A Bayes Factor Approach for Gene-Based Analysis of Rare Variants Combining Conjugate Priors and Bayesian Variable Selection

Laurent Briollais[1,2]

[1] *Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada, laurent@lunenfeld.ca*

[2] *Biostatistics division, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada*

Abstract. A common approach for detecting rare variants (RVs) associated with complex human diseases is to perform a gene-based or a region-based test of association. However, including all the RVs within a gene-based test might reduce its power since most RVs are not associated with the outcome of interest. As a shift to this paradigm, we propose to add a variable selection step to choose the RVs that compose the gene-based test statistic as a way to enhance the power of the test. We propose a Bayes Factor (BF) test statistic derived from the generalized linear model and its conjugate prior where functional annotation at the RV level can easily be integrated and the extent in prior belief can also be accommodated. A key component of our approach is the selection of the important RVs within a gene, which is performed through a novel scalable birth-death MCMC algorithm. Through simulation studies, we show that the proposed BF outperformed competing approaches both in terms of gene ranking and power to detect gene-based associations. The power of BF was improved by the use of functional annotation but interestingly, even when no annotation was included, substantial power gain was obtained from the variable selection procedure. Our application to a large whole-exome sequencing data set comparing 1,658 individuals with lung cancer to 1,492 healthy controls was able to identify new genes associated with lung cancer and pointed towards interesting cancer-related pathways.

Keywords. Bayes Factor; Generalized linear model; Bayesian variable selection; Gene-based analysis; Rare variant; Sequencing Studies

1 Introduction

With the increasing use of Next Generation Sequencing (NGS) technology in the past decade, several statistical methods for rare variant (RV) association testing have emerged. A first step is often to perform a gene-based (or region-based) test to narrow down potential genes/regions harbouring causal variants since an exhaustive search of single RV associations might lack power once correction for the millions of tests conducted is applied (Xu et al. (2021), Xu et al. (2023)). Burden and variance component (e.g. SKAT) test statistics have been the most popular gene-based methods. A strategy to increase the power of gene-based tests is

to incorporate some prior information at the RV level instead of at the gene level. Weighted approaches that use various annotation strategies to define the weights have the risk of prioritizing RVs not associated with the outcome while missing potentially causal RVs (i.e., when the weights do not correlate with the true association). As a shift to this paradigm, we propose here an alternative solution that is to perform a variable selection of RVs that should compose the gene-based test.

2 Model

2.1 Model setting

Our framework is based on a Bayesian generalized linear regression model and its conjugate prior proposed by Chen and Ibrahim (2023). For the individual i, $i \in \{1, .., n\}$, let Y_i denote a phenotype (e.g., disease outcome) following an exponential family distribution $p(\theta_i, \tau)$, where θ_i denotes the canonical parameter and τ denotes the scale parameter. The density function of Y_i is written as

$$p(y_i|\theta_i, \tau) = \exp\{a_i^{-1}(\tau)(y_i\theta_i - b(\theta_i)) + c(y_i, \tau)\}, \tag{1}$$

where a_i, b and c are known functions and determine a particular distribution type. For the simplicity purpose, we set $\tau = 1$ and $a_i^{-1}(\tau) = 1$ in equation (1), which leads to a form of natural exponential family distribution (e.g., Normal, Poisson, Gamma with known shape parameter, binomial and negative binomial distribution). Thus, the Y_i density function in equation (1) can be simplified as

$$p(y_i|\theta_i) = \exp\{y_i\theta_i - b(\theta_i) + c(y_i)\}. \tag{2}$$

A generalized linear model (GLM) with a θ-link function $\theta(.)$ is constructed to assess the association between k RVs ($G_{i1}, ..., G_{ik}$ denote k genotypes for individual i) within a gene (or a specific region on a chromosome) and the phenotype Y_i,

$$\theta_i = \theta(\boldsymbol{X_i}\boldsymbol{\beta}) = \theta(\beta_0 + \sum_{j=1}^{k} \beta_j G_{ij}). \tag{3}$$

In this model, $\boldsymbol{X} \equiv (\mathbb{1}, \boldsymbol{G}_1, ..., \boldsymbol{G}_k)$ is an $n \times (k+1)$ covariate matrix including a vector of ones $\mathbb{1}$ and k RVs. The ith row of $\boldsymbol{X}$ is denoted as $\boldsymbol{X}_i \equiv (1, \boldsymbol{G}_{i.}) \equiv (1, G_{i1}, ..., G_{ik})$. The coefficients $(\beta_1, ..., \beta_k)$ represent the effect sizes of the k RVs. The conjugate prior density function of $\boldsymbol{\beta} \equiv (\beta_0, \beta_1, ..., \beta_k)^T$ is written as

$$\pi(\boldsymbol{\beta}|a_0, \boldsymbol{y_0}) \propto \exp\left[a_0\{\boldsymbol{y_0}'\theta(\boldsymbol{X}\boldsymbol{\beta}) - \mathbb{1}'b(\theta(\boldsymbol{X}\boldsymbol{\beta}))\}\right], \tag{4}$$

where $\boldsymbol{y_0} \equiv (y_{01}, ..., y_{0n})$ is an $n \times 1$ vector of prior parameters and $a_0 > 0$ is a scalar prior parameter. The $\boldsymbol{y_0}$ parameter could be interpreted as a prior guess for $E(\boldsymbol{Y})$, e.g. each individual's probability of disease when Y_i is binary outcome. We define $y_{0i} = \exp(\boldsymbol{G}_{i.}\boldsymbol{w})/(1+\exp(\boldsymbol{G}_{i.}\boldsymbol{w}))$ for the binary phenotype Y, where $\boldsymbol{w} \equiv (w_1, ..., w_k)^T$ are weights given to the k

RVs. For instance, some genetic annotation priors can assign higher weights to coding variants (more likely to possess biological functions) than noncoding variants. The a_0 parameter is a precision parameter (or prior sample size) that quantifies the strength of our prior belief in $\boldsymbol{y_0}$ measuring individuals' prior prediction of Y.

Given this conjugate prior $(\boldsymbol{\beta}|a_0, \boldsymbol{y}_0) \sim D(\boldsymbol{y}_0, a_0)$ and equation (4), the posterior density of β can be written as

$$(\boldsymbol{\beta}|\boldsymbol{y}, a_0, \boldsymbol{y}_0) \sim D(\frac{a_0\boldsymbol{y}_0 + \boldsymbol{y}}{a_0 + 1}, a_0 + 1),$$

and

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}, a_0, \boldsymbol{y_0}) \propto \exp\left[(\boldsymbol{y} + a_0\boldsymbol{y_0})'\theta(\boldsymbol{X\beta}) - (a_0 + 1)\mathbb{1}'b(\theta(\boldsymbol{X\beta}))\right]. \tag{5}$$

In this study, we focus on a binary outcome, where $Y_i = 0$ or 1 represents healthy controls and patients with disease, respectively. Accordingly, assuming $Y_i \sim Bernoulli(p_i)$, a logistic regression is built for the RV association analysis,

$$\log(\frac{p_i}{1 - p_i}) = \beta_0 + \sum_{j=1}^{k} \beta_j G_{ij}. \tag{6}$$

2.2 Bayes factor for gene-based association

To assess the association between a group of RVs in the same gene and a disease outcome, we propose a test of hypothesis that compares a regression model including a chosen set of RVs to a model without any RV. We explain later how the set of RVs is chosen in the former model. Under the null hypothesis H_0, there is no association between any of the RVs and the outcome, so only the intercept β_0 is included in the regression model. Under the alternative hypothesis H_1, there is an association between a set of RVs and the outcome. Under H_1, assuming that a regression model M is built, $G_{(M)}$ denotes the set of RVs and $k_M = |G_{(M)}|$ represents the total number of RVs in the model M. We write the covariates as $X_{(M)} = (\mathbb{1}, G_{(M)1}, ..., G_{(M)k_M})$ and vector of coefficients $\boldsymbol{\beta}_{(M)} \equiv (\beta_0, \boldsymbol{\beta}_{k_M})^T$, which include an intercept, β_0, and the coefficients associated with the RVs, $\boldsymbol{\beta}_{k_M} \equiv (\beta_{(M)1}, ..., \beta_{(M)k_M})^T$.

The Bayes factor (BF) plays a dual role in this framework. First, it is used to select the best model (i.e., the model that includes the best subset of RVs) by comparing alternative models using the scalable birth-death MCMC (SBDMCMC) algorithm, that we recently developed (Wang et al. (2023)). Second, once the best model has been identified, it is compared to a null model without any RV and the resulting BF is used as a gene-based association test statistic. Following Chen et al. (2008), the BF comparing the (best) model under H_1 (model M) to the model under H_0 (null model) can be defined as

$$BF_{M0} = \frac{C_{H_1}(D)/C_{H_0}(D)}{C_{H_1}(\boldsymbol{y}_0)/C_{H_0}(\boldsymbol{y}_0)}, \tag{7}$$

where $C_{H_1}(D)$ and $C_{H_1}(\boldsymbol{y}_0)$ represent normalizing constants of $\boldsymbol{\beta}$ posterior distribution $C_{H_1}(D) = \int \pi(\boldsymbol{\beta}_{(M)}|\boldsymbol{y}, \boldsymbol{y}_0, a_0)d\boldsymbol{\beta}_{(M)}$ and prior distribution $C_{H_1}(\boldsymbol{y}_0) = \int \pi(\boldsymbol{\beta}_{(M)}|\boldsymbol{y}_0, a_0)d\boldsymbol{\beta}_{(M)}$, under H_1,

respectively; while $C_{H_0}(D)$ and $C_{H_0}(\boldsymbol{y}_0)$ represent normalizing constants of $\boldsymbol{\beta}$ posterior distribution $C_{H_0}(D) = \int \pi(\beta_0|\boldsymbol{y}, \boldsymbol{y}_0, a_0)d\beta_0$ and prior distribution $C_{H_0}(\boldsymbol{y}_0) = \int \pi(\beta_0|\boldsymbol{y}_0, a_0)d\beta_0$, under H_0, respectively.

2.3 RV selection algorithm

A key component of our approach is the selection of the important RVs within a gene. Since the number of RVs within each gene is large, pairwise comparison between models is impossible. MCMC approaches are commonly used to find the model with the largest posterior probability. For this problem, we propose here to use a more efficient local Bayesian variable selection method that we recently developed (Wang et al. (2023)), the SBDMCMC algorithm .

The SBDMCMC is a continuous time Markov process in Bayesian model selection problems. As given in Section 2.2, we would like to use the SBDMCMC algorithm to select the best covariates (RVs) from the logistic regression model. In the logistic regression (6), the model space is denoted as the power set of G: 2^G, and $G = (G_1, ..., G_k)$ is the full set of k RVs. For any $M \in 2^G$, $G_{(M)}$ denotes the set of RVs in model M. The SBDMCMC process explores the model space by adding and removing variables (RVs) corresponding to birth and death jumps. Given the current model M_1 and its RV set $G_{(M_1)}$, the birth and death events are defined by the following independent Poisson processes:

- Birth event: each variable $G_r \notin G_{(M_1)}$ is born independently of other variables as a Poisson process with rate $B_r(M_1)$. If this birth event of variable G_r happens, the process jumps to the new state M_2 with $G_{(M_2)} = G_{(M_1)} \cup G_r$.
- Death event: each variable $G_s \in G_{(M_1)}$ dies independently of other variables as a Poisson process with rate $D_s(M_1)$. If this death event of variable G_s happens, the process jumps to the new state M_2 with $G_{(M_2)} = G_{(M_1)} \setminus G_s$.

The waiting time to the next birth/death jump from the current model M_1 follows an exponential distribution with mean

$$w(M_1) = \frac{1}{\sum_{G_r \notin G_{(M_1)}} B_r(M_1) + \sum_{G_s \in G_{(M_1)}} D_s(M_1)},$$

and the probability of the birth and death events are respectively

$$\begin{aligned} p_{M_1}(G_r) &= \frac{B_r(M_1)}{\sum_{G_{r'} \notin G_{(M_1)}} B_{r'}(M_1) + \sum_{G_s \in G_{(M_1)}} D_s(M_1)}, \quad G_r \notin G_{(M_1)}, \\ q_{M_1}(G_s) &= \frac{D_s(M_1)}{\sum_{G_r \notin G_{(M_1)}} B_r(M_1) + \sum_{G_{s'} \in G_{(M_1)}} D_{s'}(M_1)}, \quad G_s \in G_{(M_1)}. \end{aligned} \tag{8}$$

Assuming the posterior probability of model M as $P(M|D) \equiv P(G_{(M)}|D)$, the birth and death rates are computed as

$$B_r(M_1) = \frac{1}{k}\frac{P(G_{(M_1)} \cup G_r|D)}{P(M_1|D)} = \frac{1}{k}\frac{P(M_2|D)}{P(M_1|D)}, \quad \forall G_r \notin G_{(M_1)}, G_{(M_2)} = G_{(M_1)} \cup G_r \tag{9}$$

and

$$D_s(M_1) \;=\; \frac{1}{k}\frac{P(G_{(M_1)}\backslash G_s|D)}{P(M_1|D)} = \frac{1}{k}\frac{P(M_2'|D)}{P(M_1|D)}, \quad \forall G_s \in G_{(M_1)}, G_{(M_2')} = G_{(M_1)} \setminus G_s, \tag{10}$$

respectively, which are based on the ratio between the posterior probability of the new model (M_2) and old model (M_1), $\frac{P(M_2|D)}{P(M_1|D)} = \frac{P(D|M_2)\pi(M_2)}{P(D|M_1)\pi(M_1)}$ (description of M_2' as a new model is same as M_2, and it is omitted for simplification hereafter). Here, $\pi(M_1)$ and $\pi(M_2)$ denote the prior probability of model M_1 and M_2, respectively. Further, $BF_{21} \equiv \frac{P(D|M_2)}{P(D|M_1)}$ is the Bayes factor between model M_2 and M_1, which can be computed exactly using the conjugate prior as given in section 2.1 and 2.2. In order to avoid over-fitting, we can apply a sparse model space prior for our SBDMCMC algorithm by using $\pi(M) \propto \alpha^{k_M}, \quad \alpha \in (0, 1]$, where k_M is the number of RVs in model M. The parameter α controls the sparsity of the prior distribution. The smaller the value α is, the larger probability a sparse model gets over a dense model. If $\alpha = 1$, all the models are equally distributed in the model space, which is the assumption we used in our simulations and read data analysis.

After applying the SBDMCMC algorithm to generate a sequence of samples from the model space: $M_1, M_2, \cdots, M_P$, we calculate their corresponding posterior probability $P(M_p|D)$, which is proportional to the waiting time $w(M_p)$. Thus, by applying the Bayesian model averaging, the posterior probability for RV G_r being selected in the logistic regression is computed as

$$p(G_r) = \sum_{p=1}^{P} \mathbf{1}_{G_r \in G_{(M_p)}} P(M_p|D), \quad r = 1, ..., k.$$

This probability is also called the posterior inclusion probability(PIP). In this paper, we use 0.5 as the cutting value to decide the inclusion of each variable, i.e. RV G_r is selected if $PIP(G_r) \geq 0.5$. One can use different cutting value to control the sparsity of variable selection results.

2.4 BF Computation and SBDMCMC algorithm

As described in section 2.3, we incorporated BF value in the SBDMCMC algorithm as the criterion for model fit to help select important RVs for a gene-based test, which we call as "BF-SBDMCMC" algorithm. Hereafter, BF is used as a short term for "BF-SBDMCMC" algorithm.

The BF-SBDMCMC algorithm is summarized as follow (see Algorithm 1 below), which is conducted for the RV selection purpose.

3 Simulation study

Our simulation studies show that the proposed BF-SBDMCMC approach outperforms competing approaches both in terms of gene ranking and power to detect gene-based associations.

The power of BF was improved by the use of functional annotation but interestingly, even when no annotation was included, substantial power gain was obtained from the variable selection procedure.

4 Application

4.1 Design

We applied BF-SBDMCMC to the WES study of the International Lung Cancer Consortium (ILCCO) data to identify new genes that are associated with lung cancer. Four independent substudies conducted at 4 sites are included in the ILCCO data, including Harvard University School of Public Health/Massachusetts General Hospital (426 cases vs. 270 controls), University Health Network and Mount Sinai Hospital in Toronto (259 cases vs. 258 controls), University of Liverpool in UK (64 cases vs. 69 controls) and International Agency for Research on Cancer (293 cases vs. 284 controls). Our application was able to identify new genes associated with lung cancer and pointed towards interesting cancer-related pathways.

5 Conclusion

Our BF approach adds to the current methodologies on RV gene-based (or region-based) association tests. It allows for an easy integration of functional annotations through the elegant formulation of the conjugate prior proposed for GLM and was developed here specifically for sequencing association studies. Besides, the extent of prior belief is parametrized in this formulation through a_0 and can be decided by the user based on the confidence in prior annotation. The power of the gene-based statistic is improved by the use of functional annotation(s) but interestingly, even when no annotation was included, substantial power gain was obtained from the variable selection procedure, which retains only RVs with the highest PIP to be considered in the BF test statistic. This is a key result since in sequencing studies, reliable information on RVs is not always available and many RVs have unknown biological significance. We will also discuss how our approach can be used to define historical priors in a replication study.

Bibliographie

Chen, M-H. and Ibrahim, J.G. (2003), Conjugate priors for generalized linear models, *Statistica Sinica*, 13, pp. 461-476.

Chen, M-H. and Huang, L. and Ibrahim, J.G. and Kim, S. (2008), Bayesian variable selection and computation for generalized linear models with conjugate priors, *Bayesian analysis*, 3, pp. 585-614.

Wang, N. and Massam, H. and Gao, X. and Briollais, L. (2023), The scalable birth–death MCMC algorithm for mixed graphical model learning with application to genomic data integration, *Annals of Applied Statistics*, 17, pp. 1958-1983.

Xu, J. and Xu, W. and Briollais, L. (2021), A Bayes factor approach with informative prior for rare genetic variant analysis from next generation sequencing data, *Biometrics*, 77, pp. 316-328.

Xu, J. and Xu, W. and Choi, J. and Brhane, Y. and Christiani, D. et al. (2023), Large-scale whole exome sequencing studies identify two genes, CTSL and APOE, associated with lung cancer, *PLoS genetics*, 19, pp. e1010902.

Algorithm 1 BF-SBDMCMC algorithm for RV selection.

Data: $\boldsymbol{Y}, \boldsymbol{G} = (G_1, G_2, \cdots, G_k)$

Result: Select the best covariates in $\boldsymbol{G}$ (RVs) to fit the data in the logistic regression (6)

Step 1(Birth-death process part): Given the current model M_1,

1. For each $G_r \notin G_{(M_1)}$, denote $G_{(M_2)} = G_{(M_1)} \cup G_r$. Compute the BF_{21} as given in Section 2.3, and calculate the birth rate as

$$B_r(M_1) = \frac{1}{k} BF_{21} \frac{\pi(M_2)}{\pi(M_1)},$$

2. For each $G_s \in G_{(M_1)}$, denote $G_{(M_2')} = G_{(M_1)} \setminus G_j$. Compute the BF_{21} as given in Section 2.3, and calculate the death rate as

$$D_s(M_1) = \frac{1}{k} BF_{21} \frac{\pi(M_2')}{\pi(M_1)},$$

3. Calculate the waiting time as

$$w(M_1) = \frac{1}{\sum_{G_r \notin G_{(M_1)}} B_r(M_1) + \sum_{G_s \in G_{(M_1)}} D_s(M_1)},$$

4. Simulate the birth/death jump based on the birth/death probabilities of equation (8), and jump to new model M_2.

Step 2(Sampling part): Starting from the empty model M_0, repeat Step 1 P times or until the variable selection reaches the local mode, generating P samples, M_1, M_2, ... M_P. In our simulations and application, we chose $P = 30$.

Step 3(Variable selection part):

1. Calculate the PIP for each variable in G:

$$p(G_r) = \sum_{p=1}^{P} \mathbf{1}_{G_r \in G_{(M_p)}} P(M_p|D), \quad r = 1, ..., k.$$

2. Select the random variable G_r if $p(G_r) \geq 0.5$.