Identifying Rare Variants Using a Bayesian Regression Approach

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

| Citation       | Yan, Aimin, Nan M. Laird, and Cheng Li. 2011. Identifying rare variants using a Bayesian regression approach. BMC Proceedings 5(Suppl 9): S99. |
|----------------|----------------------------------------------------------------------------------------------------------------------------------|
| Published Version | doi://10.1186/1753-6561-5-S9-S99                                                                                                 |
| Citable link    | http://nrs.harvard.edu/urn-3:HUL.InstRepos:8605302                                                                                 |
| Terms of Use   | This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA |
Identifying rare variants using a Bayesian regression approach

Aimin Yan1,2*, Nan M Laird1, Cheng Li1,2*

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...
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:

\[ y_{\text{nc1}} = \mu^* 1_{\text{nc1}} + X_{\text{ncg}(\theta_{\text{pc1}} \circ \alpha_{\text{pc1}})} + \varepsilon_{\text{nc1}}, \quad (1) \]

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

Posterior estimations

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

\[ (w - x_j \hat{\alpha}_j)'(w - x_j \hat{\alpha}_j) \]

\[ \frac{1}{v_\varepsilon + n} \]

and degrees of freedom \( v_\varepsilon + n \), where:

\[ w = y - \mu^* 1 - \sum_{j=1, j \neq j}^p x_j \hat{\alpha}_j. \quad (3) \]

In expression (2), \( x_j \) is an \( n \times 1 \) vector that corresponds to the SNP \( j \).

The FCPF for \( \mu \) is the normal distribution with mean:

\[ \frac{y - X \hat{\alpha}}{n} \]

and variance \( \frac{\hat{\alpha}^2}{\sigma^2_{\varepsilon} n} \).

The FCPF for \( \alpha_j \) is the normal distribution with mean \( \hat{\alpha}_j \) and variance \( \frac{\sigma^2_{\alpha}}{C_j} \), where:

\[ \hat{\alpha}_j = \frac{x_j'(y - \mu^* 1 - \sum_{j=1, j \neq j}^p x_j \hat{\alpha}_j)}{C_j} \]

and:

\[ C_j = \frac{\sigma^2_{\alpha} x_j x_j' + \sigma^2_{\varepsilon}}{\sigma^2_{\alpha} + \sigma^2_{\varepsilon}} \]

It is clear that the \( \hat{\alpha}_j \) is conditional on all other \( \hat{\alpha}_j \).

The FCPF for \( \sigma^2_{\varepsilon} \) of each locus is the scale inverse chi-square distribution with scale parameter:

\[ \frac{\hat{\alpha} C_j + v_\alpha S_\alpha}{\hat{k} + v_\alpha} \]

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:

\[ f(\alpha_j^* \mid \theta_j^* = 0) f(\theta_j^* = 0) \]

\[ f(\alpha_j^* \mid \theta_j^* = 1) f(\theta_j^* = 1) \]

In expression (8), \( \alpha_j^* \) indicates that the \( \alpha_j \) are not in the model in the \( s \)th Markov chain Monte Carlo (MCMC) iteration; and \( f(\alpha_j^* \mid \theta_j^* = 0) \) is the likelihood that the \( \alpha_j \) are not in the model in the \( s \)th MCMC iteration and is given by:

\[ f(\alpha_j^* \mid \theta_j^* = 0) = \frac{1}{(2\pi)^{n/2} \sigma^2_{\alpha}^{1/2}} \exp \left[ -\frac{\alpha_j^*}{2\sigma^2_{\alpha}} \right] \]

In addition, \( \alpha_j^* \) indicates that the \( \alpha_j \) are in the model in the \( s \)th MCMC iteration; and \( f(\alpha_j^* \mid \theta_j^* = 1) \) is the likelihood that the \( \alpha_j \) are in the model in the \( s \)th MCMC iteration and is given by:

\[ f(\alpha_j^* \mid \theta_j^* = 1) = \frac{1}{(2\pi)^{n/2} \sigma^2_{\alpha}^{1/2}} \exp \left[ -\frac{\alpha_j^*}{2\sigma^2_{\alpha}} \right] \]
In expression (8), \( f(\theta_j^1 = 1) = \pi_j \), and the posterior distribution for \( \pi_j \) is the Beta distribution \((1 + \hat{k}_j, 1 + p - \hat{k}_j)\). From these likelihoods and the sampled \( \pi \), we compute the value of expression (8), and use this value to determine whether \( \theta_j^1 \) 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 \( \sigma_a^2 \) from its FCPF and \( \mu \) from its FCPF. Next, for \( \alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_p, \ldots, \alpha_p \) we sample from the FCPF for \( \alpha_j \). Whether \( \alpha_j \) is included in the model or not is determined, and \( \sigma_a^2 \) is updated by \( \sigma_a^2 \). \( \hat{k} \) is estimated. Next we sample from the FCPF for \( \sigma_a^2 \). Finally, we sample \( \pi \) from 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 \( \alpha_j \) is based on the proportion of \( \theta_j = 1 \) 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 9 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  | C156533| 0.478 (6)               | 0.043 (4)| 0.011478|
| KDR   | C451884| 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, C156533, and C451884 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, C156533 is located in gene ARNT, and C451884 is located in gene KDR.

For the Q2 trait, the range of the estimated number of SNPs associated with 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.

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 R01 GM077122-0251. 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/S9.

Author details
1Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA
2Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 44 Binney Street, Mailstop CL511007, 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, Pertsemidis 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, Pertsemidis 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, 10:416-427.
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 JV 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.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at www.biomedcentral.com/submit