A novel region-based Bayesian approach for genetic association with next generation sequencing (NGS) data

Jingxiong Xu[1,2] & Wei Xu[2,3] & Laurent Briollais[1,2]

Jingxiong.xu@mail.utoronto.edu & laurent@lunenfeld.ca

[1] *Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, ON, Canada, M5T 3L9.*
[2] *Dalla Lana School of Public Health, University of Toronto, Toronto, ON, M5T 3M7.*
[3] *University of Health Network, Princess Margaret Hospital, Toronto, ON, M5G 2M9.*

Résumé. La découverte de variants génétiques rares à partir du séquençage de nouvelle génération devient un probléme très complexe dans le domaine de la génétique humaine. Nous proposons ici une nouvelle statistique pour tester une région chromosomique donnée basée sur le facteur de Bayes (FB) afin de mettre en évidence l'association entre un ensemble de variants rares situés sur cette région et une maladie. La vraisemblance marginale est calculée sous l'hypothèse nulle et alternative en supposant une distribution binomiale pour le nombre de variants rares dans la région. Une distribution Beta ou un mélange de Dirac et une distribution Beta est spécifiée pour la distribution *a priori*. Les hyper-paramètres sont déterminés de manière à ce que la distribution nulle du FB ne varie pas en fonction de la taille des gènes. Un test de permutations ou la statistique *False Discovery Rate* (FDR) sont utilisés pour l'inférence. Nos études de simulations ont montré la supériorité du FB comparé à des méthodes standards dans la plupart des situations envisagées. Notre application sur données réelles concernant le cancer du poumon a mis en évidence l'enrichissement en variants rares de nouveaux gènes.

Mots-clés. Inférence Bayésienne, Facteur de Bayes, Association génétique, études de sequençage, Cancer du poumon.

Abstract. The discovery of rare genetic variants through Next Generation Sequencing (NGS) is becoming a very challenging issue in the human genetic field. We propose here a novel region-based statistical test based on a Bayes Factor (BF) approach to assess evidence of association between a set of rare variants located on this region and a disease outcome. Marginal likelihood is computed under the null and alternative hypotheses assuming a binomial distribution for the rare variants count in the region. A Beta distribution or a mixture of Dirac and Beta distribution is specified for the prior distribution. The hyper-parameters are determined to ensure the null distribution of BF does not vary across genes with different sizes. A permutation test or False Discovery Rate (FDR) statistic are used for inference. Our simulations studies showed that the new BF statistic outperforms standard methods under most situations considered. Our real data application to a lung cancer study found enrichment for rare variants in novel genes.

Keywords. Bayesian inference, Bayes Factor, Genetic association, Sequencing studies, Lung cancer.

1 Introduction

The emergence of new high-throughput genotyping technologies, such as Next Generation Sequencing (NGS), allows the study of the human genome at an unprecedented depth and scale (Lee et al., 2014). The discovery of rare variants through NGS is becoming a very challenging issue in human genetics. Because rare variants occur too infrequently in the general population, single-variant association tests lack power. We propose here a novel region-based statistic based on a Bayes Factor (BF) approach to assess evidence of association between a set of rare variants located on same chromosomal region and a disease outcome.

2 Model

2.1 The NGS Data

We focus on the bi-allelic variant sites (genetic locus with two possible alleles) with minor allele frequency (MAF) less than 1%. For one genetic locus (site), the genotype of one individual is usually coded as 0 or 1 or 2, representing the number of minor alleles an individual carries. In our study, it is recoded as 0 or 1 as the genotype 2 is too rare to be observed.

2.2 Model Setting

We propose a region-based statistic by modelling the count of rare variants in a specific chromosomal region, e.g. a gene. Let X_{ijk} be the count of rare variants in the region i, for group j and individual k within group j, with $i \in \{1, ..., m\}$, $j \in \{1, 2\}$ (1 for the control group, 2 for the case group) and $k \in \{1, ..., N_j\}$. We assume that the occurrence of a rare variant at any given site of the region follows an independent Bernouilli process. The distribution of X_{ijk} is therefore Binomial

$$X_{ijk} \sim Binomial(n_{ijk}, p_{ijk})$$

where p_{ijk} is the true, unobserved rate of rare variant at a single locus of the region and n_{ijk} is total number of sites in the region i for group j and individual k.

We suppose that p_{ijk} varies across genetic regions and individuals, according to a prior density function $g(p_{ijk}|\boldsymbol{\theta}_{ij})$, with $\boldsymbol{\theta}_{ij} \equiv \boldsymbol{\theta}_{i1}$ if j is in the control group and $\boldsymbol{\theta}_{ij} \equiv \boldsymbol{\theta}_{i2}$ if j is in the case group. Our goal is to assess whether there is a difference in rare variant counts

between cases and controls in a particular region i by comparing : $H_{i0} : \boldsymbol{\theta}_{i1} = \boldsymbol{\theta}_{i2} = \boldsymbol{\theta}_i$ vs. $H_{i1} : \boldsymbol{\theta}_{i1} \neq \boldsymbol{\theta}_{i2}$ using the Bayes Factor (BF) statistic.

2.3 BF derivation under case-control design

By definition, the BF is the ratio of the marginal likelihoods of the observed data under $H_1(m_1(\boldsymbol{X}))$ and $H_0(m_0(\boldsymbol{X}))$. We derive the BF assuming a case-control sampling design. We omit the index i for sake of presentation.

Let $\boldsymbol{X} \equiv \boldsymbol{X}_N = (X_1, ..., X_N)$ be the vector of rare variant counts and $\boldsymbol{P} \equiv \boldsymbol{P}_N = (p_1, ..., p_N)$ the vector of rare variant rates over $N(= N_1 + N_2)$ individuals. Under H_0, the marginal likelihood is

$$\begin{aligned} m_0(\boldsymbol{X}) &= \int f(\boldsymbol{X}, \boldsymbol{P}) d\boldsymbol{P} \\ &= \int f(\boldsymbol{X}|\boldsymbol{P}) g(\boldsymbol{P}) d\boldsymbol{P} \\ &= \int f(\boldsymbol{X}|\boldsymbol{P}) \int g(\boldsymbol{P}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} d\boldsymbol{P} \end{aligned}$$

where f denotes binomial distribution probability mass function, g the prior density function for $\boldsymbol{P}$, and π the density function for the parameter $\boldsymbol{\theta}$ that we are interested to compare between cases and controls.

Under H_1, the marginal likelihood is written as a product of two marginal likelihood functions over cases and controls

$$\begin{aligned} m_1(\boldsymbol{X}) &= \int f(\boldsymbol{X}_1, \boldsymbol{P}_1) d\boldsymbol{P}_1 \int f(\boldsymbol{X}_2, \boldsymbol{P}_2) d\boldsymbol{P}_2 \\ &= \int f(\boldsymbol{X}_1|\boldsymbol{P}_1) g(\boldsymbol{P}_1) d\boldsymbol{P}_1 \int f(\boldsymbol{X}_2|\boldsymbol{P}_2) g(\boldsymbol{P}_2) d\boldsymbol{P}_2 \\ &= \int f(\boldsymbol{X}_1|\boldsymbol{P}_1) \int g(\boldsymbol{P}_1|\boldsymbol{\theta}_1) \pi(\boldsymbol{\theta}_1) d\boldsymbol{\theta}_1 d\boldsymbol{P}_1 \int f(\boldsymbol{X}_2|\boldsymbol{P}_2) \int g(\boldsymbol{P}_2|\boldsymbol{\theta}_2) \pi(\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2 d\boldsymbol{P}_2 \end{aligned}$$

where $\boldsymbol{X}_1$ and $\boldsymbol{P}_1$ are the vector of rare variant counts and rates in controls and $\boldsymbol{X}_2$, $\boldsymbol{P}_2$ in cases. The marginal likelihoods are calculated using the Laplace approximation.

2.4 Prior definition

Since the proportions of rare variants among individuals can be anywhere between 0 and 1, we assume that the proportions for each genomic region within each group of individuals follow a beta distribution or a mixture of Dirac and beta distribution. The

beta distribution has long been a natural choice to model binomial proportions as it is a conjugate prior distribution of the binomial distribution and is the most flexible distribution with a support interval of $[0, 1]$.
For the beta prior, we assume

$$p_{ijk}|\boldsymbol{\theta}_{ij} \sim Beta(\eta_{ij}, K_{ij}),$$

Here the beta distribution is parametrized in terms of mean (denoted by η_{ij}) and precision (denoted by K_{ij}). Compared with the traditional parameterization of the $Beta(\alpha, \beta)$ distribution, the parameters have the following relationship:

$$\eta = \frac{\alpha}{(\alpha + \beta)}, \quad K = \alpha + \beta.$$

In this hierarchical model, the biological variation among replicates is captured by the beta distribution and the variation due to the random sampling of DNA segments during sequencing is captured by the binomial distribution.

For the mixture prior, we assume that p_{ijk} follows a mixture distribution of a point mass at zero and a beta distribution with parameters η_{ij} and K_{ij}, with probability w_{0ij} and $w_{1ij} = 1 - w_{0ij}$, respectively. The distribution of X_{ijk} becomes

$$X_{ijk} = \begin{cases} 0, & \text{if } p_{ijk} = 0 \text{ with } P(p_{ijk} = 0) = w_{0ij} \\ X_{ijk} \sim Bin(n_{ijk}, p_{ijk}), & \text{if } p_{ijk} > 0 \text{ with } P(p_{ijk} > 0) = 1 - w_{0ij} \end{cases}$$

Also when $p_{ijk} > 0$, the prior density for p_{ijk} is $Beta(\eta_{ij}, K_i)$.

For both the Beta prior or mixture prior, we assume $(\eta_{ij}, K_{ij}) \equiv (\eta_{i1}, K_i)$ if j is in the control group or $(\eta_{ij}, K_{ij}) \equiv (\eta_{i2}, K_i)$ if j is in the case group. In addition, for the mixture prior, we also have $w_{0ij} \equiv w_{0i1}$ or $w_{0ij} \equiv w_{0i2}$ if j is in the control or case group, respectively.

The precision parameter K_i captures the variation of the proportion of rare variants relative to the group mean. For the simple beta prior, K_i was fixed and similar across regions $K_i \equiv K$, while for the mixed prior we allow K_i to vary across genomic regions. The precision parameter K can be estimated by the method of moment or the MLE. In our application, we chose K that maximizes the marginal likelihood and assessed in the simulation the sensitivity of BF to the choice of K (see simulation study).

2.5 Hyper-parameter Specification

We assume a hyper-prior Beta distribution for each hyper-parameter defined above: η, η_1, η_2, w_{01}, w_{02}, and w_0. These new Beta distributions are also function of a mean and precision parameter that are estimated empirically from the data. They are determined

such that the null distribution of the BF over the multiple regions tested, does not depend on the number of sites of each region under both the Beta prior and mixture prior. This ensures the validity of permutation testing procedures.

2.6 False Discovery Rate (FDR)

We propose here a Bayesian FDR control approach for our BF statistic to assess the genome wide significance of the multiple genomic regions tested.

3 Simulation Study

The genetic variants data in a specific chromosomal region were simulated with the software simuPOP (Peng and Kimmal, 2005). The size of the region was 100k base pairs including 6372 sites (i.e. where a variant can be observed). We removed variants with MAF greater than 1% or less than 0.05% leaving 147 sites for our analysis. The disease status was generated using Logistic regression, with 15 randomly selected causal variants. The effect size of each causal variant is inversely proportional to its MAF (Wu, et al. 2011). In our current simulation study, we assume all causal variants are deleterious (i.e increase the disease risk). We used 500 cases and 500 controls.

Table 1: Power study of different versions of BF and SKAT/Burden statistics with 5000 replicates.

Statistical Test			Power[&] (%)
Bayes Factor		K	
Beta prior	Compare η	100	50.9
		200	60.1
		300	63.4
		400	65.0
		500	66.2
Mixed prior	Compare w_0	-	44.1
	Compare η	-	76.1
	Compare w_0 and η	-	74.8
SKAT			25.5
Burden			40.6

[&]Power calculation is done at the 5% level for all methods

The simulation results show that the BF approach outperforms standard methods

such as SKAT (Wu et al., 2011) and the Burden test (Li & Leal, 2008). These results still hold when considering different region sizes and sample sizes.

4 Real data application

In the application analysis, we used a whole-exome-sequencing study from Toronto that included 258 lung cancer patients and 257 healthy matched controls. We applied two versions of BF to this data, the Beta prior and the mixture prior where we compared η under both approaches. We were able to replicate some known candidate genes, such as TERT, BRCA2 and CHRNA5, and discovered new ones. It's noteworthy that many of these genes could not be found with standard methods such as SKAT and Burden test.

5 Conclusion

Our new BF statistic is a sensitive approach to detect rare variants associated with complex diseases using the newly developed NGS technology. Both simulations and real data application showed the good performances of this approach. The use of empirical Bayes priors along with a Bayesian control of FDR offer a comprehensive framework to make genome-wide statistical inference about the important chromosomal regions associated with the disease of interest. Finally, our BF approach is implemented in the R package $rareBF$.

Bibliography

[1] Lee, S. Abecasis, G.R., Boehnke, M., Lin, X. (2014) Rare-variant association analysis: study designs and statistical tests. *American Journal of Human Genetics* **95**, 5-23.
[2] Mardis, E.R. (2009). Cancer genome sequencing: a review. *Human Molecular Genetics* **18(R2),** R163-R168.
[3] Peng, B. and Kimmal, M. (2005). simuPOP: a forward-time population genetics simulation environment. *bioinformatics*, **21(18)** 3686-3687.
[4] Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., Lin, X.(2011). Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *American Journal of Human Genetics*, **89(1)** 82-93.
[5] Li B., Leal S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American Journal of Human Genetics*, **83(3)** 311-321.