RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

# Post hoc Analysis for Detecting Individual Rare Variant Risk Associations Using Probit Regression Bayesian Variable Selection Methods in Case-Control Sequencing Studies

Nicholas B. Larson,[1]* Shannon McDonnell,[1] Lisa Cannon Albright,[2] Craig Teerlink,[2] Janet Stanford,[3] Elaine A. Ostrander,[4] William B. Isaacs,[5] Jianfeng Xu,[6] Kathleen A. Cooney,[7,8] Ethan Lange,[9] Johanna Schleutker,[10] John D. Carpten,[11] Isaac Powell,[12] Joan Bailey-Wilson,[13] Olivier Cussenot,[14] Geraldine Cancel-Tassin,[14] Graham Giles,[15,16] Robert MacInnis,[15,16] Christiane Maier,[17] Alice S. Whittemore,[18] Chih-Lin Hsieh,[19] Fredrik Wiklund,[20] William J. Catolona,[21] William Foulkes,[22,23] Diptasri Mandal,[24] Rosalind Eeles,[25] Zsofia Kote-Jarai,[25] Michael J. Ackerman,[26] Timothy M. Olson,[26] Christopher J. Klein,[27] Stephen N. Thibodeau,[28] and Daniel J. Schaid[1]

*[1]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, United States of America; [2]Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, Utah, United States of America; [3]Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America; [4]National Human Genome Research Institute, Bethesda, Maryland, United States of America; [5]Department of Urology, Johns Hopkins Hospital, Baltimore, Maryland, United States of America; [6]NorthShore University HealthSystem Research Institute, Chicago, Illinois, United States of America; [7]Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, Michigan, United States of America; [8]Department of Urology, University of Michigan Medical School, Ann Arbor, Michigan, United States of America; [9]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, United States of America; [10]Department of Medical Biochemistry and Genetics, Institute of Biomedicine, University of Turku, Finland; [11]Integrated Cancer Genomics Division, The Translational Genomics Research Institute, Phoenix, Arizona, United States of America; [12]Wayne State University, Detroit, Michigan, United States of America; [13]Statistical Genetics Section, National Human Genome Research Institute, Bethesda, Maryland, United States of America; [14]CeRePP, Hopital Tenon, Paris, France; [15]Cancer Epidemiology Centre, Cancer Council Victoria, University of Melbourne, Melbourne, Australia; [16]Centre for Epidemiology and Biostatistics, School of Population and Global Health, University of Melbourne, Melbourne, Australia; [17]Department of Urology, University of Ulm, Ulm, Germany; [18]Department of Health Research and Policy, Stanford University, Stanford, California, United States of America; [19]Department of Urology, University of Southern California, Los Angeles, California, United States of America; [20]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [21]Feinberg School of Medicine, Northwestern University, Chicago, Illinois, United States of America; [22]Department of Oncology and Human Genetics, Montreal General Hospital, Montreal, QC, Canada; [23]Department of Human Genetics, Montreal General Hospital, Montreal, QC, Canada; [24]Department of Genetics, LSU Health Sciences Center, New Orleans, Louisiana, United States of America; [25]Genetics and Epidemiology, Institute of Cancer Research, Sutton, Surrey, United Kingdom; [26]Department of Pediatric and Adolescent Medicine, Mayo Clinic, Rochester, Minnesota, United States of America; [27]Department of Neurology, Mayo Clinic, Rochester, Minnesota, United States of America; [28]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, United States of America*

Received 16 December 2015; Revised 22 April 2016; accepted revised manuscript 27 April 2016.
Published online 17 June 2016 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21983

**ABSTRACT:** Rare variants (RVs) have been shown to be significant contributors to complex disease risk. By definition, these variants have very low minor allele frequencies and traditional single-marker methods for statistical analysis are underpowered for typical sequencing study sample sizes. Multimarker burden-type approaches attempt to identify aggregation of RVs across case-control status by analyzing relatively small partitions of the genome, such as genes. However, it is generally the case that the aggregative measure would be a mixture of causal and neutral variants, and these omnibus tests do not directly provide any indication of which RVs may be driving a given association. Recently, Bayesian variable selection approaches have been proposed to identify RV associations from a large set of RVs under consideration. Although these approaches have been shown to be powerful at detecting associations at the RV level, there are often computational limitations on the total quantity of RVs under consideration and compromises are necessary for large-scale application. Here, we propose a computationally efficient alternative formulation of this method using a probit regression approach specifically capable of simultaneously analyzing hundreds to thousands of RVs. We evaluate our approach to detect causal variation on simulated data and examine sensitivity and specificity in instances of high RV dimensionality as well as apply it to pathway-level RV analysis results from

*Correspondence to: Nicholas B. Larson, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905. E-mail: Larson.nicholas@mayo.edu

**a prostate cancer (PC) risk case-control sequencing study. Finally, we discuss potential extensions and future directions of this work.**
Genet Epidemiol 40:461–469, 2016. 

**KEY WORDS**: Next-generation sequencing; MCMC; prostate cancer; burden testing

## Introduction

With advancements in next-generation sequencing technologies, there has been a reinvigorated interest in the roles that rare variants (RVs) play in the genetic etiology of complex diseases [Cirulli and Goldstein, 2010]. Due to low minor allele frequencies (MAFs), traditional single-variant risk association analysis methods on RVs suffer from low statistical power for even relatively large sample sizes, and specialized strategies are necessary to identify RV associations. This has led to the development of multimarker aggregation strategies that are predicated on the notion that causal RVs may cluster in biologically relevant functional domains, such as genes [Bansal et al., 2010]. There are a growing number of multimarker omnibus methods available for RV association analysis that evaluate a priori defined target regions of interest (ROIs) to localize clustering of causal RVs. These include various burden-based collapsing methods [Dering et al., 2011], as well as variance component tests such as the C-alpha test [Neale et al., 2011] and sequence kernel association test (SKAT) [Lee et al., 2012; Wu et al., 2011].

A notable caveat for these omnibus tests is that they do not provide any inference at the marker level as to which RVs may be driving a given multimarker association. An alternative strategy is to simultaneously assess all of the RVs under consideration and apply some form of variable selection. One approach to identifying these RVs is to apply Bayesian variable selection procedures (for review, see O'Hara and Sillanpaa [2009]). Use of these methods in marker association studies have the potential to be more powerful than other model selection procedures [Quintana and Conti, 2013; Wilson et al., 2010], and additionally provide relevant posterior quantities of interest for variable inclusion. Recently, Bayesian model uncertainty (BMU) strategies have been proposed for RV association analysis in case-control studies, referred to as the Bayesian risk index (BRI) [Quintana et al., 2011]. The BRI method utilizes an aggregation and collapsing risk index parameterization of the selected RVs in a logistic regression framework, which we hereafter refer to as L-BRI. The authors' simulation results not only indicate increased power over traditional omnibus approaches for global association, but powerful detection of individual RVs driving an association signal through the derivation of marginal Bayes Factors (BFs).

A drawback of selecting the logit link function for the generalized linear model is that no closed-form solutions exist for the full conditional densities of the model parameters. Moreover, the Metropolis-Hastings (MH) algorithm for sampling from the model space in L-BRI applies a single-component proposal procedure to the variable inclusion vector. This can result in a computationally intensive algorithm requiring many hours to run to fully explore the model space for higher RV counts, reserving practical applications to smaller regions of the genome. Recent findings from large-scale sequencing studies indicate that, from a population-based perspective, RV sites can be quite common [Nelson et al., 2012]. Consequently, sufficient sample size and sequence content could yield a computationally burdensome quantity of RVs for the L-BRI method. An illustrative example of potentially high RV dimensionality is a targeted sequencing study of the *DISC1* locus investigating association with psychiatric traits [Thomson et al., 2014], which identified over 2000 validated RVs (MAF < 1%) across the ROI. Moreover, most sequencing studies are underpowered for gene-based analyses, prompting multigenic analyses that aggregate RVs across related genes in a given pathway [Wu and Zhi, 2013]. Targeted analysis of multiple genes within a gene set could yield similarly extreme quantities of RVs. These applications may not be tenable for the L-BRI or similar approaches without application of strict exclusion criteria that could inadvertently filter out causal variation.

An alternative strategy to handling high dimensional RV analysis would be to apply Bayesian variable selection in a post hoc fashion to identify potential causal variation driving an association finding from frequentist testing. One reformulation of the BRI approach would be to instead utilize the probit link function for the generalized linear model in combination with alternative MH algorithms that permit effective exploration of the model space. A key advantage of the Bayesian probit regression model is that closed forms of the full conditional distributions exist for appropriately selected conjugate priors using data augmentation techniques [Tanner and Wing, 1987], resulting in efficient Gibbs sampling (GS). The use of probit regression with Bayesian variable selection methods for high-dimensional modeling has been demonstrated to be quite powerful in the analysis of gene expression [Baragatti, 2011; Lee et al., 2003; Leon-Novelo et al., 2012; Yang and Song, 2010], capable of simultaneous consideration of hundreds to thousands of probesets. The utility of the probit regression approach relative to logistic regression for variant analysis in case-control sequencing studies was recently demonstrated by Kang et al. [2014].

Here, we propose a fully Bayesian probit regression BRI (P-BRI) method for detection of individual RV risk associations and define strategies for instances of high variant dimensionality. We outline the basic sampling algorithm, which is an adaptation of existing Bayesian variable selection procedures for probit regression. We then evaluate the power of our approach at detecting causal RVs via simulation studies, detailing sensitivity and specificity under varying conditions against L-BRI, as well as apply P-BRI to high dimensional

variant scenarios. To illustrate our method using real data, we apply our P-BRI approach to a prostate cancer (PC) case-control whole-exome sequencing (WES) analysis of the previously detected RV pathway associations. Finally, we discuss the advantages of our approach and outline extensions and future research directions.

## Methods

### Model Definition

Consider a case-control rare variation association study with $N$ subjects consisting of $N_D$ cases and $N_C$ controls, and let $\boldsymbol{Y}$ be an $N \times 1$ vector of corresponding binary responses indicating affected status, such that $Y_i = 1$ if the $i$th subject is a case and $Y_i = 0$ if a control. Let $\boldsymbol{Z}$ be an $N \times p$ RV genotype matrix, where $z_{ij} \equiv \boldsymbol{Z}[i, j]$ represents the minor allele count for subject $i$ at RV position $j$ for $j = 1, \ldots p$. We also define the $N \times q$ design matrix $\boldsymbol{X}$ consisting of $q$ additional adjustment covariates, such as age or gender. In general, it is assumed that the proportion of truly causal RVs in $\boldsymbol{Z}$ is relatively small and that some form of model selection is desired to identify a subset of the total RVs that are associated with the trait of interest. For our approach, we apply variable selection on the set of RVs in $\boldsymbol{Z}$ to characterize an RV load defined by the selected RVs. As such, each possible model $\mathcal{M}_{\boldsymbol{\gamma}}$ within the model space $\mathcal{M}$ can be characterized through a variable inclusion vector $\boldsymbol{\gamma}$, a $p \times 1$ vector of indicators such that $\gamma_j = 1$ denotes that the $j$th RV is included in the aggregation measure, yielding $2^p$ total possible models. For even moderate values of $p$, enumeration of all $2^p$ models $\mathcal{M}_{\boldsymbol{\gamma}} \in \mathcal{M}$ is not feasible.

To account for the effects of RVs on disease risk, we apply a risk index approach that considers the aggregate effect of multiple RVs by the collapsed measure $z_{\gamma,i} = \boldsymbol{Z}_i'\boldsymbol{\gamma}$, where $\boldsymbol{Z}_i$ is a column vector corresponding to the $i$th row of $\boldsymbol{Z}$. The scalar quantity $z_{\gamma,i}$ is the summation of minor alleles over the selected RVs in the model for subject $i$ and indicates the subject-wise RV burden, and we denote $\boldsymbol{Z}_{\boldsymbol{\gamma}} = (z_{\gamma,1}, \ldots, z_{\gamma,N})'$. We define the binary regression model, such that

$$\Pr\left(Y_i = 1 | \boldsymbol{X}_i, Z_{\gamma,i}\right) = g^{-1}(\eta_i),$$

$$\eta_i = \boldsymbol{X}_i'\boldsymbol{\beta} + z_{\gamma,i}\beta_\gamma,$$

where $g(\mu)$ is a link function and $\eta_i$ denotes the linear predictor. For our approach, we select the probit link, such that $g^{-1}(\mu) = \Phi(\mu)$, where $\Phi(\mu)$ represents the standard Gaussian cumulative probability distribution function. The model likelihood can then be written as:

$$\prod_i [\Phi(\eta_i)]^{y_i} [1 - \phi(\eta_i)]^{1-y_i},$$

which does not initially provide analytical solutions for the model parameter posteriors. However, Albert and Chib [1993] proposed a data augmentation solution to computing probit regression posterior distributions by introducing the additional vector of independent latent variables $\tilde{\boldsymbol{Y}}$ corresponding to $\boldsymbol{Y}$, such that

$$Y_i = \begin{cases} 1 & \text{if } \widetilde{Y}_i > 0 \\ 0 & \text{if } \widetilde{Y}_i \leq 0 \end{cases}$$

and

$$\widetilde{Y}_i | \boldsymbol{X}_i, Z_{\gamma,i} \sim \mathcal{N}\left(\boldsymbol{X}_i'\boldsymbol{\beta} + Z_{\gamma,i}\beta_\gamma, 1\right),$$

where $\mathcal{N}(\mu, \sigma^2)$ indicates a Gaussian distribution with mean $\mu$ and variance $\sigma^2$. Thus, the observed dichotomous variable $\boldsymbol{Y}$ is indicative of the sign of the latent random variable $\tilde{\boldsymbol{Y}}$, which is modeled via linear regression with fixed variance.

### Prior Distributions

We opt for traditional conjugate priors where applicable in order to attain full conditional distributions. We first define the prior distribution on the vector of design covariate parameters, $\boldsymbol{\beta}$, to be a $q$-dimensional multivariate Gaussian distribution such that

$$\boldsymbol{\beta} \sim \mathcal{N}_q\left(0, N\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right),$$

which is a conventional g-prior distribution [Zellner, 1983] in probit regression coefficients for blocked GS. We similarly place a standard Gaussian prior on the BRI coefficient $\beta_\gamma$, such that $\beta_\gamma \sim \mathcal{N}(0, 1)$.

We specify the prior probability of a given model $\mathcal{M}_{\boldsymbol{\gamma}} \in \mathcal{M}$, $\Pr(\mathcal{M}_{\boldsymbol{\gamma}})$ through the individual variable prior inclusion probabilities $\Pr(\gamma_j = 1) = \pi_j$, such that $\Pr(\mathcal{M}_{\boldsymbol{\gamma}}) = \prod_{j=1}^{r} \pi_j^{\gamma_j}(1 - \pi_j)^{1-\gamma_j}$. We define the model probability in this fashion via the assumption that the probabilities that given RVs are included in the model are independent, because low linkage disequilibrium is expected among RV sites [Pritchard, 2001]. The vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)'$ can either reflect no differential prior belief of inclusion, such that $\pi_1 = \ldots = \pi_p = \pi$, or may differ based upon available functional data that informs potential RV functionality in relation to the trait of interest. Similar to Quintana et al. [2011], we specify the default prior on $\gamma_j$ to be $\pi_j = (1 - (\frac{1}{2})^{\frac{1}{p}})$, such that the prior probability of the global null model $\Pr(\mathcal{M}_0) = \Pr(\boldsymbol{\gamma} = 0_{\boldsymbol{p}\times 1}) = \prod_j (1 - \pi_j) = \frac{1}{2}$ to account for the potential of a Type I error as well as render the models equitable in this regard.

### Bayesian Sampling Algorithm

To obtain estimates of the posterior quantities of interest, we apply a Markov Chain Monte Carlo (MCMC) approach [Hastings, 1970], whereby samples from the respective posterior distributions of the model parameters are iteratively drawn using GS and MH methods. To define our sampler, we first must characterize the full conditional distributions of the model parameters, which include $f(\tilde{\boldsymbol{Y}}|\boldsymbol{Y}, \boldsymbol{\beta}, \beta_\gamma, \boldsymbol{\gamma})$, $f(\boldsymbol{\beta}|\tilde{\boldsymbol{Y}}, \beta_\gamma, \boldsymbol{\gamma})$, $f(\beta_\gamma|\tilde{\boldsymbol{Y}}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, and $f(\boldsymbol{\gamma}|\tilde{\boldsymbol{Y}}, \boldsymbol{\beta}, \boldsymbol{\beta}_\gamma)$. The full conditional distributions for the first three can easily be derived, such that

- $\widetilde{Y}_i | Y_i = 1 \sim \mathcal{N}(\boldsymbol{X}_i'\boldsymbol{\beta} + Z_{\gamma,i}\beta_\gamma, 1)$ left truncated at 0.

- $\widetilde{Y}_i | \; Y_i = 0 \sim \mathcal{N}(\boldsymbol{X}_i'\boldsymbol{\beta} + Z_{\gamma,i}\beta_\gamma, 1)$ right truncated at 0.
- $\boldsymbol{\beta} | \tilde{\boldsymbol{Y}}, \alpha, \boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{X}'(\tilde{\boldsymbol{Y}} - \boldsymbol{Z}_{\boldsymbol{\gamma}}\beta_\gamma), \boldsymbol{V}_{\boldsymbol{\beta}})$, where $\boldsymbol{V}_{\boldsymbol{\beta}} = \frac{N}{N+1}(\boldsymbol{X}'X)^{-1}$.
- $\beta_\gamma | \tilde{\boldsymbol{Y}}, \boldsymbol{\beta}, \boldsymbol{\gamma} \sim \mathcal{N}(v_\gamma \boldsymbol{Z}_{\boldsymbol{\gamma}}'(\tilde{\boldsymbol{Y}} - \boldsymbol{X}\beta), v_\gamma)$, where $v_\gamma = \frac{1}{\sigma_\beta^{-2} + \boldsymbol{Z}_{\boldsymbol{\gamma}}'\boldsymbol{Z}_{\boldsymbol{\gamma}}}$.

Because these distributions are properly defined, GS methods can be used for iterative updating. However, under our BMU procedure, the full conditional distribution of $\boldsymbol{\gamma}$ cannot be directly simulated easily, requiring a Metropolis-within-Gibbs approach. To sample from the distribution of $\boldsymbol{\gamma}$, we adopt a marginalization strategy [Liu, 1994], which is based upon the integrated distribution of the full conditional of $\boldsymbol{\gamma}$ over $\beta_\gamma$, $f(\boldsymbol{\gamma} | \tilde{\boldsymbol{Y}}, \boldsymbol{\beta})$. It can be shown using Bayesian linear model theory that $f(\boldsymbol{\gamma} | \tilde{\boldsymbol{Y}}, \boldsymbol{\beta})$ is proportional to:

$$\exp\left(-\frac{1}{2}\left((\tilde{\boldsymbol{Y}} - \boldsymbol{X}\beta)'\left(\boldsymbol{I}_N - \frac{\boldsymbol{Z}_{\boldsymbol{\gamma}}\boldsymbol{Z}'_{\boldsymbol{\gamma}}}{\sigma_\beta^{-2} + \boldsymbol{Z}'_{\boldsymbol{\gamma}}\boldsymbol{Z}_{\boldsymbol{\gamma}}}\right)(\tilde{\boldsymbol{Y}} - \boldsymbol{X}\beta)\right)\right) \times \prod_{j=1}^{p} \pi_j^{\gamma_j}(1-\pi_j)^{1-\gamma_j},$$

which we use to define a MH algorithm for updating $\boldsymbol{\gamma}$, directly followed by simulation of $\beta_\gamma$ from its full conditional distribution.

There are a number of options for proposing new values of $\boldsymbol{\gamma}$ in the MH step of the MCMC sampler. Quintana et al. [2011] elected a single-step addition/deletion MH algorithm for model selection in L-BRI, whereby the proposed vector $\boldsymbol{\gamma}$ is generated by switching the binary value of a randomly chosen variable inclusion indicator $\gamma_j$. However, in instances of higher RV dimensionality, this approach requires a prohibitively large number of iterations to adequately explore the model space $\mathcal{M}$, resulting in relatively poor mixing. In contrast, updating each $\gamma_i$ in a component-wise fashion can significantly improve mixing and convergence and may result in overall better performance [Johnson et al., 2013]. Consequently, we apply a component-wise multistep MH algorithm, similar to that applied by Lee et al. [2014] for imaging data, that iteratively updates each element in $\boldsymbol{\gamma}$. This is conducted in a modified metropolised Gibbs framework, such that the proposal for $\gamma_i$is always the opposite of the current state, yielding more efficient mixing [Liu, 1996]. The unique formulation of the risk index as a product of a fixed design matrix $\boldsymbol{Z}$ and variable inclusion vector $\boldsymbol{\gamma}$ permits computationally efficient component-wise MH updating, which is generally infeasible for high dimensional problems. At each iteration of the MCMC algorithm, we randomize the updating order of MH step for $\boldsymbol{\gamma}$, and convergence to the stationary distribution may be checked by running multiple chains from different initial values and comparing posterior samples.

Given that the defined prior on the BRI parameter $\beta_\gamma$ has positive support over the entire real line, it is possible for the sampler to draw negative values of $\beta_\gamma$ despite it characterizing risk. One simple solution is to constrain the prior distribution to the positive real line by using a truncated normal prior. By Gelfand et al. [1992], we can accommodate this prior by adding a rejection step to the Gibbs sampler for $\beta_\gamma$, accepting new draws of $\beta_\gamma$, $\beta_\gamma{}^{(\star)}$, only if $\beta_\gamma{}^{(\star)} > 0$.

## Posterior Measures of Interest

Conditional on evidence against the global null model $\mathcal{M}_0$ (e.g., from a previously conducted test), a primary motivation is identifying an interesting subset of variants associated with the disease of interest for follow-up analyses. In the case of variable selection problems, the marginal posterior probabilities of inclusion are useful for such inference. Denote $\zeta_j = \Pr(\gamma_j = 1 | \boldsymbol{Y})$ to be the marginal posterior probability of inclusion for $j$th RV in $\boldsymbol{Z}$. Quintana et al. [2011] derive the marginal BFs to isolate RVs that may be driving an association, such that

$$BF\left[\gamma_j = 1 : \gamma_j = 0\right] = \frac{\Pr(\gamma_j = 1 | \boldsymbol{Y})}{\Pr(\gamma_j = 0 | \boldsymbol{Y})} \times \frac{\Pr(\gamma_j = 0)}{\Pr(\gamma_j = 1)} = \frac{\zeta_j}{1-\zeta_j} \times \frac{1-\pi_j}{\pi_j}.$$

We estimate $\zeta_j$in a Monte Carlo fashion from the $T$ posterior samples of $\boldsymbol{\gamma}$, such that $\hat{\zeta}_j = \frac{\sum_r \gamma_j^{(t)}}{T}$. Decisions of relative importance of each RV can then be made with respect to common thresholds (e.g., >10 or 31.6) defined by Jeffreys' [1961] grades of evidence using these marginal BFs.

## Simulations

To evaluate the performance of P-BRI at identifying individual risk associated RVs, we considered a hypothetical case-control genetic association study with $N = 1,000$ total subjects ($N_D = N_C$). To simulate the RV genotype data conditional on disease status, we employed the model developed by Li and Leal [ 2008] and algorithmically defined by Zhou et al. [2010]. This model is based upon the conditional Poisson binomial whereby any of the $\nu$ risk RVs can independently cause the disease status, defined though the MAFs, prevalence, and relative risks of RVs. The MAFs for all RVs were randomly generated uniformly on the interval (0.005, 0.01), and all simulated RVs that resulted in an empirical MAF of zero were excluded from analysis. Prevalence was fixed at 0.01 and no additional covariates were included in the simulation model. Simulations under the null (i.e., no causal variation) simply involved random assignment of case-control status to randomly generated RV genotype vectors.

We first compared the performance of the P-BRI relative to the original L-BRI at detecting causal RVs in scenarios that were computationally reasonable for either method, fixing $p = 50$. Software implementation of the L-BRI method is available via the R package *BVS*, which we applied under default settings unless otherwise noted. For simulations involving causal variation, we considered the quantity of truly causal RVs, $\nu$, to range from 5 to 15, and applied both the P-BRI and L-BRI methods to detect the associated RVs. All causal RVs were attributed a relative risk (RR) of 1.5, 2.5, or 5, with all remaining RVs being neutral (RR = 1). Convergence of the L-BRI was evaluated by running two parallel chains and comparing output marginal BFs, as per the method's documentation, with convergence defined by the root mean square error between the two sets of BFs to be <1. To

evaluate convergence of P-BRI, the Gelman-Rubin diagnostic was applied to MCMC posterior samples of $\beta_\gamma$ for two parallel chains with different starting values, with convergence declared if the upper 95% confidence limit was <2. For the P-BRI method, we sampled a total of 30,000 iterations, treating the first 15,000 as a burn-in, while for the L-BRI method we sampled 100,000 iterations and treated the first 50,000 as a burn-in. If convergence was not achieved at these iteration counts, additional posterior samples were drawn until convergence criteria were met. Marginal BFs were also computed for P-BRI in order to compare the relative false-positive (FPR) and true-positive rates (TPR) based upon detection of causal variant status across all simulation iterations ($50 \times 500 = 25{,}000$ total variants). For P-BRI, instances where RVs had corresponding posterior inclusion probabilities (PIPs) estimates, $\hat{\zeta}_j = 1$ were adjusted to $\hat{\zeta}_j = \frac{T-1}{T}$ to avoid division by zero in the marginal BF. For purposes of comparing performance between L-BRI and P-BRI relative to TPR and FPR, we computed bootstrapped 95% confidence intervals on these metrics and/or their differential across methods using the R package *fbroc*, based upon 1,000 bootstrap samples.

To additionally examine the performance of P-BRI under high RV dimensionality, we increased the total RV counts to $p = 500$ and 1, 000 and fixed the number of true deleterious RVs to $\nu = 25$, such that the causal RV proportions were 5.0% ($p = 500$) and 2.5% ($p = 1000$), respectively. Given the larger quantity of RVs under simultaneous consideration, we focused on identification of larger effect sizes and examined performance for RRs of 2.5, 5, and 10 for causal RVs. For these applications, the first 15,000 MCMC samples were discarded as a burn-in, resulting in a posterior sample size of 15,000.

### Data Application: PC Risk

A WES study of men with PC was conducted by the International Consortium of Prostate Cancer Genetics (ICPCG). The ICPCG has identified and sampled the most informative high-risk PC pedigrees known throughout the world. With the goal of identifying PC susceptibility loci utilizing this extraordinary collection of families, WES was performed on 539 familial cases of PC derived from 366 families all having at least three affected men with PC: 257 cases from 84 families (the majority having three sequenced/family) and 282 singleton cases. WES was performed using the Agilent (Santa Clara, CA) 50Mb SureSelect Human All Exon chip or the Agilent SureSelect V4+UTR kit. Bioinformatics analysis was performed using GenomeGPS, a comprehensive analysis pipeline developed at Mayo Clinic, which performs alignment using Novoalign (v.07.13), realignment, and recalibration using the Genome Analysis Tool Kit (GATK, v3.3), germline single nucleotide and small insertion/deletion variant calling using GATK HaplotypeCaller, and Variant quality score recalibration (VQSR), following GATK best practices v3 [DePristo et al., 2011; McKenna et al., 2010; Van der Auwera et al., 2013]. Population-based controls were selected from samples that were sequenced at Mayo Clinic using similar library preparation and sequencing to the cases. We identified 494 samples from four studies that met our inclusion criteria (germline sequencing using Agilent V2 or V4+UTR capture and with initial alignment performed using the same version of Novoalign. Samples included 89 unselected samples from the Mayo Clinic Community Biobank, 355 samples from two studies of cardiovascular phenotypes, and 50 samples from a study of neuropathy. All samples were reprocessed using the bioinformatics pipeline described above and underwent the same stringent quality control analyses.

We conducted a pathway-directed RV case-control study (see Supporting Information for details) to evaluate the role of RVs in risk of PC using 860 gene-set definitions from kyoto encyclopedia of genes and genomes [Kanehisa, 2002] and Reactome [Joshi-Tope et al., 2005]. For our purposes, we restricted our analyses to unrelated subjects by randomly selecting single individuals from pedigrees with multiple sequenced subjects. After sample exclusions for quality control or relatedness, a total of 333 cases and 349 controls remained. In our analyses, we identified multiple highly overlapping gene sets related to the Lands cycle (Reactome IDs R-HSA-1482922.1, R-HSA-1483226.1, R-HSA-1482788.1, R-HSA-1482839.1, R-HSA-1482925.1) to be significantly associated ($P < 5.8 \times 10^{-5}$) using SKAT-O [Lee et al., 2012] and burden-based testing. The Lands cycle is involved in the acyl-chain remodeling of a variety of phospholipids, and the union of the associated pathways constitutes 26 genes involving 438 unique observed variants with empirical MAF < 0.05. To investigate which RVs may be driving the association, we applied the P-BRI approach to the data, including additional covariate adjustment for WES capture kit and five leading principal components derived from the complete genetic data. Similar posterior sampling procedures that were used in the simulations were applied and no additional information was used to alter the priors on $\boldsymbol{\gamma}$.

## Results

### Simulation Analysis

The TPRs for RV associations declared at a marginal BF threshold of BF $\geq$ 10 are presented in Table 1. Overall, we observed higher TPR as well as FPRs for L-BRI relative to P-BRI, indicating marginal BFs to be larger in general for the L-BRI approach and rendering performance comparisons difficult. When evaluating TPRs at a fixed FPR of 0.01 (Table 2), we noted comparable performance. We additionally observed reduced TPR at fixed RR effect sizes as the proportion of causal variants increased, regardless of method. This is likely due to the fact that models encompassing a larger number of causal variants are less likely under the default prior distribution on the model space $\mathcal{M}$. In general, performance was comparable between the two approaches, with P-BRI tending to perform better under conditions of lower effect size and smaller proportion of causal variants and L-BRI under large effect sizes and higher causal variant proportion.

Marginal TPR and FPR results at BF thresholds of 10 and 31.6 for the high RV dimensionality simulations are presented

**Table 1. Simulation results for empirical TPRs and FPRs and corresponding bootstrapped 95% confidence intervals for causal RV detection across simulation replications with fixed marginal BF threshold set to 10 ("strong" evidence)**

| | | $\nu = 5$ | | $\nu = 10$ | | $\nu = 15$ | |
|---|---|---|---|---|---|---|---|
| Method | RR | TPR | FPR | TPR | FPR | TPR | FPR |
| P-BRI | 1.5 | 0.341 (0.323, 0.360) | 0.012 (0.011, 0.014) | 0.331 (0.317, 0.344) | 0.014 (0.013, 0.016) | 0.313 (0.303, 0.324) | 0.015 (0.013, 0016) |
| L-BRI | | 0.314 (0.295, 0.333) | 0.017 (0.015, 0.019) | 0.405 (0.391, 0.417) | 0.030 (0.028, 0.033) | 0.434 (0.423, 0.445) | 0.039 (0.036, 0.042) |
| P-BRI | 2.5 | 0.694 (0.677, 0.713) | 0.014 (0.012, 0.015) | 0.643 (0.630, 0.656) | 0.014 (0.013, 0.016) | 0.569 (0.558, 0.581) | 0.014 (0.012, 0016) |
| L-BRI | | 0.761 (0.746, 0.779) | 0.035 (0.033, 0.038) | 0.806 (0.796, 0.817) | 0.058 (0.055, 0.061) | 0.784 (0.774, 0.793) | 0.070 (0.067, 0.074) |
| P-BRI | 5.0 | 0.972 (0.966, 0.978) | 0.013 (0.012, 0.015) | 0.929 (0.923, 0.936) | 0.014 (0.013, 0.016) | 0.867 (0.859, 0.875) | 0.014 (0.013, 0016) |
| L-BRI | | 0.986 (0.981, 0.990) | 0.039 (0.036, 0.041) | 0.978 (0.974, 0.982) | 0.061 (0.058, 0.064) | 0.965 (0.961, 0.969) | 0.090 (0.086, 0.095) |

**Table 2. Simulation results for empirical TPRs for causal RV detection across simulation replications at a fixed FPR of 0.01. For comparisons across methods, the difference in TPR (Δ) and corresponding bootstrap 95% confidence interval are reported**

| | | TPR | | |
|---|---|---|---|---|
| Method | RR | $\nu = 5$ | $\nu = 10$ | $\nu = 15$ |
| P-BRI | 1.5 | 0.324 | 0.288 | 0.265 |
| L-BRI | | 0.253 | 0.262 | 0.250 |
| Δ (95% CI) | | 0.070 (0.056, 0.086) | 0.026 (0.012, 0.038) | 0.015 (0.002, 0.027) |
| P-BRI | 2.5 | 0.660 | 0.603 | 0.523 |
| L-BRI | | 0.638 | 0.612 | 0.560 |
| Δ (95% CI) | | 0.022 (0.014,0.034) | −0.009 (−0.020, 0.001) | −0.037 (−0.049, −0.022) |
| P-BRI | 5.0 | 0.966 | 0.919 | 0.845 |
| L-BRI | | 0.968 | 0.929 | 0.877 |
| Δ (95% CI) | | −0.002 (−0.006, 0.001) | −0.010 (−0.015, −0.007) | −0.032 (−0.039, −0.024) |

**Table 3. Simulation results for P-BRI in high RV dimensionality ( $p = 500, 1, 000$) for marginal TPR and FPRs at traditional BF thresholds (10 and 31.6), along with bootstrap 95% confidence intervals**

| | Marginal (BF > 10) | | Marginal (BF > 31.6) | |
|---|---|---|---|---|
| RR | TPR | FPR | TPR | FPR |
| | | $p = 500$ | | |
| 2.5 | 0.379 (0.368, 0.385) | 0.014 (0.014, 0.015) | 0.235 (0.228, 0.241) | 0.004 (0.004, 0.004) |
| 5.0 | 0.605 (0.596, 0.613) | 0.014 (0.013, 0.014) | 0.477 (0.468, 0.486) | 0.004 (0.004, 0.004) |
| 10.0 | 0.833 (0.826, 0.840) | 0.012 (0.012, 0.013) | 0.762 (0.754, 0.770) | 0.004 (0.004, 0.004) |
| | | $p = 1, 000$ | | |
| 2.5 | 0.378 (0.369, 0.386) | 0.013 (0.013, 0.014) | 0.238 (0.231, 0.245) | 0.004 (0.003, 0.004) |
| 5.0 | 0.586 (0.577, 0.594) | 0.014 (0.014, 0.014) | 0.454 (0.445, 0.461) | 0.004 (0.004, 0.004) |
| 10.0 | 0.801 (0.794, 0.808) | 0.013 (0.012, 0.013) | 0.717 (0.709, 0.725) | 0.004 (0.004, 0.004) |

in Table 3. We observed similar patterns of performance with respect to underlying RR and causal variant proportions as observed in the low RV count simulations, with higher global TPRs for $p = 500$ relative to $p = 1, 000$ for a fixed causal variant effect size. Marginal RV detection evaluated by TPR and FPR was comparable across differing total number of evaluated variants at a fixed BF threshold, with increasing TPR at higher effect sizes with the FPR remaining relatively fixed.

The above simulation results do not take into account the likely high degree of multiple testing that would likely occur prior to post hoc evaluation, as the simulations only consider the case where true causal variation is present. Consequently, false-positive rates may be higher than reported, depending upon how Type I error was controlled at the first stage of testing. To evaluate the behavior of P-BRI under false-positive testing results, we conducted an additional 500 simulations for each of the high variant dimensionality conditions where none of the simulated variants were associated with case/control status. At a BF threshold of 10, variant-level false-positive rates were commensurate with those reported in Table 3 (0.012 for $p = 500$; 0.013 for $p = 1, 000$).

## Data Application

The marginal BFs for the 438 RVs analyzed in the PC risk analysis are present in Figure 1. A total of four variants in three separate genes corresponded to a BF >10 (Table 4), including a splice-site variant in gene *PLA2G4F* (hg19 chr15:42448635A→T) with a corresponding marginal BF of 2787.7 and PIP of 0.815, occurring in 19 cases, but only one control. Both *PLA2G4D* and *PLA2G4F* encode proteins

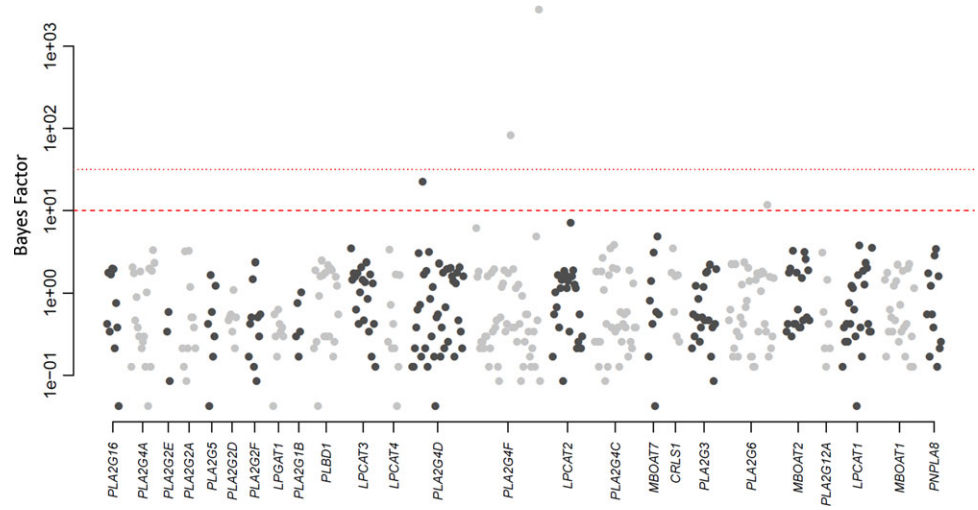


**Figure 1.** Depiction of BFs (*y*-axis on log10 scale) for all 438 RVs within 26 genes related to pathways previously identified to have significant associations with PC risk in the data set. Colors of the points alternate by gene membership from dark gray to light gray, and the *y*-axis is annotated in the original scale. Horizontal lines depict BF thresholds of 10 and 31.6.

**Table 4. Annotation and marginal BFs for RVs with BF > 10 in the post hoc analysis of 438 RVs implicated in pathway RV analysis in PC risk**

| Chromosome | Position | Gene | Reference Allele | Alternate Allele | Effect | BF |
|---|---|---|---|---|---|---|
| 15 | 42,364,046 | *PLA2G4D* | T | C | Missense | 22.5 |
| 15 | 42,439,921 | *PLA2G4F* | G | C | Nonsense | 82.9 |
| 15 | 42,448,635 | *PLA2G4F* | A | C | Splice-site | 2787.7 |
| 22 | 38,559,586 | *PLA2G6* | G | C | Intronic | 11.6 |

that selectively hydrolyze glycerophospholipids, and dysregulation of lipid metabolism has been noted in many cancers [Huang and Freter, 2015].

To evaluate the MCMC mixing for the data application, we computed the model mutation rate as the proportion of posterior samples that resulted in model state transitions (77.5%). Computational runtime for the full 30,000 iterations was approximately 20 min. Similar application of L-BRI resulted in only 188 accepted model transitions (mutation rate = 1.25%) for the same number of iterations. After 100,000 iterations (∼1 hr runtime) for two independent runs with a 50,000 burn-in, examination of the marginal BF output from L-BRI still indicated lack of convergence.

## Discussion

In this paper, we have presented a regression-based Bayesian variable selection strategy for post hoc analysis of aggregative RV associations in disease risk via a reformulation of the BRI method for case-control RV association analysis. By modeling the probability of affected status using a probit link function, in contrast to a logistic regression approach, we have demonstrated the method to be feasible for high-dimensional applications. We have also proposed a component-wise MH algorithm for updating the variable inclusion vector $\boldsymbol{\gamma}$, which results in rapid exploration of the model space. Our simulation results comparing L-BRI and P-BRI for moderate RV counts indicate that their ability to detect causal RVs is comparable for a variety of conditions, while P-BRI was also capable of detecting causal variation under very high RV dimensionality. This renders P-BRI a powerful method for dense post hoc RV association analyses, as evidenced by both our large-scale simulations and our PC risk analysis of 438 RVs within genes involved in the Lands cycle. The application of our approach indicates the significant associations previously detected by pathway-based analyses may be driven by variants within three phospholipase genes and additional targeted sequencing of these genes may be warranted in future research.

From a computational perspective, our probit approach benefits from a multistep MH algorithm for updating variable inclusion vector $\boldsymbol{\gamma}$. Execution runtimes for P-BRI in our simulation study under conditions where $p = 50$ and $N = 1,000$ averaged 6.2 min, while runtimes for our larger simulations where $p = 1,000$ and $N = 1,000$ at 30,000 iterations were approximately 75 min on average. The latter analyses were not feasible for L-BRI in our simulations due to the high model space dimensionality and single-step updating of $\boldsymbol{\gamma}$. These timings are based upon working code written in the R statistical language and executed on a modern workstation equipped with a Quad-Core AMD Opteron$^{TM}$ Processor and 16 Gb of RAM. We anticipate that computational burden for the P-BRI method may be further reduced substantially with alternative Bayesian variable selection methods, such as objective Bayes model selection [Leon-Novelo et al., 2012] and particle stochastic search [Shi and Dunson, 2011] approaches, as well as implementation of parts of the current MCMC algorithm in more computationally efficient computer languages such as C++.

A simplifying assumption of risk index methods in general is that each included RV contributes an equal effect to the RV burden, or rather that it models the mean effect of the selected RVs. Although this assumption permits efficient sampling, it may not accurately reflect the effects of the individual RV associations. It is possible to utilize existing structural definitions, such as genes or exons, as a grouping mechanism and assign separate burden-based parameters, although careful consideration is necessary to avoid singular design matrices if the number of included elements exceeds the sample size. If protective RV's are present, they would not be appropriately modeled by our approach. However, the P-BRI method could be simply modified by increasing the support of $\boldsymbol{\gamma}$ to include negative indicators, as in the MixBRI approach by Quintana et al. [2011].

Although the P-BRI method permits efficient exploration of high-dimensional model spaces, alternative MH algorithms for sampling from the model space $\mathcal{M}$ may be useful in extreme scenarios where the RV dimensionality renders the component-wise MH algorithm computationally infeasible. One approach is to consider a subset of the model space $\mathcal{M}$, $\mathcal{M}^b$, by defining the MH transition kernel such that the number of included RVs in any model $\mathcal{M}_\gamma \in \mathcal{M}^b$ is invariant and equal to an a priori defined quantity $b$. This approach is comparable to the MCMC algorithm outlined in Baragatti [2011] and preliminary simulations indicate feasibility for P-BRI with $p \geq$10,000, although further work is necessary to formally develop these methods. Adaptive algorithms designed for high-dimensional sampling in GWAS may also be of utility [Peltola et al., 2012].

There are a variety of promising extensions from our development of the P-BRI method for post hoc RV analysis. An added benefit of this work is that application to quantitative traits is trivial, because the algorithms are already in place through the latent variable $\tilde{\boldsymbol{Y}}$, although variational Bayesian methods have previously demonstrated high computational efficiency in this area [Logsdon et al., 2014]. We could also extend the regression procedure to include common variants in the model selection for a comprehensive association analysis, as well as easily adopt the integrative variable selection procedures in Quintana and Conti [2013] for informed model selection based upon existing variant annotation. Finally, we are actively evaluating methods to estimate global null model posterior probabilities using sampling procedures implemented by Liang and Xiong [2013] for association inference, as well as integrating P-BRI methods with curated pathway databases to facilitate genome-wide exploratory RV gene-set analysis.

## Acknowledgments

This research was supported by the US Public Health Service, National Institutes of Health (NIH), contract grant number GM065450 (D.J.S.) and National Cancer Institute, grant number U01 CA 89600 (S.N.T.). There are no conflicts of interest to declare.

## References

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

Bansal V, Libiger O, Torkamani A, Schork NJ. 2010. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 11(11):773–85.

Baragatti M. 2011. Bayesian variable selection for probit mixed models applied to gene selection. *Bayesian Anal* 6(2):209–229.

Cirulli ET, Goldstein DB. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Revi Genet* 11(6):415–425.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M and others. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491–498.

Dering C, Hemmelmann C, Pugh E, Ziegler A. 2011. Statistical analysis of rare sequence variants: an overview of collapsing methods. *Genet Epidemiol* 35:S12–S17.

Gelfand AE, Smith AFM, Lee TM. 1992. Bayesian-analysis of constrained parameter and truncated data problems using Gibbs sampling. *Am Stat Assoc* 87(418):523–532.

Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57(1):97–109.

Huang C, Freter C. 2015. Lipid metabolism, apoptosis and cancer therapy. *Int J Mol Sci* 16(1):924–949.

Jeffreys H. 1961. *Theory of probability*. Oxford: Clarendon Press.

Johnson AA, Jones GL, Neath RC. 2013. Component-wise Markov chain Monte Carlo: uniform and geometric ergodicity under mixing and composition. *Stat Sci* 28(3):360–375.

Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L and others. 2005. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 33(Database Issue):D428–D432.

Kanehisa M. 2002. The KEGG database. *Novartis Found Sympo* 247:91–101; discussion 101–103, 119–128, 244–252.

Kang G, Bi W, Zhao Y, Zhang JF, Yang JJ, Xu H, Loh ML, Hunger SP, Relling MV, Pounds S and others. 2014. A new system identification approach to identify genetic variants in sequencing studies for a binary phenotype. *Hum Hered* 78(2):104–116.

Lee KE, Sha NJ, Dougherty ER, Vannucci M, Mallick BK. 2003. Gene selection: a Bayesian variable selection approach. *Bioinformatics* 19(1):90–97.

Lee KJ, Jones GL, Caffo BS, Bassett SS. 2014. Spatial bayesian variable selection models on functional magnetic resonance imaging time-series data. *Bayesian Anal* 9(3):699–732.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X. 2012. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 91(2):224–237.

Leon-Novelo L, Moreno E, Casella G. 2012. Objective Bayes model selection in probit models. *Stat Med* 31(4):353–365.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83(3):311–321.

Liang FM, Xiong MM. 2013. Bayesian detection of causal rare variants under posterior consistency. *PloS One* 8(7):e69633.

Liu JS. 1994. The collapsed Gibbs sampler in Bayesian computations with applications to a gene-regulation problem. *J Am Stat Assoc* 89(427):958–966.

Liu JS. 1996. Peskun's theorem and a modified discrete-state Gibbs sampler. *Biometrika* 83(3):681–682.

Logsdon BA, Dai JY, Auer PL, Johnsen JM, Ganesh SK, Smith NL, Wilson JG, Tracy RP, Lange LA, Jiao S and others. 2014. A variational Bayes discrete mixture test for rare variant association. *Genet Epidemiol* 38(1):21–30.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M and others. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297–1303.

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. 2011. Testing for an unusual distribution of rare variants. *Plos Genet* 7(3):e1001322.

Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D and others. 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337(6090):100–104.

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

Peltola T, Marttinen P, Vehtari A. 2012. Finite adaptation and multistep moves in the Metropolis-Hastings algorithm for variable selection in genome-wide association analysis. *PloS One* 7(11):e49445.

Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 69(1):124–137.

Quintana MA, Berstein JL, Thomas DC, Conti DV. 2011. Incorporating model uncertainty in detecting rare variants: the Bayesian risk index. *Genet Epidemiol* 35(7):638–649.

Quintana MA, Conti DV. 2013. Integrative variable selection via Bayesian model uncertainty. *Stat Med* 32(28):4938–4953.

Shi MH, Dunson DB. 2011. Bayesian variable selection via particle stochastic search. *Stat Probab Lett* 81(2):283–291.

Tanner MA, Wing HW. 1987. The calculation of posterior distributions by data augmentation—rejoinder. *J Am Stat Assoc* 82(398):548–550.

Thomson PA, Parla JS, McRae AF, Kramer M, Ramakrishnan K, Yao J, Soares DC, McCarthy S, Morris SW, Cardone L and others. 2014. 708 Common and 2010 rare DISC1 locus variants identified in 1542 subjects: analysis for association with psychiatric disorder and cognitive traits. *Mol Psychiatry* 19(6):668–675.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J and others. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43:11.10.1–33.

Wilson MA, Iversen ES, Clyde MA, Schmidler SC, Schildkraut JM. 2010. Bayesian model search and multilevel inference for SNP association studies. *Ann Appl Stat* 4(3):1342–1364.

Wu G, Zhi D. 2013. Pathway-based approaches for sequencing-based genome-wide association studies. *Genet Epidemiol* 37(5):478–494.

Wu MC, Lee S, Cai TX, Li Y, Boehnke M, Lin XH. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89(1):82–93.

Yang AJ, Song XY. 2010. Bayesian variable selection for disease classification using gene expression data. *Bioinformatics* 26(2):215–222.

Zellner A. 1983. Applications of Bayesian-analysis in econometrics. *Statistician* 32(1-2):23–34.

Zhou H, Sehl ME, Sinsheimer JS, Lange K. 2010. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* 26(19):2375–2382.