Yan et al. BMC Proceedings 2011, 5(Suppl 9):S99
http://www.biomedcentral.com/1753-6561/5/S9/S99

BMC Proceedings

PROCEEDINGS Open Access

Identifying rare variants using a Bayesian regression approach

Aimin Yan[1,2*], Nan M Laird[1], Cheng Li[1,2*]

From Genetic Analysis Workshop 17
Boston, MA, USA. 13-16 October 2010

Abstract

Recent advances in next-generation sequencing technologies have made it possible to generate large amounts of sequence data with rare variants in a cost-effective way. Statistical methods that test variants individually are underpowered to detect rare variants, so it is desirable to perform association analysis of rare variants by combining the information from all variants. In this study, we use a Bayesian regression method to model all variants simultaneously to identify rare variants in a data set from Genetic Analysis Workshop 17. We studied the association between the quantitative risk traits Q1, Q2, and Q4 and the single-nucleotide polymorphisms and identified several positive single-nucleotide polymorphisms for traits Q1 and Q2. However, the model also generated several apparent false positives and missed many true positives, suggesting that there is room for improvement in this model.

Background

Rare variants are genetic variants that have a minor allele frequency (MAF) less than 1%. Many previous studies have suggested that rare variants generally have larger effects on a trait than common variants. Therefore identification of rare variants has become an important research topic in recent genome-wide association studies. Several statistical approaches have been developed to tackle this problem. These methods include the weighted sum statistic [1], combined multivariate and collapsing [2], the comparison of rare variants found exclusively in case subjects to those found only in control subjects [3,4], and the kernel-based adaptive cluster [5]. Overall, the results observed from these studies indicate that multiple rare variants collectively contribute to the variations of the trait, suggesting that it is desirable to use all variants together to identify the associated genetic variants with a given phenotype.

Bayesian regression models have been used in animal breeding to predict breeding values based on all available single-nucleotide polymorphisms (SNPs) [6]. Many extensions of Bayesian regression models in this field have been discussed by Gianola et al. [7]. Bayesian stochastic variable selection methods have also provided an alternative approach to genome-wide association studies [8,9]. Srivastava and Chen [10] compared the performance of a Bayesian stochastic variable selection method with that of a penalized sparse regression method and demonstrated that the Bayesian stochastic variable selection outperformed the sparse regression and also the single-SNP-based method. Yi and Zhi [11], in a recent study, used Bayesian stochastic variable selection for the identification of rare variants. However, the studies by Srivastava and Chen [10] and Yi and Zhi [11] did not estimate the probability that the SNP will be associated with the phenotype given the data.

In the current study, we model common variants and rare variants simultaneously using a Bayesian stochastic variable selection method. We calculate the regression coefficient and posterior probability of association of each SNP and use them to measure the association between each SNP and the given trait. The difference between our method and those of Srivastava and Chen [10] and Yi and Zhi [11] is that we estimate the posterior probability that the SNP will be associated with the

* Correspondence: aimin@jimmy.harvard.edu; cli@hsph.harvard.edu
[1]Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA
Full list of author information is available at the end of the article

phenotype and use that probability to estimate the number of SNPs associated with the given trait. We apply this method to study the association for the quantitative risk factors Q1, Q2, and Q4 in the Genetic Analysis Workshop 17 (GAW17) data and successfully identify several SNPs associated with the Q1 and Q2 traits.

Methods

Overall model

First, let us introduce the model and some notation. The model is:

$$\mathbf{y}_{n\times 1} = \mu^{*}\mathbf{1}_{n\times 1} + \mathbf{X}_{n\times p}(\theta_{p\times 1} \circ \alpha_{p\times 1}) + \varepsilon_{n\times 1}, \quad (1)$$

where $\varepsilon \infty \mathrm{N}(0, \sigma_{\varepsilon}^{2} * \mathrm{I})$, n is the number of individuals, p is the number of SNPs, $\mathbf{y}$ is an $n \times 1$ phenotype vector, $\mathbf{X}$ is an $n \times p$ matrix with entries being 0, 1, and 2 encoded for the genotypes *AA*, *AB*, and *BB*, respectively, $\boldsymbol{\theta}$ is a $p \times 1$ latent variable vector with entries being 0 and 1 to perform variable selection, and $\boldsymbol{\alpha}$ is a $p \times 1$ regression coefficient vector. $\boldsymbol{\theta} \circ \boldsymbol{\alpha}$ indicates the element-wise product between $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$. If $\theta_j = 1$, then α_j (SNP j) is included in the model; if $\theta_j = 0$, then α_j is excluded from the model. $P(\theta_j = 1) = \pi$ is the prior probability that the SNP will be associated with the phenotype in question, where π is the same for all SNPs. We assume that the prior probability for k is Binomial ($B(p, \pi)$), where k is the number of SNPs that are associated with a phenotype. α_j is normally distributed with mean 0 and variance σ_{α}^{2}, and σ_{α}^{2} is the scaled inverse chi-square distribution with scale parameter S_{α} and degrees of freedom v_{α}. σ_{ε}^{2} follows the scaled inverse chi-square distribution with scale parameter S_{ε} and degrees of freedom v_{ε}.

Posterior estimations

Based on Eq. (1), one can obtain the full conditional probability functions (FCPFs) for the parameters (derivations not shown). The FCPF for σ_{ε}^{2} is the scaled inverse chi-square distribution with scale parameter:

$$\frac{(\mathbf{w} - \mathbf{x}_j\hat{\alpha}_j)'(\mathbf{w} - \mathbf{x}_j\hat{\alpha}_j)}{v_{\varepsilon} + n} \quad (2)$$

and degrees of freedom $v_{\varepsilon} + n$, where:

$$\mathbf{w} = \mathbf{y} - \mu^{*}\mathbf{1} - \sum_{j'=1, j'\neq j}^{p} \mathbf{x}_{j'}\hat{\alpha}_{j'}. \quad (3)$$

In expression (2), $\mathbf{x}_j$ is an $n \times 1$ vector that corresponds to the SNP j.

The FCPF for μ is the normal distribution with mean:

$$\frac{\mathbf{1}'_{1\times n}(\mathbf{y} - \mathbf{X}\hat{\alpha})}{n} \quad (4)$$

and variance $\hat{\sigma}_{\varepsilon}^{2} / n$.

The FCPF for α_j is the normal distribution with mean $\hat{\alpha}_j$ and variance $\hat{\sigma}_{\varepsilon}^{2} / C_j$, where:

$$\hat{\alpha}_j = \frac{\mathbf{x}'_j\left(\mathbf{y} - \mu^{*}\mathbf{1} - \sum_{j'=1, j'\neq j}^{p} \mathbf{x}_{j'}\hat{\alpha}_{j'}\right)}{C_j} \quad (5)$$

and:

$$C_j = \frac{\hat{\sigma}_{\alpha}^{2}\mathbf{x}'_j\mathbf{x}_j + \hat{\sigma}_{\varepsilon}^{2}}{\hat{\sigma}_{\alpha}^{2}}. \quad (6)$$

It is clear that the $\hat{\alpha}_j$ is conditional on all other $\hat{\alpha}_{j'}$.

The FCPF for σ_{α}^{2} of each locus is the scale inverse chi-square distribution with scale parameter:

$$\frac{\hat{k}\hat{\alpha}'\hat{\alpha} + v_{\alpha}S_{\alpha}}{\hat{k} + v_{\alpha}} \quad (7)$$

and degrees of freedom $\hat{k} + v_{\alpha}$, where $\hat{\alpha}$ is a vector in which the $\hat{\alpha}_j$ is not 0. To obtain $\hat{k}$, we need to decide whether each SNP should be included in the model or not. To make this decision, we need to calculate:

$$\frac{f\left(\alpha_{j\_}^{s} \middle| \theta_j^{s} = 0\right) f\left(\theta_j^{s} = 0\right)}{f\left(\alpha_j^{s} \middle| \theta_j^{s} = 1\right) f\left(\theta_j^{s} = 1\right)}. \quad (8)$$

In expression (8), $\alpha_{j\_}^{s}$ indicates that the α_j are not in the model in the sth Markov chain Monte Carlo (MCMC) iteration; and $f\left(\alpha_{j\_}^{s} \middle| \theta_j^{s} = 0\right)$ is the likelihood that the α_j are not in the model in the sth MCMC iteration and is given by:

$$f\left(\alpha_{j\_}^{s} \middle| \theta_j^{s} = 0\right) = \frac{1}{\left(2\pi\mathbf{x}'_j\mathbf{x}_j\hat{\sigma}_{\varepsilon}^{2}\right)^{1/2}} \exp\left[\frac{-\hat{\alpha}_j^{2}}{2\left(\mathbf{x}'_j\mathbf{x}_j\hat{\sigma}_{\varepsilon}^{2}\right)}\right]. \quad (9)$$

In addition, α_j^{s} indicates that the α_j are in the model in the sth MCMC iteration; and $f\left(\alpha_j^{s} \middle| \theta_j^{s} = 1\right)$ is the likelihood that the α_j are in the model in the sth MCMC iteration and is given by:

$$f\left(\alpha_j^{s} \middle| \theta_j^{s} = 1\right) = \frac{1}{\left\{2\pi\left[\left(\mathbf{x}'_j\mathbf{x}_j\right)^{2}\hat{\sigma}_{\alpha}^{2} + \mathbf{x}'_j\mathbf{x}_j\hat{\sigma}_{\varepsilon}^{2}\right]\right\}^{1/2}} \exp\left\{\frac{-\hat{\alpha}_j^{2}}{2\left[\left(\mathbf{x}'_j\mathbf{x}_j\right)^{2}\hat{\sigma}_{\alpha}^{2} + \mathbf{x}'_j\mathbf{x}_j\hat{\sigma}_{\varepsilon}^{2}\right]}\right\}. \quad (10)$$

In expression (8), $f(\theta_j^s = 1) = \pi$, and the posterior distribution for π is the Beta distribution ($\text{Beta}(1+\hat{k}, 1+p-\hat{k})$). From these likelihoods and the sampled π, we compute the value of expression (8), and use this value to determine whether θ_j^s is 1 or 0. Posterior estimations are based on the samples from the given FCPFs using MCMC sampling.

MCMC sampling

MCMC sampling works as follows. For each MCMC iteration, we first sample σ_ε^2 from its FCPF and μ from its FCPF. Next, for α_1, α_2, α_3, ..., α_j, ..., α_p, we sample from the FCPF for α_j. Whether α_j is included in the model or not is determined, and σ_j^2 is updated by $\hat{\sigma}_q^2$. $\hat{k}$ is estimated. Next we sample from the FCPF for σ_α^2. Finally, we sample π from $\text{Beta}(1+\hat{k}, 1+p-\hat{k})$.

We performed 15,000 MCMC iterations and used the first 1,000 iterations as the burn-in period. The inclusion probability for α_j is based on the proportion of θ_j = 1 in all the MCMC samples after the burn-in period. This probability is used as the posterior probability of association (PPA) for each SNP.

Data set

The GAW17 data set includes 697 unrelated individuals; each individual has 24,487 SNPs. The MAFs of the SNPs range from 0.0717% to 49.9283% [12]. Our analysis is based on quantitative traits Q1, Q2, and Q4. The GAW17 answers show that Q1 is associated with 39 SNPs in 9 genes from the vascular endothelial growth factor (VEGF) pathway, that Q2 is influenced by 72 SNPs in 13 genes related to cardiovascular risk and inflammation, and that Q4 is not affected by any of the available SNPs. There are 200 data replications for each trait. We perform an analysis for each replication and obtain the average regression coefficients and PPAs for each SNP from the 200 data replications. We then use the different cutoffs of the regression coefficients and PPAs to compute a series of true-positive rates (TPRs) and false-positive rates (FPRs). We use the receiver operating characteristic (ROC) to compare the TPR and the FPR as the cutoffs change and the area under the curve (AUC) to measure the performance of the model.

Results and discussion

We analyzed the association between quantitative traits Q1, Q2, and Q4 and the SNPs in the GAW17 data. For each trait, we assigned a relative rank for each SNP on the basis of the sorting of the absolute values of the average regression coefficients and PPAs of all SNPs in decreasing order. In our model, we estimated the number of SNPs associated with a trait. Using this number ($\hat{k}$), we identified the SNPs whose ranks are within the range of this number.

Table 1 Association analysis for trait Q1

Gene	SNP	Regression coefficient	PPA	MAF
FLT1	C13S431	2.501 (2)	0.243 (3)	0.017217
	C13S522	2.188 (3)	0.264 (2)	0.027977
	C13S523	9.027 (1)	0.998 (1)	0.066714
ARNT	C1S6533	0.478 (6)	0.043 (4)	0.011478
KDR	C4S1884	0.126 (42)	0.016 (8)	0.020803

The third and fourth columns are the average regression coefficients and posterior probabilities of association (PPAs) out of 200 replications, respectively. The numbers in parentheses for the regression coefficients indicate the rank of the SNP based on sorting the absolute values of average regression coefficients in decreasing order. The numbers in parentheses for the PPAs indicate the rank of the SNP based on sorting the PPAs in decreasing order. We use the same notation in Table 2.

For the Q1 trait, the range of the estimated number of SNPs associated with Q1 is 3 to 8. Based on this range, the SNPs whose average regression coefficients and PPAs are within the top eight rankings are considered associated with Q1 (Table 1). The ranks of C13S431, C13S522, C13S523, C1S6533, and C4S1884 are within the top eight. The GAW17 answers confirmed that all five SNPs are truly associated with Q1. C13S431, C13S522, and C13S523 are located in gene *FLT1*, C1S6533 is located in gene *ARNT*, and C4S1884 is located in gene *KDR*.

For the Q2 trait, the range of the estimated number of SNPs associated with the Q2 is 2 to 6. Based on this range, the SNPs whose average regression coefficients and PPAs are within the top six rankings are considered associated with Q2 (Table 2). We found that the ranks of C6S5380, C6S5449, C6S5441, C8S442, and C10S3050 are within the top six. C6S5380 is located in gene *VNN1*, C6S5449 and C6S5441 are in gene *VNN3*, C8S442 is in gene *LPL*, and C10S3050 is a rare variant in gene *SIRT1* with a MAF = 0.002152. The GAW17 answers confirmed that all five SNPs are truly associated with Q2. In our analysis, C10S3051 is also identified as being associated with Q2. Compared with the GAW17 answers, this finding is a false-positive association. However, we found that C10S3051 is a synonymous mutation and is identical to C10S3050. The positions of the two SNPs are close together, suggesting that the two SNPs may be in high linkage disequilibrium.

Table 2 Association analysis for trait Q2

Gene	SNP	Regression coefficient	PPA	MAF
VNN1	C6S5380	0.251 (1)	0.077 (1)	0.170732
VNN3	C6S5449	0.194 (2)	0.015 (3)	0.010043
	C6S5441	0.038 (25)	0.010 (4)	0.098278
LPL	C8S442	0.152 (5)	0.016 (2)	0.015782
SIRT1	C10S3050	0.170 (3)	0.008 (6)	0.002152

See Table 1 notes for an explanation of the notation.

For the Q4 trait, the range of the estimated number of SNPs associated with Q4 is 3 to 6. Based on this range, the SNPs whose average regression coefficients and PPAs are within the top six rankings are considered associated with Q4. Compared with the GAW17 answers, these SNPs are false positives. We observed that the correlation coefficient between Q1 and Q4 is −0.293 and that there are 39 SNPs that are associated with Q1, which could be the reason for these false positives.

The results for Q1 and Q2 demonstrate that our model identified several true SNPs associated with Q1 and Q2 but missed many true positives for these two traits. To assess the model's performance for identifying rare variants, based on the association results using all SNPs, we extract the SNPs with a MAF less than 1% and calculate the AUCs using all SNPs and using the rare variants only for Q1 and Q2. Table 3 shows that the AUCs using all SNPs range from 0.774 to 0.808; the AUCs using only the rare variants range from 0.699 to 0.724. We obtained a reasonable power to detect rare variants using this model. However, it is obvious that the power of our model to identify rare variants is less than the power to identify common variants. These results could be due to the small effects of SNPs with lower MAFs, and our model shrinks their regression coefficients to 0.

Several other factors could also have played a role in causing the false negatives and false positives. First, we observed that there is an outlier for the Q1 trait. Several studies have shown that removing this outlier might increase the detection power. Our analyses did not consider these outliers, so we expected that we could increase the power by removing the outliers in the subsequent analyses. Second, the structure information of the SNPs was not included in the model, although all SNPs were modeled simultaneously. Many studies have shown that collapsing SNPs into blocks based on linkage disequilibrium, a gene, or a biological pathway can increase the power to detect associations. In a future study, we plan to model the correlations between the SNPs or to collapse the SNPs into blocks first and then apply this model to the blocks to see whether the detection power of this method can be increased.

Table 3 AUC of the model using all SNPs and the rare variants only for traits Q1 and Q2

Trait	AUC based on regression coefficients	AUC based on PPAs
Q1	0.791 (all SNPs)	0.808 (all SNPs)
	0.699 (RV only)	0.724 (RV only)
Q2	0.776 (all SNPs)	0.774 (all SNPs)
	0.724 (RV only)	0.718 (RV only)

"All SNPs" indicates our analysis based on all SNPs; "RV only" indicates our analysis based on only the rare variants (MAF < 0.01).

Conclusions

In the present study, we modeled all SNPs simultaneously to study the association between the SNPs and the quantitative risk traits Q1, Q2, and Q4 using a Bayesian regression method. Some true associated SNPs for Q1 and Q2 were identified using this method. However, our model missed many true positives and generated several false positives, suggesting that there is room for improvement.

Acknowledgments

We thank Giovanni Parmigiani and Cheng Li's group for discussion. AY and CL acknowledge the support of National Institutes of Health (NIH) grant 3R01 GM077122-02S1. AY received a travel award from Genetic Analysis Workshop 17. The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences. This article has been published as part of *BMC Proceedings* Volume 5 Supplement 9, 2011: Genetic Analysis Workshop 17. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/5?issue=S9.

Author details

[1]Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA. [2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 44 Binney Street, Mailstop CLS11007, Boston, MA 02115, USA.

Authors' contributions

AY performed data analysis and drafted the manuscript. NML and CL helped with data analysis and manuscript writing. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 29 November 2011

References

1. Madsen BE, Browning SR: **A groupwise association test for rare mutations using a weighted sum statistic.** *PLoS Genet* 2009, **5**:e1000384.
2. Li B, Leal SM: **Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data.** *Am J Hum Genet* 2008, **83**:311-321.
3. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH: **Multiple rare alleles contribute to low plasma levels of HDL cholesterol.** *Science* 2004, **305**:869-872.
4. Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, Grundy SM, Hobbs HH: **Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels.** *Proc Natl Acad Sci USA* 2006, **103**:1810-1815.
5. Liu DJ, Leal SM: **A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions.** *PLoS Genet* 2010, **6**:e1001156.
6. Meuwissen TH, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157**:1819-1829.
7. Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R: **Additive genetic variability and the Bayesian alphabet.** *Genetics* 2009, **183**:347-363.
8. Stephens M, Balding DJ: **Bayesian statistical methods for genetic association studies.** *Nat Rev Genet* 2009, **10**:681-690.
9. Fridley BL: **Bayesian variable and model selection methods for genetic association studies.** *Genet Epidemiol* 2009, **33**:27-37.

10. Srivastava S, Chen L: **Comparison between the stochastic search variable selection and the least absolute shrinkage and selection operator for genome-wide association studies of rheumatoid arthritis.** *BMC Proc* 2009, **3(suppl 7)**:S21.
11. Yi N, Zhi D: **Bayesian analysis of rare variants in genetic association studies.** *Genet Epidemiol* 2011, **35**:57-69.
12. Almasy LA, Dyer TD, Peralta JM, Kent JW Jr., Charlesworth JC, Curran JE, Blangero J: **Genetic Analysis Workshop 17 mini-exome simulation.** *BMC Proc* 2011, **5(suppl 9)**:S2.

doi:10.1186/1753-6561-5-S9-S99
Cite this article as: Yan *et al.*: Identifying rare variants using a Bayesian regression approach. *BMC Proceedings* 2011 **5**(Suppl 9):S99.