega)$, which is called the energy function in terms of physics. To apply the SAMC algorithm to estimate the Bayes factor, we partition the sample space as follows: Treat $\boldsymbol{\Omega}_0$ as a single subregion, i.e., setting $E_0=\{\omega : |\xi|=0, \omega\in\boldsymbol{\Omega}_0\times\{H_0\}\}$, and partition $\boldsymbol{\Omega}_1$ according to the energy function into m subregions: $E_1=\{\omega : U(\omega)\le u_1, \omega\in\boldsymbol{\Omega}_1\times\{H_1\}\}$, $E_2=\{\omega : u_1<U(\omega)\le u_2, \omega\in\boldsymbol{\Omega}_1\times\{H_1\}\}$, $\ldots$, $E_{m-1}=\{\omega : u_{m-2}<U(\omega)\le u_{m-1}, \omega\in\boldsymbol{\Omega}_1\times\{H_1\}\}$, $E_m=\{\omega : U(\omega)>u_{m-1}, \omega\in\boldsymbol{\Omega}_1\times\{H_1\}\}$, where $u_1,\ldots,u_{m-1}$ are pre-specified numbers. The sample space $\boldsymbol{\Omega}_1$ can also be partitioned according to the value of $|\xi|$. However, when P is large, this alternative partition often leads to a slower convergence of SAMC, as which encourages SAMC to sample the models of different sizes instead of those of low energy values.

SAMC seeks to draw samples from each of the subregions with a pre-specified frequency. For the time being, we assume that all the $m+1$ subregions are non-empty; that is, $\int_{E_i}\psi(\omega)d\omega>0$ for $i=0,1,\ldots,m$. Let $p=(p_0,p_1,\ldots,p_m)$ denote the vector of desired sampling frequencies of the $m+1$ subregions, where $0<p_i<1$ and $\sum_{i=0}^m p_i=1$. Henceforth, p is called the desired sampling distribution. Let $\theta_i=\log(\int_{E_i}\psi(\omega)d\omega/p_i)$ for $i=0,1\ldots,m$, let $\boldsymbol{\theta}=(\theta_0,\theta_1,\ldots,\theta_m)$, and let Θ denote the domain of $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}^{(t)}=(\theta_0^{(t)},\theta_1^{(t)},\ldots,\theta_m^{(t)})$ denote the working estimate of $\boldsymbol{\theta}$ obtained at iteration t. Let $\omega^{(t+1)}$ denote a sample drawn at iteration $t+1$ from the MH kernel $K_{\boldsymbol{\theta}^{(t)}}(\omega^{(t)},\cdot)$, which is constructed with the proposal distribution $T(\omega^{(t)},\cdot)$ and admits (11) as the invariant distribution:

$$f_{\boldsymbol{\theta}^{(t)}}(\omega)\propto\sum_{i=0}^{m}\frac{\psi(\omega)}{e^{\theta_i^{(t)}}}I(\omega\in E_i). \tag{11}$$

Define $R(\boldsymbol{\theta}^{(t)},\omega^{(t+1)})=\boldsymbol{e}^{(t+1)}-\boldsymbol{\pi}$, where $\boldsymbol{e}^{(t+1)}=(e_0^{(t+1)},\ldots,e_m^{(t+1)})$ and $e_i^{(t+1)}=1$ if $\omega^{(t+1)}\in E_i$ and 0 otherwise. Note that the dependence of $R(\cdot,\cdot)$ on $\theta^{(t)}$ is implicit through the sample $\omega^{(t+1)}$. To have the algorithm complied with the notation of stochastic approximation, $\boldsymbol{\theta}^{(t)}$ is still included in the function $R(\cdot,\cdot)$. Let $\{a_t\}$ be a positive, non-decreasing sequence satisfying the conditions,

$$(i)\ \sum_{t=0}^{\infty}a_t=\infty, \qquad (ii)\ \sum_{t=0}^{\infty}a_t^{\tau}<\infty, \tag{12}$$

for some $\tau\in(1,2]$. In the context of stochastic approximation, $\{a_t\}_{t\ge 0}$ is called the gain factor sequence.

In this paper, we assume that Θ is compact; that is, assuming that the sequence $\{\boldsymbol{\theta}^{(t)}\}$ can be kept in a compact set. Extension of this algorithm to the case that $\Theta=\mathbb{R}^{m+1}$ is trivial with the technique of varying truncations studied in [23,24], which ensures, almost surely, that the sequence $\{\boldsymbol{\theta}^{(t)}\}$ remains in a compact set. In simulations, we can set Θ to a huge set, e.g., $\Theta=[-10^{100},10^{100}]^{m+1}$, which, as a practical matter, is equivalent to setting $\Theta=\mathbb{R}^{m+1}$. Let $J(\omega)$ denote the index of the subregion that the sample ω belongs to, which takes values in $\{0,1,\ldots,m\}$. With the above notations, one iteration of SAMC can be described as follows.

Algorithm 0.1 (The SAMC algorithm)

(a) (Sampling) Simulate a sample $\omega^{(t+1)}$ *by a single MH update with the target distribution as defined in (11):*

(a. 1) Generate ω' *according to a proposal distribution* $T(\omega^{(t)},\omega')$*. Refer to File S1 (Section S2) for the definition of* $T(\omega^{(t)},\omega')$*.*

(a. 2) Calculate the ratio

$$r=e^{\theta_{k_t}^{(t)}-\theta_{k'}^{(t)}}\frac{\psi(\omega)T(\omega',\omega^{(t)})}{\psi(\omega^{(t)})T(\omega^{(t)},\omega')}, \tag{13}$$

where $k_t=J(\omega^{(t)})$ *and* $k'=J(\omega')$ *are the indices of the subregions that* $\omega^{(t)}$ *and* ω' *belong to, respectively.*

(a. 3) Accept the proposal with probability $\min(1,r)$*. If it is accepted, set* $\omega^{(t+1)}=\omega'$*; otherwise, set* $\omega^{(t+1)}=\omega^{(t)}$*.*

(b) ($\boldsymbol{\theta}$*-updating) Set*

$$\boldsymbol{\theta}^{(t+\frac{1}{2})}=\boldsymbol{\theta}^{(t)}+a_{t+1}R(\boldsymbol{\theta}^{(t)},\omega^{(t+1)}). \tag{14}$$

If $\boldsymbol{\theta}^{(t+\frac{1}{2})}\in\Theta$*, set* $\boldsymbol{\theta}^{(t+1)}=\boldsymbol{\theta}^{(t+\frac{1}{2})}$*; otherwise, find a value of* c *such that* $\boldsymbol{\theta}^{(t+\frac{1}{2})}+c1_{m+1}\in\Theta$ *and set* $\boldsymbol{\theta}^{(t+1)}=\boldsymbol{\theta}^{(t+\frac{1}{2})}+c1_{m+1}$*, where* 1_{m+1} *denotes a constant* $(m+1)$*-vector of ones.*

SAMC is an adaptive MCMC algorithm for which the invariant distribution of the MH kernel changes from iteration to iteration. Due to the adaptive change of the invariant distributions, SAMC possesses a self-adjusting mechanism: If a proposal is rejected, then the sample $\omega^{(t+1)}$ will be retained in the current subregion, the θ-value associated with the current subregion will be adjusted to a larger value, and the overall rejection probability of the next iteration will be reduced. This mechanism warrants the algorithm not to be trapped by local energy minima. The SAMC algorithm represents a significant advance in simulations of complex systems for which the energy landscape is rugged.

The proposal distribution $T(\omega,\omega')$ is usually assumed to satisfy the local positive condition: For every $\omega\in\boldsymbol{\Omega}$, there exist $\epsilon_1>0$ and $\epsilon_2>0$ such that

$$\|\omega-\omega'\|\le\epsilon_1\Rightarrow T(\omega,\omega')\ge\epsilon_2, \tag{15}$$

where $\|\omega-\omega'\|$ denotes a distance norm between ω and ω'. This is a natural condition in MCMC theory. In practice, this kind of proposals can be easily designed for both discrete and continuum systems as discussed in the literature [22]. Regarding the convergence of SAMC, [22] established the following result: Under the conditions (12) and (15) and some regularity conditions, for all non-empty subregions,

$$\theta_i^{(t)} \to C_0 + \log\left(\int_{E_i} \psi(\omega)d\omega\right) - \log(p_i + p_e), \quad a.s., \tag{16}$$

as $t \to \infty$, where $p_e = \sum_{j \in \{i: E_i = \varnothing\}} p_j/(m+1-m_0)$, $m_0 = \#\{i : E_i = \varnothing\}$ is the number of empty subregions, and C_0 is a constant which can be determined by imposing a constraint on $\theta_i^{(t)}$, e.g., $\sum_{i=0}^{m} \exp(\theta_i^{(t)}) = 1$.

For global association tests, we set the desired sampling distribution to be uniform, i.e., setting $p_0 = p_1 = \cdots = p_m = 1/(m+1)$. For mathematical simplicity, we have constrained $\boldsymbol{\Omega}_0$ and $\boldsymbol{\Omega}_1$ to two large compact sets by restricting $(\alpha_0, \alpha, \beta_\xi)$ to the set $[-10^{100}, 10^{100}]^{1+q+|\xi|}$, which, as a practical matter, is equivalent to $\mathbb{R}^{1+q+|\xi|}$. The gain factor sequence $\{a_t\}$ is set in the form

$$a_t = \frac{t}{\max\{t, t_0\}}, \tag{17}$$

where $t_0 > 0$ is a user-specified number. It is easy to verify that (17) satisfies the condition (12). A large value of t_0 will allow the SAMC sampler to reach all subregions quickly, even when m is large. The proposal distribution $T(\omega, \omega')$ is described in File S1 (Section S2). It is easy to see that it satisfies the condition (15). Then, by (16), we have the following result:

$$\frac{\sum_{i=1}^{m} e^{\theta_i^{(t)}}}{e^{\theta_0^{(t)}}} \to BF(H_1 : H_0), \quad a.s., \tag{18}$$

as $t \to \infty$. That is, SAMC provides a consistent estimator for the Bayes factor.

Rare variant detection

In this section, we describe how to detect rare variants when the global association test shows positive support for the hypothesis H_1.

Identification of important variables based on the marginal inclusion probability has been widely used in Bayesian variable selection, see, for example, [25] for the case of large-n-small-P normal linear models, and [26] for small-n-large-P generalized linear models. Let q_j denote the marginal inclusion probability of variable j. A conventional rule is to choose the variables for which the marginal inclusion probability is greater than a threshold value $\hat{q}$; i.e., setting $\widehat{\xi}_{\hat{q}} = \{x_j : q_j > \hat{q}, j = 1, \ldots, P_n\}$ as an estimator of ξ_*, the set of true model variables. Based on [26], we show in Lemma 2 of File S1 (Section S1) that this rule possesses the properties of sure screening and consistency for rare variant detection under the priors given in Section 0. The sure screening property implies that for some choice of $\hat{q} \in (0,1)$,

$$P(\xi_* \subset \widehat{\xi}_{\hat{q}}) \to 1,$$

as the sample size n tends to infinity. The property of variant selection consistency implies that

$$P(\xi_* = \widehat{\xi}_{0.5}) \to 1,$$

as the sample size n tends to infinity.

To implement the rule $\widehat{\xi}_{\hat{q}}$ for causal variant detection, one needs a consistent estimator for the marginal inclusion probability under H_1 and a method for determining the threshold value $\hat{q}$. In SAMC, the marginal inclusion probability can be consistently estimated as follows. Let $(\omega^{(1)}, \boldsymbol{\theta}^{(1)}), \ldots, (\omega^{(N)}, \boldsymbol{\theta}^{(N)})$ denote the samples drawn by SAMC in a run. Liang [27] showed that SAMC is actually a dynamic importance sampling algorithm and for any integrable function $\rho(\omega)$, as $N \to \infty$,

$$\frac{\sum_{t=1}^{N} e^{\theta_{J(\omega^{(t)})}^{(t)}} \rho(\omega^{(t)})}{\sum_{t=1}^{N} e^{\theta_{J(\omega^{(t)})}^{(t)}}} \to E_\pi \rho(\omega), \quad a.s., \tag{19}$$

where $E_\pi \rho(\omega)$ denotes the expectation of $\rho(\omega)$ with respect to the target distribution $\pi(\omega|\mathcal{D})$. This result implies

$$\hat{q}_j = \frac{\sum_{t=1}^{N} e^{\theta_{J(\omega^{(t)})}^{(t)}} I(\boldsymbol{x}_j \in \xi_t)}{\sum_{t=1}^{N} e^{\theta_{J(\omega^{(t)})}^{(t)}} I(|\xi_t| \geq 1)} \to q_i, \quad a.s., \tag{20}$$

as N goes to infinity; that is, the estimator $\hat{q}_j$ is consistent.

To determine the threshold $\hat{q}$, [26] proposed a multiple hypothesis testing-based procedure based on the work [28]. This procedure is adopted in the paper and briefly described in File S1 (Section S3).

Empirical Power Simulations

To explore the power of the proposed method versus other alternative methods for the global association tests and rare variant detection, we simulated 200 datasets, with 100 simulated under H_0 and 100 under H_1. Each dataset consists of 250 cases and 250 controls, and each subject consists of $q = 2$ covariates. The first covariate is binary, which mimics the gender of the subjects. The second covariate is drawn uniformly from the interval [10,85], which mimics the age of the subjects. The regression coefficients of the two covariates are set to $\alpha_1 = 0.25$ and $\alpha_2 = 0.01$, respectively. The genotypes of each subject are simulated by resampling from a haplotype dataset given in the package *SKAT*. The haplotype dataset is generated by the calibrated coalescent model with a mimicking linkage disequilibrium (LD) structure of European ancestry. To emphasize rare events, the variants with MAF greater than 5% have been removed from the haplotype dataset before resampling. For the 100 datasets simulated under H_1, the first 10 variants are assumed to be causal with the regression coefficients given by (2.09,1.90,1.85,1.82,1.57,1.96,1.40,1.93,2.20,2.00), which represents a random sample drawn from $N(2, 0.25^2)$. Then we remove the zero-MAF variants from the resampled dataset and keep only the first 600 non-zero MAF variants for further analysis. Because of this deletion step, the number of causal variants becomes a random variable for each dataset. For the 100 datasets simulated under H_1, the number of causal variants ranges from 5 to 9, and has a mean value of 7.81 with standard deviation 0.92. The average MAF of the first 9 variants is 0.833% with standard deviation 0.0012. Among the first 9 variants, the maximum MAF is 1.155%. Variants 1 and 2 have very low MAFs, which are 0.183% and 0.293%, respectively. Due to their low MAFs, identification of the causal variants, especially for variants 1 and 2, has put a great challenge on the existing methods.

Comparison with Other Methods

We compare the BRVD with the competing Bayesian method *Bayesian risk index (BRI)* for both global association tests and causal variant detection. We also compare BRVD with the commonly used non-Bayesian methods, including CMC, WSS, SKAT, and RARECOVER, for global association tests. Among the four non-Bayesian methods, CMC and WSS belong to the class of variant pooling methods, SKAT belongs to the class of random effect model-based methods, and RARECOVER belongs to the class of variable selection methods. These methods can be briefly described as follows.

- Bayesian risk index (BRI) [14]: For a model M_ξ, the BRI defines the risk index as the sum of the selected variants, i.e.,

$$R_\xi = \boldsymbol{X}\,\boldsymbol{\delta}_\xi,$$

where $\delta_\xi = (\delta_\xi(1), \ldots, \delta_\xi(P))'$ is a binary vector which indicates the variants included in the model M_ξ. Then it conducts an approximate Bayesian analysis for the model

$$\mathrm{logit}\, P(\boldsymbol{Y}=1|M_\xi) = \alpha_0 + \boldsymbol{Z}\,\boldsymbol{\alpha} + R_\xi \beta_\xi,$$

under a Beta-Binomial prior for the model size. The prior specification for $(\alpha_0, \boldsymbol{\alpha}, \beta_\xi)$ is avoided in BRI, as it directly works on the marginal likelihood $P(\boldsymbol{Y}|M_\xi)$ with the parameters $(\alpha_0, \boldsymbol{\alpha}, \beta_\xi)$ replaced by their MLE. The significance of global association is determined using the Bayes factor calculated in (4) with posterior samples. The rare variants are selected based on the marginal Bayes factor which, for any two variants, is defined as the ratio of the odds of their posterior marginal inclusion probabilities to the odds of their prior marginal inclusion probabilities.

- *Combined multivariate and collapsing (CMC) test* [4]: CMC is a variant pooling method in which the rare variants are grouped according to their allele frequency. After grouping, the rare variants are collapsed into an indicator variable, and then a multivariate test such as Hotelling's T^2 test is applied to the collection formed by the common variants and the collapsed super-variant.
- *Weighed sum Statistic (WSS) test* [5]: WSS is a variant pooling method. It first calculates for each subject a genetic score, which accumulates the rare variants counts within the same gene with a weighting term that emphasizes alleles with a low frequency in controls. Then the scores for all subjects are ordered, and the WSS is computed as the sum of the ranks for the cases. The significance is determined by a permutation procedure.
- *Sequence kernel association (SKAT) test* [6]: SKAT is a random effect model-based method. It assumes a common distribution for the genetic effects of different variants and test for the null hypothesis that the distribution has zero variance.
- *RARECOVER* [9]: RARECOVER is a variable selection-based method. It selects variants in a manner of forward variable selection: Starting from a null model without any genetic variants, the variants are added into the model one by one based on their statistical significance. The significance of global association is determined by a permutation procedure.

The implementation of BRI is available in the R package BVS, the implementation of SKAT is available in the R package SKAT, and the implementations of CMC, WSS, and RARECOVER are available in the R package AssotesteR. In this paper, all the methods are run under their default settings unless otherwise stated.

Results

Global Power

We first aim to examine the power of the BRVD versus alternative methods for global association tests. The BRVD has a prior hyperparameter γ^L to tune. To determine the value of γ^L, we tried the values 0.4, 0.5, …, 0.9, and 0.99 for all the 200 simulated datasets. For each dataset and each value of γ^L, SAMC was run for 5.05×10^6 iterations, where the first 50000 iterations were for the burn-in process and the samples generated from the remaining iterations were used for inference. The gain factor sequence was set in (17) with $t_0 = 1000$, and the sample space $\boldsymbol{\Omega}_1$ was partitioned into $m = 99$ equally spaced (in energy values) subregions with $u_1 = 341$ and $u_{m-1} = 439$. Figure 1 (a) & (b) show the average posterior probability $\bar{\pi}(H_1|\mathcal{D},\gamma^L)$ versus γ^L for the datasets simulated under H_1 and H_0, respectively, where the average is calculated over 100 datasets. To indicate the dependency of the average posterior probability on γ^L, we include γ^L in the notation. For the datasets simulated under H_1, $\bar{\pi}(H_1|\mathcal{D},\gamma^L)$ attains its maximum at $\gamma_L = 0.6$; and for the datasets simulated under H_0, $\bar{\pi}(H_1|\mathcal{D},\gamma^L)$ attains its maximum at $\gamma_L = 0.99$. This is interesting: A small value of γ_L encourages selection of variants, while a large value of γ_L discourages selection of variants. This is consistent with our design of the study: More variants are preferred to be selected for the datasets simulated under H_1. Figure 1 shows $\bar{\pi}(H_1|\mathcal{D},\gamma^L)$ versus different values of γ^L: $\bar{\pi}(H_1|\mathcal{D},\gamma^L)$ changes only about %2 over the interval $0.4 \le \gamma^L \le 0.99$ for the datasets simulated under H_1, and changes only about 7% for the datasets simulated under H_0. Therefore, we may conclude that the posterior probability $\pi(H_1|\mathcal{D},\gamma^L)$ is quite robust to the choice of γ^L.

Since BRVD, BRI and SKAT are all developed under the regression setting, they are able to adjust for covariates, such as age, gender, race, etc. For this reason, we first compare the powers of these three methods with the simulated covariates adjusted in regression. Figure 2 compares the ROC curves for the global association test, which plots the global false-positive rate (gFPR) versus global true-positive rate (gTPR) as the global BF threshold varies for BRVD and BRI, and the *p*-value threshold varies for SKAT. As in BRI, the gFPR is calculated as the ratio of the number of null datasets (the datasets simulated under H_0) for which a global association has been detected versus the total number of null datasets, and the gTPR is calculated as the number of associated datasets (the datasets simulated under H_1) for which a global association has been detected versus the total number of associated datasets. Figure 2(a) shows that for this example, BRVD has about the same power as SKAT and much greater power than BRI to detect a global association. Note that in this plot, we have followed the procedure suggested in Section 2.1 to calculate the gFPR for the null datasets with $\gamma^L = 0.99$ and calculate gTPR for the associated datasets with $\gamma^L = 0.6$. To show the performance of BRVD is robust to the choice of γ^L, we plot in Figure 2(b) a few ROC curves, where for each curve both gFPR and gTPR were calculated at the same value of γ^L. The plot indicates that the BRVD is very robust to the choice of γ^L for global association tests.

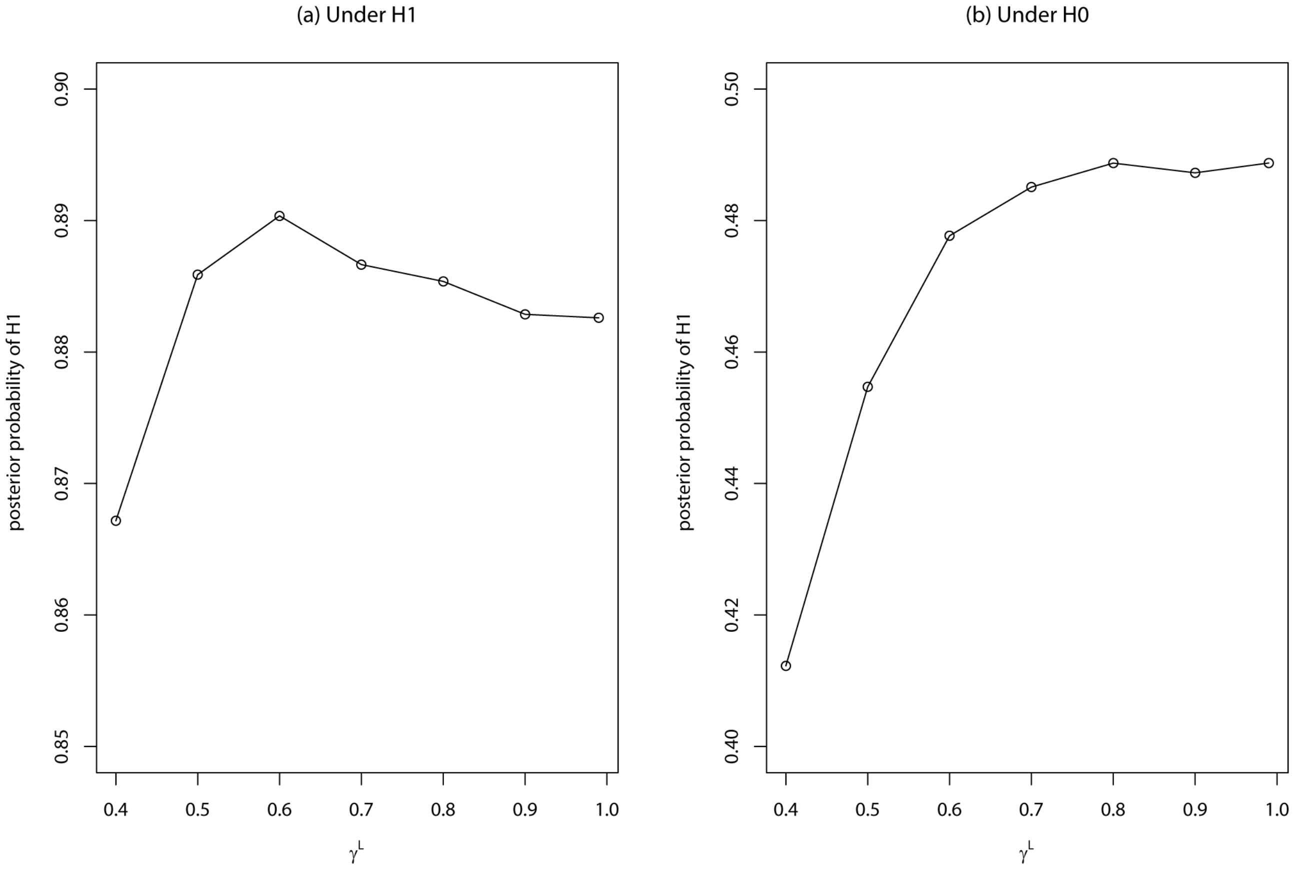

Figure 1. The average posterior probability $\bar{\pi}(H_1|\mathcal{D},\gamma^L)$ versus γ^L for the datasets simulated under H_1 (plot (a)) and under H_0 (plot (b)).
doi:10.1371/journal.pone.0069633.g001

The CMC, WSS and RARECOVER cannot be adjusted for covariates. To compare with them, we re-run the BRVD, BRI and SKAT methods on the simulated datasets with the covariates omitted. The effect of covariate omission on test power has been discussed in the literature [29,30,31]. The results seem mixed. Under certain situations, such as rare diseases and large sample sizes, omitting the covariates, which are known to affect disease susceptibility and are independent of tested genotypes, can increase the power to detect new genetic associations; whereas, for common diseases, it can decrease the power [31]. For BRVD, SAMC was run for these datasets with the same setting as for the case with covariates adjusted. Figure 3(a) compares the ROC curves of the six methods for global association tests. It shows that when covariates are omitted, BRVD has much greater power than all other methods. Compared to Figure 2(a), we may conclude that BRVD is more robust to covariate omission than the SKAT method. This is important for the success of a method, as in practice we may inevitably have some covariates omitted due to the limitation of our measurements. Figure 3(b) compares the ROC curves of BRVD calculated with different values of γ^L. It shows again that the power of BRVD is robust to the choice of γ^L for global association tests.

In addition to the power, we also explored the type-I error of the global association test based on the testing statistic $\max_{\gamma^L\in\Lambda}\pi(H_1|D,\gamma^L)$ for the simulated examples, where $\Lambda=\{0.4,0.5,\ldots,0.9,0.99\}$ and the prior probabilities $\pi(H_0)=\pi(H_1)=1/2$. The results, for both cases with and without covariate adjustment, are summarized in Figure 4. Following from Table 1, we suggest to choose 0.75 as the threshold value of $\max_{\gamma^L\in\Lambda}\pi(H_1|D,\gamma^L)$; that is, rejecting H_0 if $\max_{\gamma^L\in\Lambda}\pi(H_1|D,\gamma^L)>0.75$. With this threshold value, the resulting type-I errors are 0.01 and 0.02 for the cases with and without covariate adjustment, respectively.

Rare Variant Detection

Our next aim is to detect rare variants that are associated with the disease, provided that the global association test shows a positive support for the hypothesis H_1. Figure 5 compares the ROC curves of BRVD and BRI for rare variant detection, which are calculated based on the 100 datasets simulated under H_1. The ROC curves plot the marginal false-positive rate (mFPR) versus marginal true-positive rate (mTPR) as the marginal inclusion probability threshold varies for BRVD and the marginal BF threshold varies for BRI. As in BRI, the mFPR is calculated as the ratio of the number of non-associated variants for which a marginal association has been detected versus the total number of non-associated variants, and the mTPR is calculated as the ratio of the number of associated variants for which a marginal association has been detected versus the total number of associated variants. In drawing Figure 5, the marginal inclusion probabilities for both BRVD and BRI have been averaged over 100 datasets. The left panel of Figure 4 shows the ROC curves for the case with covariates adjusted, and the right panel shows for the case with covariates omitted. In both cases, the BRVD has much greater power than BRI for detection of causal rare variants, especially

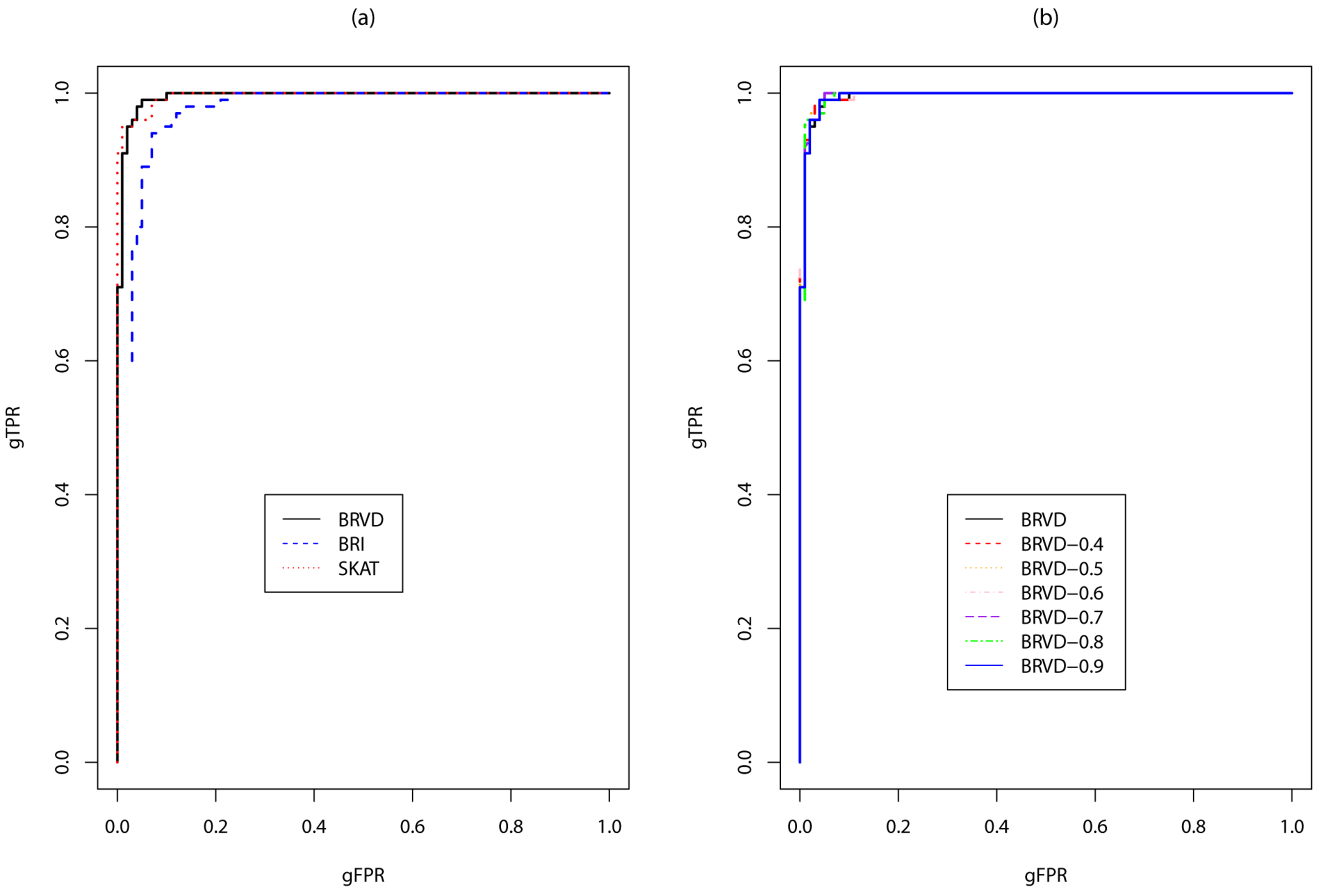

Figure 2. Global ROC curves for BRVD versus BRI and SKAT for the simulated example (with covariate adjustment). Each plot represents a ROC curve as we vary the global BF threshold for BVRD and BRI, and vary the p-value threshold for SKAT.
doi:10.1371/journal.pone.0069633.g002

when γ^L is small, e.g., $\gamma^L = 0.4$, 0.5 and 0.6. When $\gamma^L = 0.99$, under which all alleles are treated equally, the BRVD has about the same power as BRI. It is worth noting that the BRVD yields its worst result at $\gamma^L = 0.99$.

For global association tests, we suggest to choose the value of γ^L such that the Bayes factor $\mathrm{BF}(H_1 : H_0)$ is maximized. Figure 5 suggests that this is still a reasonable rule for determining the value of γ^L even when our aim is to detect causal rare variants. At $\gamma^L = 0.6$, BRVD performs reasonably well: The top 9 variants (ranked in marginal inclusion probabilities) include 7 causal variants, and variants 1 and 2 are ranked 22 and 19, respectively. For this example, we find that a smaller value of γ^L may result in a greater power of BRVD to detect causal rare variants. For example, at $\gamma^L = 0.4$, the top 10 variants include all 9 causal variants, and variants 1 and 2 are ranked 4 and 9, respectively. At $\gamma^L = 0.5$, the top 10 variants include 8 causal variants (1,3–9), and variant 2 is ranked 15. This is remarkable, as both variants 1 and 2 have very low MAFs. In BRI, although the variants 3–9 have high ranks in their marginal BFs, variants 1 and 2 are ranked 542 and 68, respectively. This implies that BRI essentially fails to detect variants 1 and 2. The results of this example suggest an alternative rule for determining the value of γ^L: If we aim to detect rare variants, we may choose a small value of γ^L such that some rare variants, such as those singleton variants, can be ranked high in their marginal inclusion probabilities, provided that the association set includes some singleton variants in *a priori* knowledge.

Figure 6 illustrates how to identify causal variants based on their marginal inclusion scores. The left panel of Figure 6 shows the result for $\gamma^L = 0.6$. At the FDR level of 0.05, 10 variants are identified as causal variants, and 7 of them (including variants 3–9) are true causal variants. At the FDR level of 0.01, 7 variants are identified and 6 of them (variants 4–9) are true. The right panel of Figure 6 shows the result for $\gamma^L = 0.5$. At the FDR level of 0.05, 11 variants are identified as causal variants, and 8 of them (variants 1, 3–9) are true. At the FDR level of 0.01, 7 variants are identified and 6 of them (variants 4–9) are true. The results for other values of γ^L are similar.

Application to the Early-Onset Myocardial Infarction (EOMI) Exome Sequence Data

The EOMI data (downloaded from dbGaP) is from the NHLBI€s Exome Sequencing Project (ESP), which was designed to identify genetic variants in coding regions (exons) of the human genome that are associated with heart, lung and blood diseases. The dataset consists of 278,263 SNPs in 905 subjects (467 cases and 438 controls) with European origin (EA). After removing the common variants (with MAF > 5%) and the variants with zero MAFs, the number of variants is reduced to 113,438. A direct application of BRVD to this dataset is time consuming as it may need an order of 10^8 iterations. In addition, the whole dataset need to be scanned once for each iteration. To resolve this issue, we propose, based on the strategy of divide-and-conquer, the following procedure:

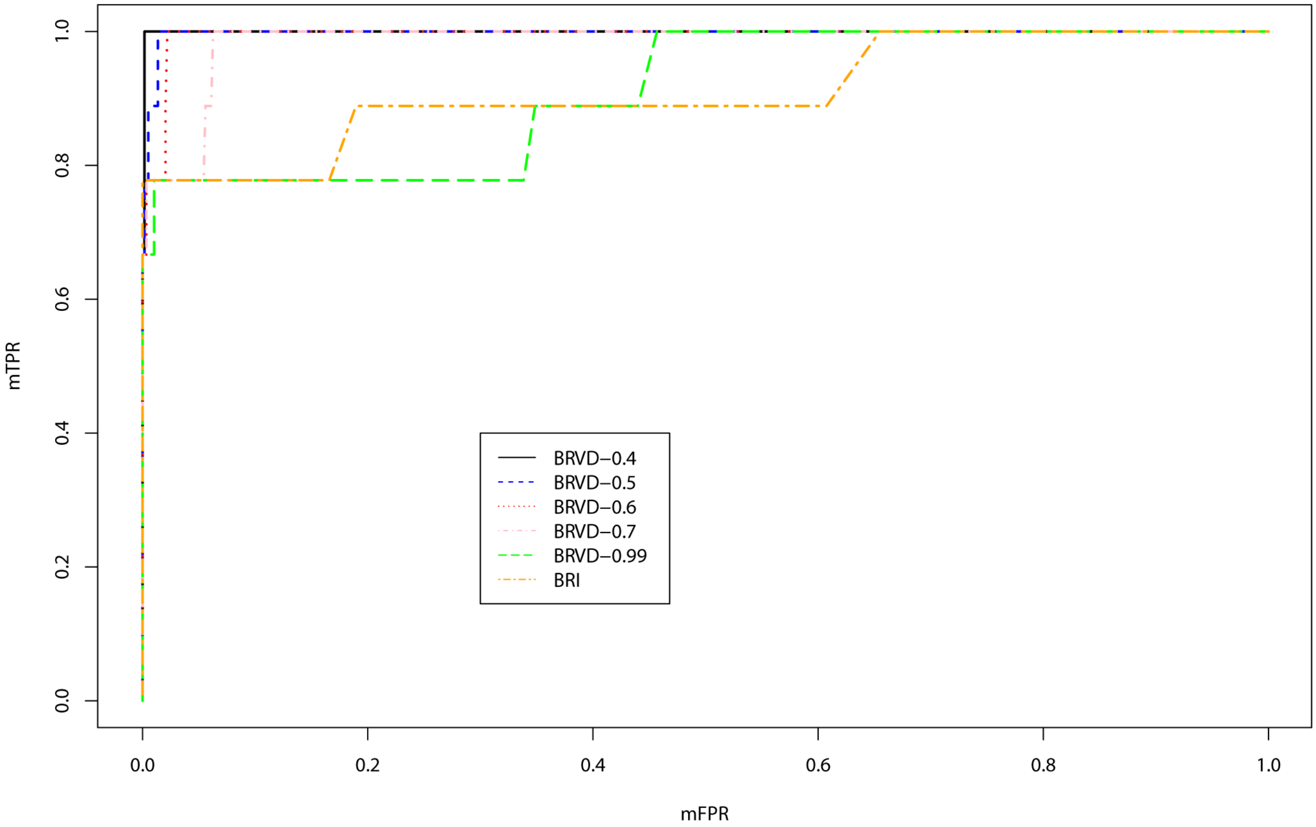

Figure 3. Global ROC curves for BRVD versus BRI, SKAT, CMC, WSS and RARECOVER for the simulated examples (without covariate adjustment). Each plot represents a ROC curve as we vary the global BF threshold for BRVD and BRI, and vary the p-value threshold for SKAT, CMC, WSS and RARECOVER.
doi:10.1371/journal.pone.0069633.g003

Parallel BRVD

(a) (Dividing) Divide the variants into subsets that are of an acceptable size in computation.

(b) (Parallel conquering) Apply BRVD to each of the subsets and identify putative associated variants from the subsets for which the hypothesis H_1 is supported.

(c) (Combining) Combine the variants identified at step (b) into a new dataset, the so-called selected subset data; and then apply BRVD to the selected subset data to identify causal rare variants.

For each subset, the logistic regression model is potentially misspecified because the causal variants located in other subsets are not included in the regression. If some causal variants are missed, we can expect that the BRVD will find some surrogate variants within the subset for the missing causal variants, and the number of surrogate variants can often be greater than the number of missing causal variants. For this reason, we suggest a high FDR level, say, 0.25 or even higher, to be used for identifying putative causal variants from each subset. For the selected subset data, we can expect that it will include the causal variants, surrogate variants of some causal variants, and some noise variants. It is obvious that Lemma 1 and Lemma 2 are still applicable to the selected subset data. By these two lemmas, the parallel BRVD can also select causal variants consistently.

The global association test can also be done on the selected subset of variants. However, a direct application of the BRVD to this subset can lead to a biased test, although for which the power can be very high. This is the same for all other testing procedures. To avoid the bias, a permutation method can be used to evaluate the p-value of the test. For example, one can permute the response variable a large number of times. For each of permuted datasets, the parallel BRVD can be applied to identify a selected subset of variants and then obtain a Bayes factor for the global association test based on the selected subset. Finally, a p-value can be calculated based on the Bayes factors of the permuted datasets.

For the EOMI dataset, we divide the variants into 22 subsets according to the chromosomes where they belong to. The numbers of variants on the 22 chromosomes range from 1,271 (on chromosome 21) to 11,491 (on chromosome 1), which are all acceptable to our current computing facility. BRVD was run 5 times for each subset at each value of $\gamma^L = 0.6$, 0.7, 0.8 and 0.9, and each run consisted of 2.5×10^7 iterations. The gain factor sequence was set in (17) with $t_0 = 5000$, and the sample space Ω_1 was partitioned into $m = 599$ equally spaced (in energy values) subregions with $u_1 = 601$ and $u_{m-1} = 1199$. Table 2 summarizes the posterior probabilities of H_1 for the 22 chromosomes. The support for the hypothesis H_1 is overwhelming: $\max_{\gamma^L \in \mathit{\Lambda}} \bar{\pi}(H_1|\mathcal{D},\gamma^L)$ is greater than 0.5 for all 22 chromosomes, where the probability $\bar{\pi}(H_1|\mathcal{D},\gamma^L)$ is calculated by averaging over 5 independent runs and $\Lambda = \{0.6,0.7,0.8,0.9\}$ denotes the set of values of γ^L we have tried. According to the value of $\max_{\gamma^L \in \Lambda} \bar{\pi}(H_1|\mathcal{D},\gamma^L)$, the chromosomes can be classified into two groups: chromosomes 13, 2, 3 and 19 are in the first group with $\max_{\gamma^L \in \Lambda} \bar{\pi}(H_1|\mathcal{D},\gamma^L) \geq 0.7$, and all other chromosomes are in the second group with $0.5 < \max_{\gamma^L \in \Lambda} \bar{\pi}(H_1|\mathcal{D},\gamma^L) < 0.57$. Among the first group chromosomes, chromosomes 13 and 2 provide "substantial" evidence for the global association.

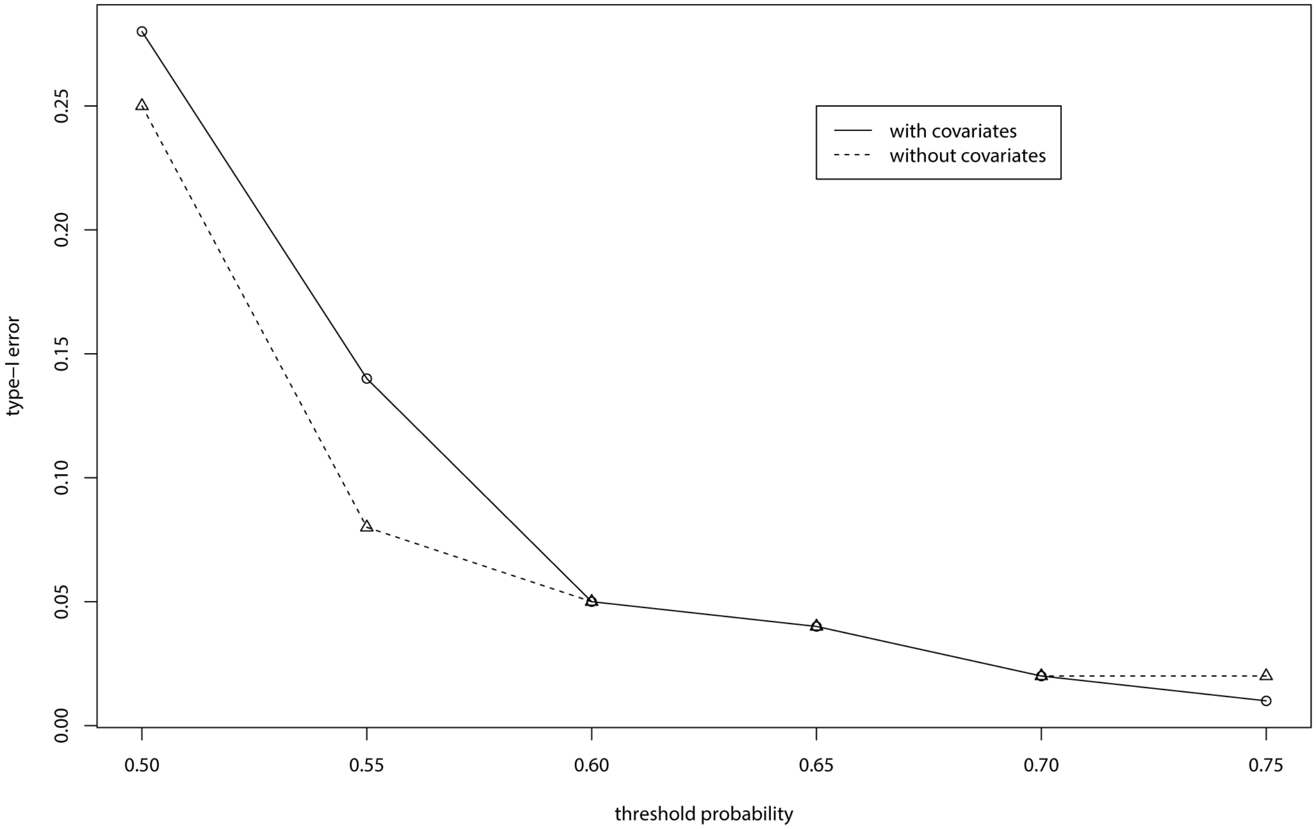

Figure 4. Type-I errors of BRVD for the simulated examples.
doi:10.1371/journal.pone.0069633.g004

Since all chromosomes show positive support for the global association, putative associated variants should be identified from each of them. For illustration, we here work on the first group chromosomes only. Figure 7 illustrates the selection of putative associated variants from chromosome 13. At a FDR level of 0.25, 24 variants were identified from this chromosome. In the same procedure, 42, 32, and 39 variants were identified from chromosomes 2, 3, and 19, respectively. Putting all the selected variants together form a selected subset of 137 variants.

The BRVD was then applied to the selected subset of variants with the same setting as described above except for sample space partitioning and Λ. For the selected subset data, Ω_1 was partitioned into $m=299$ equally spaced (in energy values) subregions with $u_1=601$ and $u_{m-1}=899$, and the values of γ^L we tried include 0.5, 0.6, …, 0.9. A smaller value of γ^L was tried here as $P=137$ is very small for the selected subset. At each value of γ^L, the BRVD shows a decisive support to the hypothesis H_1 with the estimate of the posterior probability $\pi(H_1|\mathcal{D})$ being nearly equal to 1. For example, at $\gamma^L=0.5$, the BRVD produced an estimate of $1-3.6\times10^{-75}$ for $\pi(H_1|\mathcal{D})$. As discussed above, this estimate of $\pi(H_1|\mathcal{D})$ can be biased for the global association test. At $\gamma^L=0.5$, the BRVD identified 10 variants as causal variants at the FDR level 0.1, and identified 14 variants as causal variants at the FDR level 0.2. Table 3 shows the 14 variants in the order (from high to low) of their marginal inclusion probabilities. Among the 14 variants, there are two variants with the MAF lower than 1%. The results for other values of γ^L are similar.

Our method is surprisingly successful for this example: A few rare variants identified by it have been verified in the literature. It is reported that SLC1A4 is associated with atherosclerosis [32], TMEM44 regulates low-density lipoprotein receptor (LDLR) levels which in turn is a critical factor in the regulation of blood cholesterol levels [33], GPC6 is associated with breast cancer [34], and schizophrenia and bipolar [35] and PCBP4 is associated with lung cancer [36].

For comparison, BRI and SKAT were also applied to this example. BRI was run for 50,000 iterations for each of the 22 subsets. The outputs show that only chromosome 2 provides "substantial" evidence for the global association with a Bayes factor of 7.1. The Bayes factors for all other chromosomes are less than 1. On chromosome 2, BRI identified three SNPs, rs65245292, rs179455352 and rs28827533, whose marginal Bayes factor are all greater than 10. It is interesting to point out that both SNPs, rs65245292 and rs28827533, have been identified by BRVD as shown in Table 3. Although the SNP rs179455352 is not included in Table 3, it has been selected by BRVD in the parallel conquering step.

SKAT produced a small p-value for each of the 22 subsets, ranging from 2.3×10^{-7} (chromosome 12) to 0.0016 (chromosome 21). According to the p-values, all chromosomes are associated with heart, lung and blood diseases. This result suggests that SKAT may be liberal in global association tests. To explore the relationship between the p-value and the chromosome length, we plot in Figure 8(b) the scatterplot of $\Phi^{-1}(1-p_i)$ versus $\log(L_i)$, where p_i denotes the p-value of chromosome i, L_i denotes the length of chromosome i, and Φ denotes the CDF of the standard normal distribution. The scatterplot indicates that SKAT tends to produce a smaller p-value for a longer chromosome; that is, it tends to be sensitive to the proportion of causal variants.

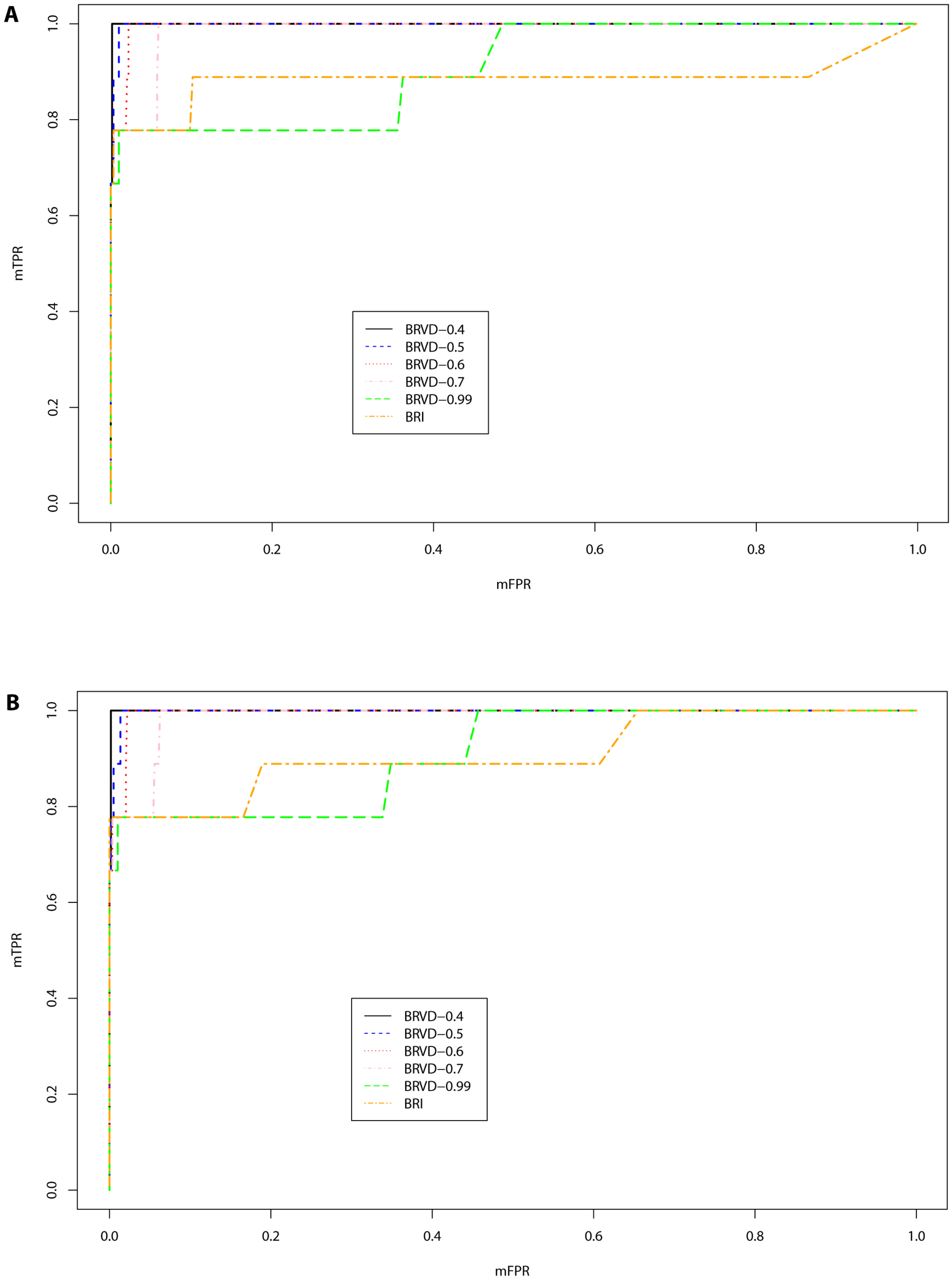
A
mTPR
0.0
0.2
0.4
0.6
0.8
1.0
mFPR
0.0
0.2
0.4
0.6
0.8
1.0
BRVD−0.4
BRVD−0.5
BRVD−0.6
BRVD−0.7
BRVD−0.99
BRI
B
mTPR
0.0
0.2
0.4
0.6
0.8
1.0
mFPR
0.0
0.2
0.4
0.6
0.8
1.0
BRVD−0.4
BRVD−0.5
BRVD−0.6
BRVD−0.7
BRVD−0.99
BRI

Figure 5. Marginal ROC curves for BRVD and BRI: Left panel: ROC curves with covariates adjusted; right panel: ROC curves with covariates omitted.
doi:10.1371/journal.pone.0069633.g005

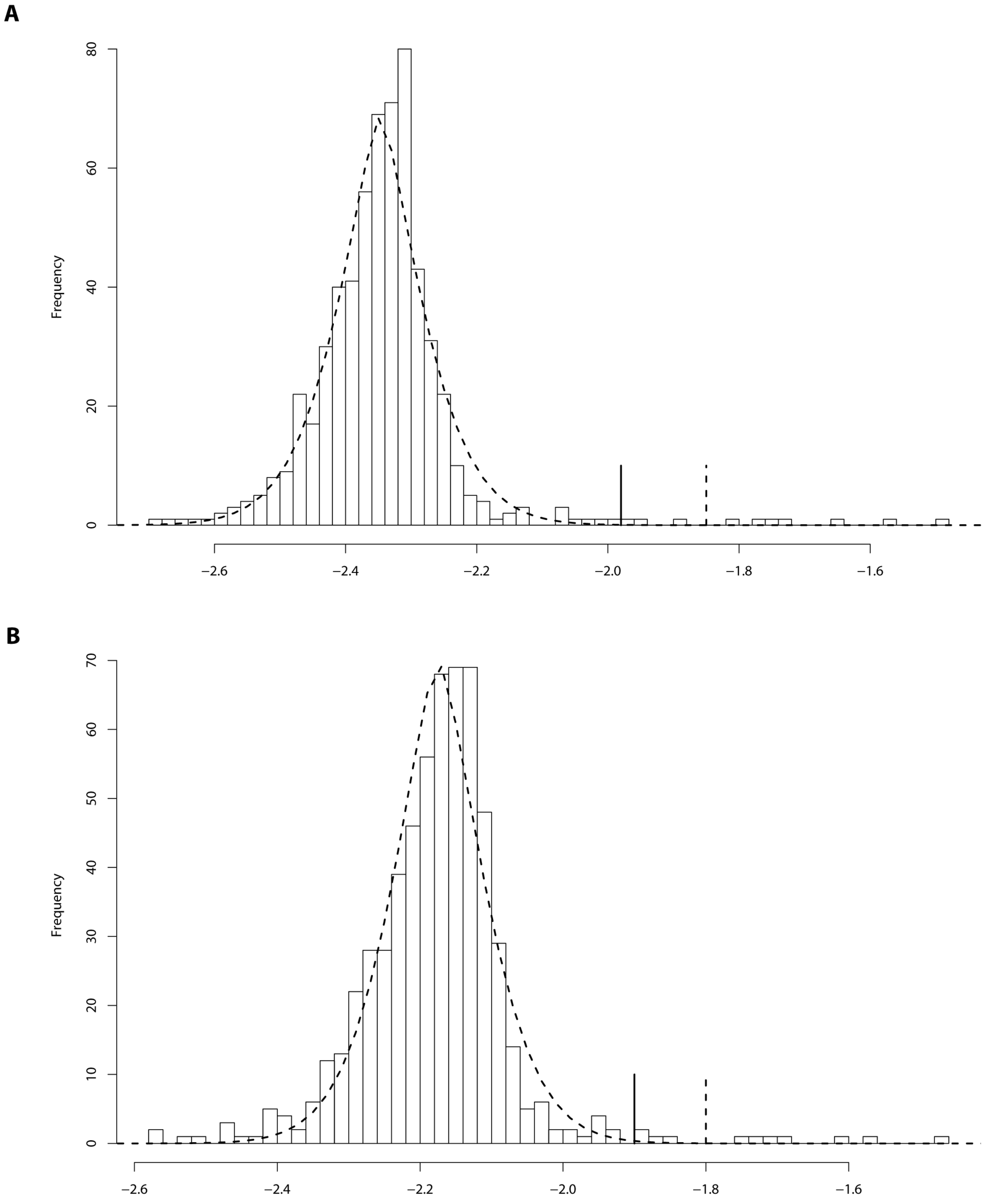

Figure 6. Illustrative plot for causal rare variants detection. The dashed curve shows the fitted density function for the marginal inclusion scores of non-associated variants, and the vertical bar shows the classification rules at the FDR level 0.05 (solid line) and the FDR level 0.01 (dashed line). The left panel is for $\gamma^L = 0.6$ and the right panel is for $\gamma^L = 0.5$.
doi:10.1371/journal.pone.0069633.g006

Table 2. BRVD results for the EOMI data.

Chromosome	size	γ^L	mean	SD
13	1811	0.9	0.9516	0.0046
2	8383	0.8	0.8059	0.0079
3	6534	0.9	0.7356	0.0080
19	8216	0.9	0.7069	0.0016
other	1271~12491	—	0.5~0.57	—

Size: the number of variants included in each chromosome; γ^L: the selected value of γ^L; mean: $\bar{\pi}(H_1|D_i,\gamma^L)$, i.e., the average value of $\pi(H_1|D_i,\gamma^L)$ over five independent runs at the selected value of γ^L; SD: standard deviation of $\bar{\pi}(H_1|D_i,\gamma^L)$.

doi:10.1371/journal.pone.0069633.t002

Similarly, we plot in Figure 8(a) the scatterplot of $\Phi^{-1}(\max_{\gamma^L\in\Lambda}\bar{\pi}(H_1|\mathcal{D}_i,\gamma^L))$ versus $\log(L_i)$ for BRVD, where $\mathcal{D}_i$ denotes the subset corresponding to chromosome i; and plot in Figure 8(c) the scatterplot of $\Phi^{-1}(\pi(H_1|\mathcal{D}_i))$ versus $\log(L_i)$ for BRI, where $\pi(H_1|\mathcal{D}_i)$ is calculated from the Bayesian factor with the prior probabilities $\pi(H_0)=\pi(H_1)=1/2$. Although BRI is not as sensitive to the chromosome length as SKAT, its results suggest that it is pretty conservatives in global association tests. As discussed above, the literature results show that chromosome 3 and chromosome 13 are also associated with heart, lung and blood diseases, but BRI failed to identify these associations. In summary, the comparison implies that BRVD outperforms both SKAT and BRI for this real-data example.

Computational time

The computation time for the BRVD depends on the sample size (n) and the number of variants (P). Table 4 recorded the CPU time cost by BRVD on an Intel Xeon E5-2690 processor for running 10^5 iterations under different settings of n and P. A linear regression analysis of the CPU time versus n and P produces a R^2 of 99.76%, which indicates an adequate fitting of the regression. Both P and n are significant for the regression, and their p-values are 4.9×10^{-6} and 7.4×10^{-4}, respectively. Figure 9 plots the CPU time of BRVD versus P for the EOMI data (with $n=905$). It indicates a strong linear relationship between the CPU time and P. Since the number of iterations is usually set to be proportional to the value of P, this analysis implies that the CPU time of the BRVD can increase as a quadratic function of P.

In analyzing the CPU time of BRVD, we fixed γ^L to 0.9. We note that the CPU time of BRVD can slightly increase as γ^L decreases for fixed values of n and P, because a smaller value of γ^L tends to result in a larger model. However, the effect of γ^L is not significant, because, under the control of multiplicity, the sizes of the selected models are always tiny compared to the value of P. The CPU time of the BRVD is dominated by the part of data scanning that needs to be performed for each iteration.

Discussion

In this paper, we have developed a new Bayesian method, the so-called BRVD, for detection of causal variants. The BRVD simultaneously addresses two issues: (i) Are there any of the variants associated with the disease, and (ii) Which variants, if any, are driving the association. The BRVD is developed based on the theory of posterior consistency, under which the causal variants can be identified consistently. The numerical results indicate that the BRVD is more powerful for global association tests than the

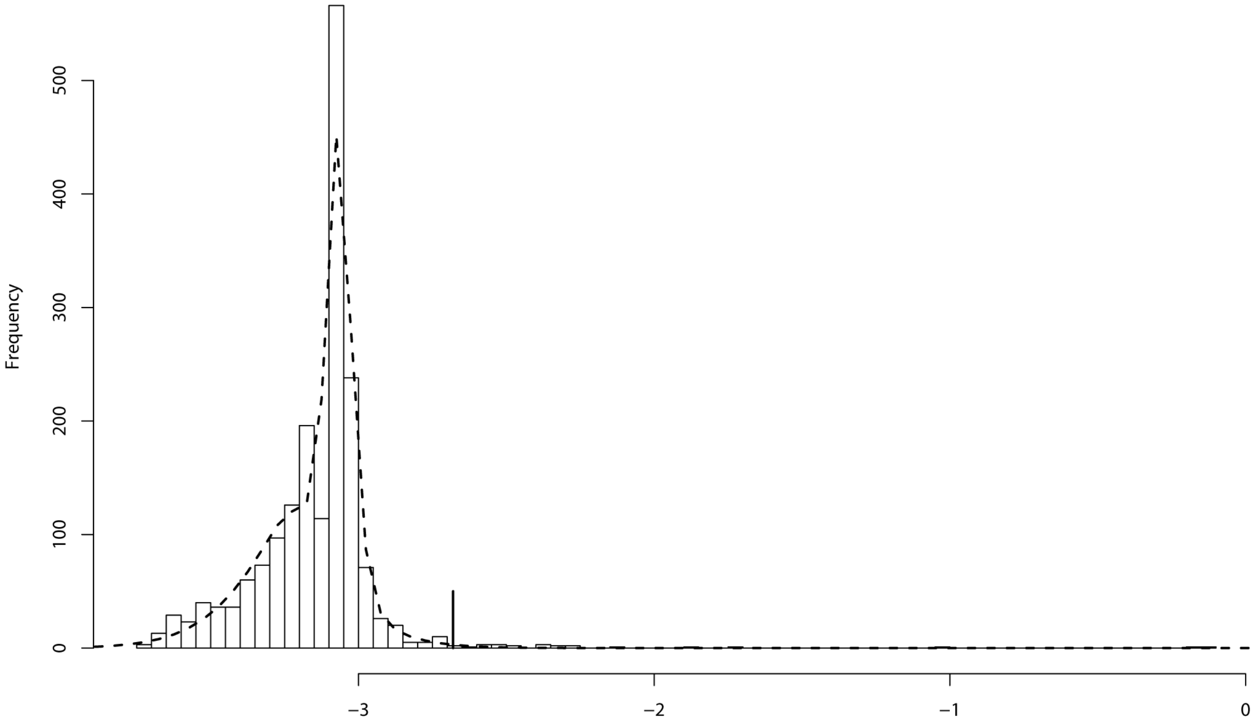

Figure 7. Variant selection from Chromosome 13 for the EOMI data: The dashed curve shows the fitted density function for the marginal inclusion scores of non-associated variants, and the vertical bar shows the classification rules at the FDR level 0.25.
doi:10.1371/journal.pone.0069633.g007

Table 3. Top 14 variants identified by BRVD for the EOMI data at a FDR level of 0.2.

No.	Variant	Gene	Chrom	MAF	No.	Variant	Gene	Chrom	MAF
1	rs65245292	SLC1A4	2	1.38%	8	rs194325058	TMEM44	3	3.26%
2	rs194408716	FAM43A	3	2.54%	9	rs28827533	PLB1	2	1.05%
3	rs39586979	C13orf23	13	2.76%	10	rs19961331	EFHB	3	4.81%
4	rs39424253	FREM2	13	3.76%	11	rs94197611	GPC6	13	0.99%
5	rs51994587	PCBP4	3	1.33%	12	rs128695828	NO-Gene	3	1.49%
6	rs39424254	FREM2	13	3.76%	13	rs242610172	ATG4B	2	1.55%
7	rs549728	GZMM	19	2.38%	14	rs57867517	ZNF304	19	0.94%

doi:10.1371/journal.pone.0069633.t003

existing methods, such as CMC, WSS, SKAT, C-alpha, RARECOVER, VT, and BRI, and also more powerful for detection of causal variants than the BRI method. In this paper, we have also developed a parallel version of BRVD based on the strategy of divide-and-conquer. The parallel BRVD can be conveniently used for the datasets for which the number of variants is extremely large.

Since the BRVD is developed under the framework of logistic regression, it can be directly applied to identify gene-gene and gene-environment interactions by including in the model some interaction terms of SNP-SNP and SNP-covariates. A gene-gene and/or gene-environment interaction network can then be constructed. This method is very flexible, depending on the specification of interaction terms. For example, to explore complex higher-order interactions, a partially linear tree-based regression model [37] may be used.

Although BRVD has a high power for both the global association tests and causal variants detection, its power can be further improved by employing a more sophisticated weighting scheme for the variants. The current weighting scheme depends on the MAF only. In the future, one may incorporate other biological information, e.g., the gene information, into the weighting scheme. This may help further to identify the causal variants whose MAFs are extremely low. In the current

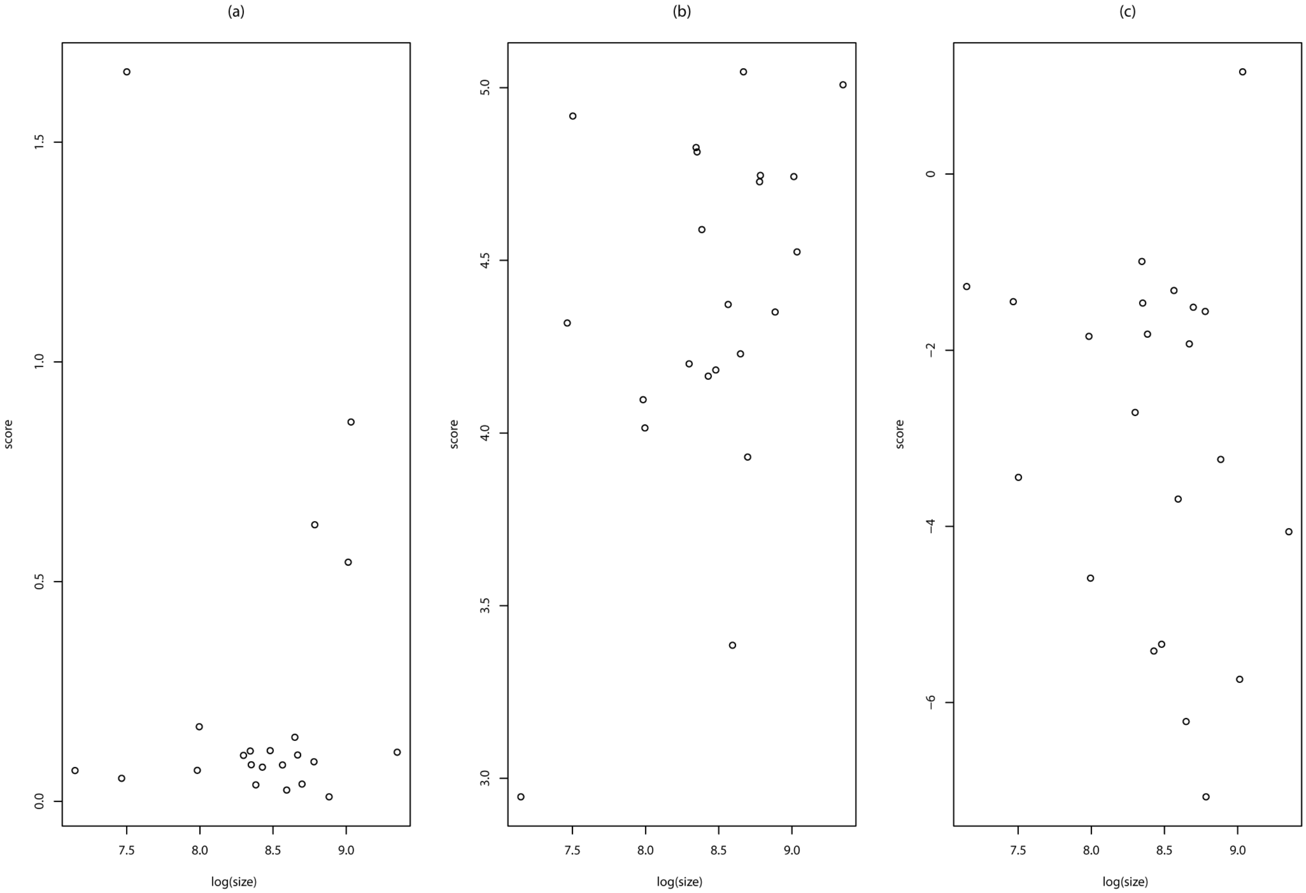

Figure 8. Significance of global association tests versus chromosome length for the EOMI data: (a) BRVD; (b) SKAT; and (c) BRI.
doi:10.1371/journal.pone.0069633.g008

Table 4. CPU time cost by the BRVD on an Intel Xeon E5-2690 processor (2.9 GHz) for running 10^5 iterations.

Case	*n*	*P*	CPU(s)
1	500	600	7.67
2	905	1,271	17.21
3	905	1,811	19.02
4	905	6,534	26.99
5	905	8,383	29.88
6	905	11,491	34.73
7	905	24,944	53.82

n: sample size; *P*: number of variants.
doi:10.1371/journal.pone.0069633.t004

implementation of the BRVD, the SAMC algorithm is used for sampling from the posterior. At each iteration, a variant is randomly selected to undergo a model update of variant addition, deletion, or exchange. In the future, a SAMC algorithm with an adaptive proposal may be used. The new version of SAMC allows one to select a variant for model update based on the working estimate of marginal inclusion probabilities. In the limit case, the new version of SAMC will update the model according to the marginal inclusion probabilities of all variants. Therefore, it can converge faster than the standard version of SAMC.

For global association tests, the BRVD can also be used in conjunction with other frequentist methods, such as SKAT, if one is interested in a *p*-value measurement for the significance of the test. One can first apply the BRVD to select a subset of variants and then conduct the association test on the selected subset of variants using the frequentist method. Since all the existing rare variant testing methods seem to be sensitive to the proportion of causal variants [38], the combined use of the BRVD and frequentist methods can generally reduce the sensitivity of the test methods to the proportion of causal variants.

The BRVD is general in the sense that it can be used for rare variants, common variants, and also a joint analysis of common and rare variants. In the case of joint analysis, its power for detecting rare variants will not be affected much if γ_i in (9) is chosen appropriately as an increasing function of MAF. We note that in the literature some other Bayesian variable selection methods have also been developed and can potentially be used for variant selection [39,40,41]. However, none of these methods is directly comparable with BRVD. The method [39] is developed for linear regression under the framework of large-*n*-small-*P*, and thus cannot be applied to the small-*n*-large-*P* logistic regression problems considered in this paper. The method [40] is developed for linear regression, although for the small-*n*-large-*P* problems; hence, it cannot be compared with BRVD for logistic regression. The method [41] aims to identify biomarkers, for which the model incorporates the biological information on known pathways and gene-gene networks. Since these information are not available for the problems considered in this paper, this method cannot be directly compared with BRVD. Also, we note that although BRVD and the methods [40,41] are all applicable to the small-*n*-large-*P* problems, BRVD has a theoretical advantage over the other two methods: BRVD is consistent, i.e., the causal variables

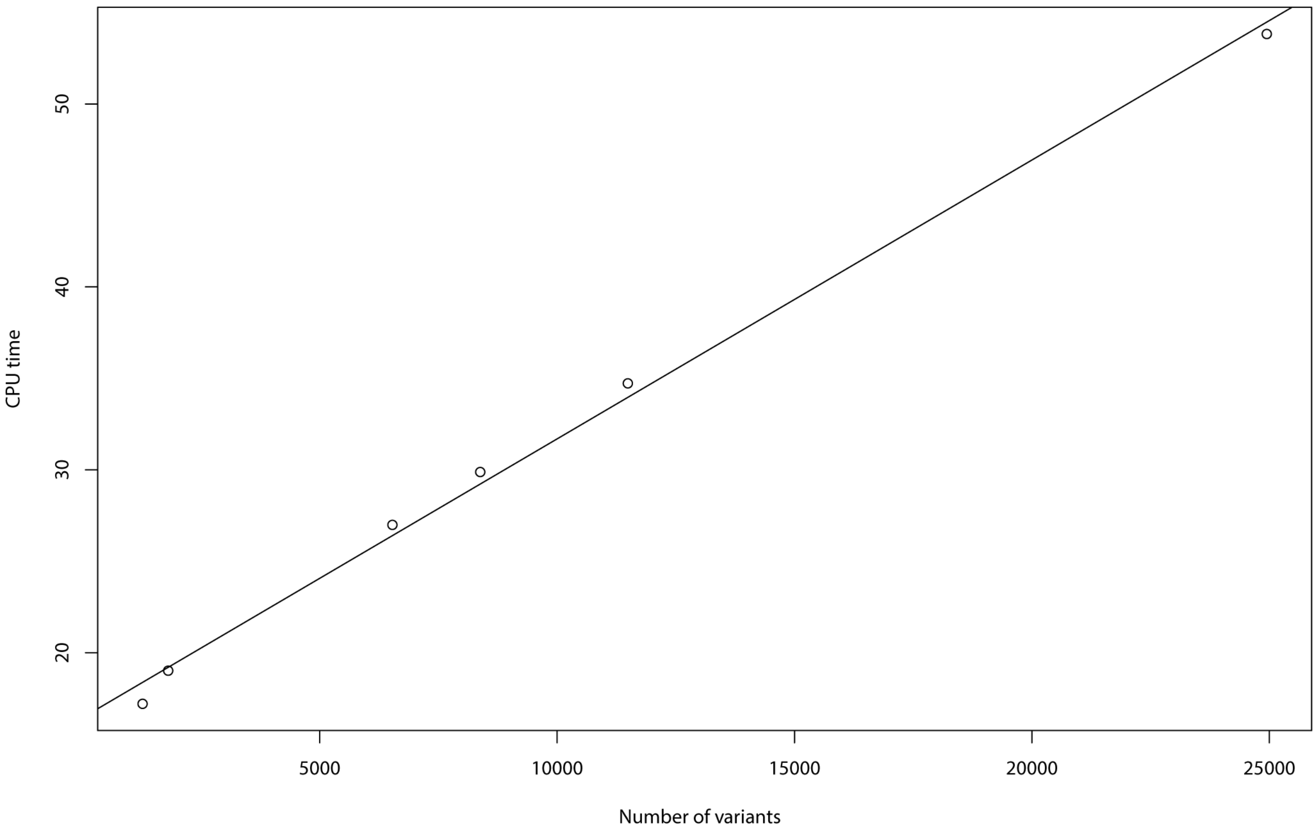

Figure 9. The CPU time of BRVD versus the number of variants for the EOMI data (with $n = 905$).
doi:10.1371/journal.pone.0069633.g009

can be identified by it in probability 1 as the sample size $n \to \infty$; while this is unclear for the other two methods.

In this paper, BRVD is developed for dichotomous phenotypes only. The framework of BRVD can be easily extended to continuous phenotypes. For continuous phenotypes, linear regression can be used to relate the phenotype to the variants, and appropriate prior distributions that lead to the posterior consistency need to be specified for the model and model specific parameters. Alternatively, one can impose a non-local prior on the model parameters as in [42]. Under the non-local prior, it can be shown that the causal variants can be consistently identified if the total number of variants is bounded by the number of subjects.

Supporting Information

File S1 Supporting Information.
(PDF)

Author Contributions

Conceived and designed the experiments: FL. Performed the experiments: FL. Analyzed the data: FL MX. Contributed reagents/materials/analysis tools: FL MX. Wrote the paper: FL.

References

1. Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40: 695–701.
2. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324: 387–389.
3. Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, et al. (2006) Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc Natl Acad Sci USA 103: 1810–1815.
4. Li B, Leal SM (2008) Methods for detecting associations with rare variants for common disease: application to analysis of sequence data. Am J Hum Genet 83: 311–321.
5. Madsen E, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLOS Genet 5: e1000384. Available: http://www.plosgenetics.org/article/info%3Adoi %2F10.1371%2Fjournal.pgen.1000384. Accessed 2013 Feb 28.
6. Wu MC, Lee S, Cai T, Li Y, Boehnke M, et al. (2011) Rare-variant association testing for sequence data with the sequence kernel association test. Am J Hum Genet 89: 82–93.
7. Han F, Pan W (2010) A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered 70: 42–54.
8. Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, et al. (2010) Extending rare-variant testing strategies: Analysis of noncoding sequence and imputed genotypes. Am J Hum Genet 87: 604–617.
9. Bhatia G, Bansal V, Harismendy O, Schork NJ, Topol EJ, et al. (2010) A covering method for detecting genetic associations between rare variants and common Phenotypes. PLoS Comput Bio 6:e1000954s. Available: http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000954. Accessed 2013 Feb 28.
10. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, et al. (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86: 832–838.
11. King CR, Rathouz PJ, Nicolae DL (2010) An evolutionary framework for association testing in resequencing studies. PLoS Genet 6: e1001202. Available: http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1001202. Accessed 2013 Feb 28.
12. Yi N, Liu N, Zhi D, Li J (2011) Hierarchical generalized linear models for multiple groups of rare and common variants: Jointly estimating group and individual-variant effects. PLoS Genet 7: e1002382. Available: http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1002382. Accessed 2013 May 20.
13. Yi N, Zhi D (2011) Bayesian analysis of rare variants in genetic association studies. Genet Epidemiol 35: 57–69.
14. Quintana MA, Berstein JL, Thomas DC, Conti DV (2011) Incorporating model uncertainty in detecting rare variants: The Bayesian risk index. Genet Epidemiol 35: 638–649.
15. Wilson MA, Iversen ES, Clyde MA, Schmidler SC, Schildkraut JM (2010) Bayesian model search and multilevel inference for SNP association studies. Ann Appl Statist 4: 1342–1364.
16. Jeffreys H (1961) Theory of probability (3rd edition). Oxford: Oxford University Press. 470 p.
17. Berger JO (1985) Statistical decision theory and Bayesian analysis. New York: Springer. 617 p.
18. Berger JO, Sellke T (1987) Testing a point null hypothesis: The irreconcilability of *p* values and evidence. J Amer Statist Assoc 82: 112–122.
19. Jiang W (2006) On the consistency of Bayesian variable selection for high dimensional binary regression and classification. Neural Comput 18: 2762–2776.
20. Jiang W (2007) Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities. Ann Statist 35: 1487–1511.
21. Scott JG, Berger JO (2010) Bayes and empirical-Bayes multiplicity adjustment in the variable selection problem. Ann Statist 38: 2587–2619.
22. Liang F, Liu C, Carroll RJ (2007) Stochastic approximation in Monte Carlo computation. J Amer Statist Assoc 102: 305–320.
23. Chen HF (2002) Stochastic approximation and its applications. Dordrecht: Kluwer Academic Publishers. 357 p.
24. Andrieu C, Moulines É, Priouret P (2005) Stability of Stochastic Approximation Under Verifiable Conditions. SIAM J Control Optim 44: 283–312.
25. Barbieri MM, Berger JO (2004) Optimal Predictive Model Selection. Ann Statist 32: 870–897.
26. Liang F, Song Q, Yu K (2013) Bayesian subset modeling for high dimensional generalized linear models. J Amer Statist Assoc. In press. doi:10.1080/01621459.2012.761942.
27. Liang F (2009) On the use of stochastic approximation Monte Carlo for Monte Carlo integration. Stat Prob Lett 79: 581–587.
28. Liang F, Zhang J (2008) Estimating the false discovery rate using the stochastic approximation algorithm. Biometrika 95: 961–977.
29. Neuhaus JM (1998) Estimation efficiency with omitted covariates in generalized linear models. J Amer Statist Assoc 93: 1124–1129.
30. Xing G, Xing C (2010) Adjusting for covariates in logistic regression models. Genet Epidemiol 34: 769–771.
31. Pirinen M, Donnelly P, Spencer CC (2012) Including known covariates can reduce power to detect genetic effects in case-control studies. Nat Genet 44: 848–851.
32. Inouye M, Ripatti S, Kettunen J, Lyytikäinen LP, Oksala N, et al. (2012) Novel Loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. PLoS Genet 8: e1002907. Available: http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1002907. Accessed 2013 Feb 28.
33. Do HT, Tselykh TV, Mäkelä J, Ho TH, Olkkonen VM, et al. (2012) Fibroblast growth factor-21 (FGF21) regulates low-density lipoprotein receptor (LDLR) levels in cells via the E3-ubiquitin ligase Mylip/Idol and the Canopy2 (Cnpy2)/Mylip-interacting saposin-like protein (Msap). J Biol Chem 287: 12602–12611.
34. Eriksson N, Benton GM, Do CB, Kiefer AK, Mountain JL, et al. (2012) Genetic variants associated with breast size also influence breast cancer risk. BMC Med Genet 13: 53. Available: http://www.biomedcentral.com/1471-2350/13/53. Accessed 2013 Feb 28.
35. Wang KS, Liu XF, Aragam N (2010) A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder. Schizophr Res 124: 192–199.
36. Pio R, Blanco D, Pajares MJ, Aibar E, Durany O, et al. (2010) Development of a novel splice array platform and its application in the identification of alternative splice variants in lung cancer. BMC Genom, 11:352. Available: http://www.biomedcentral.com/1471-2164/11/352. Accessed 2013 Feb 28.
37. Chen J, Yu K, Hsing A, Therneau TM (2007) A partially linear tree-based regression model for assessing complex joint gene-gene and gene-environment effects. Genet Epidemiol 31: 238–251.
38. Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CMT, Richards JB (2012) The empirical power of rare variant association methods: Results from sanger sequencing in 1,1998 individuals. PLoS Genet 8: e1002496. Available: http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1002496. Accessed 2013 Feb 28.
39. Liang F, Paulo R, Molina G, Clyde MA, Berger JO (2008) Mixtures of *g* priors for Bayesian variable selection. J Amer Statist Assoc 103: 410–423.
40. Guan Y, Stephens M (2011) Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Ann Appl Statist 5: 1780–1815.
41. Stingo FC, Chen YA, Tadesse MG, Vannucci M (2011) Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. Ann Appl Statist 5: 1978–2002.
42. Johnson VE, Rossell D (2012) Bayesian model selection in high-dimensional settings. J Amer Statist Assoc 107: 649–660.