A Bayesian framework for Generalized Linear Mixed Models in Genome-Wide Association Studies

Xulong Wang, Vivek Philip, Gregory W. Carter

The Jackson Laboratory

Abstract

Recent technical and methodological advances have greatly expanded genome-wide association studies (GWAS). The advent of low-cost whole-genome sequencing facilitates high-resolution variant identification, and the development of linear mixed models (LMM) allows improved identification of putatively causal variants. While essential for correcting false positive associations due to population stratification, LMMs have been restricted to numerical variables. However, phenotypic traits in association studies are often categorical, coded as binary case-control or ordered variables describing disease stages. Furthermore, optimally integrating the results of prior studies remains a methodological challenge. To address these issues, we have devised a method for genomic association studies that implements a generalized linear mixed model (GLMM) in a Bayesian framework, called Bayes-GLMM. Bayes-GLMM has four major features: support of categorical variables; cohesive integration of previous GWAS results for related traits by Bayesian modeling; correction for sample relatedness by mixed modeling; and model estimation by both MCMC sampling and maximal likelihood estimation. To demonstrate our method, we applied Bayes-GLMM to the whole-genome sequencing cohort in the Alzheimer's Disease Sequencing Project (ADSP). This study contains 576 individuals distributed across 111 families, each with Alzheimer's disease diagnosed at four confidence levels. The profound population structure in these data required a mixed model approach, and the categorical trait necessitated a generalized model. In summary, this work provides the first implementation of a flexible, generalized mixed model approach in a Bayesian framework.

Introduction

Linking genomic variants to traits is central to discover the genomic mechanisms of complex diseases. By date, NHGRI has curated 1,751 publications of genome-wide association studies (GWAS) assaying at least 100,000 single nucleotide polymorphisms (SNP, Welter et al., 2014, Manolio, 2010). Following the rapid advancements of high throughput sequencing technology, more variants are characterized in unprecedented speed. 1000 Genomes Project has characterized roughly 88 million variants by whole genome sequencing 2504 individuals from 26 populations (1000 Genomes Project Consortium, 2015). Large scale sequencing projects alike will only grow in faster pace, which stage GWAS a more prominent role for new discoveries. Meanwhile, statistical methods for GWAS have dramatically evolved, from odds ratio test, to generalized linear regression model, and a more sophisticated multivariate linear mixed model (LMM). LMM gained great attention in genetics and GWAS for its capacity to correct population structures and relatedness (Henderson, 1953). LMM-compatible computational tools in the context of GWAS are quickly expanding, including ASReml, TASSEL, EMMA, QTLRel, FaST-LMM, DOQTL, and GEMMA (Gilmour et al., 1995; Zhang et al., 2010; Kang et al., 2010; Cheng et al., 2011; Lippert et al., 2011; Gatti et al., 2014; Zhou and Stephens, 2014).

While LMMs are efficient in correcting sample relatedness, the response variables are restricted as numerical. Meanwhile, phenotypic traits in GWAS are often categorical, such as a binary variable in case-control studies or a multiple levels ordered categorical variable in describing different stages of the diseases. To model discrete response variables in the context of mixed models for population relatedness correction, generalized linear mixed models (GLMM) are required. However, fitting such a model efficiently and stably is nontrivial. Instead, categorical variables are often transformed into numerical form by a rather arbitrary rule, which is risky of distorting the phenotypic information.

Another issue in GWAS is to combine results from multiple studies of the same trait. Such efforts, refereed as meta-analysis, can dramatically boost statistical power by a much larger sample size (Kavvoura and Ioannidis, 2008). Association strengths of a given variant or a genomic range among studies of a given phenotype usually fluctuate. This is often explained by different population structures and environmental exposures. It is often difficult or impossible to merge raw data of different studies into a single association model for a meta-analysis because human genotype data are usually sensitive. Marker sets among studies, data structure of the traits and components of the conditional factors are also often different among studies. This paragraph is problematic, because it mainly argues against using priors, instead of proposing an issue that can be addressed.

To approach these problems, we proposed a method Bayes-GWAS, by exploiting the flexibility of Bayesian modeling framework and the computing efficiency of a newly developed statistical programming language Stan (http://mc-stan.org; Carpenter et al., 2015). In Bayesian statistics, model parameters are assumed to be stochastic, rather than fixed as the case in the frequentist approaches (Gelman et al., 2014). The stochastic nature of Bayesian modeling provides a coherent solution to combine published results of a related GWAS by configuring the prior distributions of the statistics of interest (Verzilli et al., 2008; Newcombe et al., 2009; Stephens and Balding, 2009). Primary interest in GWAS settings is the evidence of association between phenotypic traits and genotypes of the variants or genomic regions. Effect size and a corresponding p-value were regularly reported. In the proposed Bayes-GWAS, prior distribution of the variant’s effect was set as a normal distribution with expected mean the multiplication of the standardized effect size z and the standard deviation. Standardized effect size z was further modeled by a normal distribution with expected mean the reported standardized effect size and unit deviation.

Flexibility of the Bayesian modeling allows convenient configuration of sophisticated likelihood, such as GLMM. In Bayes-GLMM, logistic and ordered logistic regression likelihoods were used to model binary and ordered categorical variables, respectively. Conditional factors and epistasis terms can be included as model covariates. Sample relatedness was modeled by a random term that followed a multivariate normal distribution (mvNormal). Covariance matrix of the mvNormal was constrained by the kinship matrix of the samples. Model parameters can be estimated by either Hamilton Markov chain Monte Carlo (HMC) sampling, or L-BFGS maximal likelihood estimation (MLE). Association evidence of a variant-trait pair can be quantified by likelihood ratio test (LRT) using the MLE results or Watanabe-Akaike information criterion (WAIC, Vehtari and Gelman, 2014) using the MCMC samples. Finally, Bayes-GWAS method was built into an R package for public use.

Results

We applied Bayes-GWAS to a dataset from Alzheimer's disease sequencing project (ADSP). ADSP was initiated by NIA and NHGRI to discover novel genomic variants for late-onset Alzheimer's disease (LOAD). As the first cycle, ADSP released whole genome sequences (WGS) results of 576 individuals from 111 families and 3 races. Each individual was diagnosed with one of the four AD levels: no (80), possible (81), probable (360), and definite (55). Additional sample information includes family pedigree, race, ethnicity, age, sex, and APOE e2/e3/e4 genotypes. Frequencies of APOE e2, e3, and e4 are 4%, 78%, and 18%, respectively. Interquartile range of sample ages is 67 to 80. This age range constrain the ADSP study into LOAD. Interestingly, while 61% of the samples are female, females proportions in the definite and probable AD groups are 65.5% and 65.6%, suggesting female a risky factor of LOAD (Figure 1a-f).

Additive effects of age, sex, and APOE e2/e4 were tested in bayes-glmm. Kinship structure was computed by using autosomal variants (Figure 1g). Model parameters were estimated by MCMC sampling. The protective and risky effects of APOE/e2 and APOE/e4 were both significant, with 95% CI for APOE/e2 and e4 -1.45 to -0.26 and 0.42 to 1.2, respectively. Female showed as a significant risk factor of LOAD, with 95% CI of its effect 0.00 to 0.74 (Figure 1h). Surprisingly, while the risky effect of age is significant, the effect size is rather trivial (95% CI 0.02 to 0.05). This reflects that ADSP chose older peoples in general, with interquartile range 67 to 80. Further, no significant interactions between the covariates were observed by including an additional interaction terms for any covariates pairs (Supplementary Figure 1).

We defined on average 4.8 million short variants (SV) for each sample (4.4 to 5.6 million), and 45.4 million unique SV in total from all 576 samples. This accounts for 1.5% of the genome. Variants with MAF less than 0.01 were excluded from GWAS for being lack of statistical power. This leaves 12.6 million SNPs and 1.5 million INDEL (Supplementary Figure 2a-b). By stratifying variants according to genetic structures, we found intron regions have higher variant density than exon and intergenic regions.

There are 5.6 variants per kilobase in intron regions, but only 3.2 and 2.5 in the intergenic and exon regions. This suggests introns are more vulnerable to random mutation, hereby more relevant with disease development. Functional consequences of the 14.1 variants were predicted (Supplementary Figure 2c). Results showed 51.8% of the total consequences were intron-related, followed by noncoding transcript variants (14.1%) and intergenic variants (9.78%).

To identify the most risky and protective LOAD variants, additive effects of the 14.1 million variants were tested independently by bayes-glmm. For computing efficiency, effect of each variant was first estimated by MLE in a generalized linear model, in which the random term was dropped (Figure 2a). Variants with p-values less than 0.0001 was kept for the second round testing, in which a full MCMC sampling was applied on the bayes-glmm method (Figure 2b). P-values of variant effect's MLE estimations were obtained by the estimated effect size and its standard error. P-values of the variant effect's MCMC sampling estimations were obtained by its empirical posterior distribution. Kinship structures in bayes-glmm for testing a given variant was computed by taking off variants from the variant's chromosome.

A total of 244 genomic variants in 73 independent LD blocks passed the LOD threshold of 0.05 FDR (Figure 3a, Table 1). Effect size, MAF, and LOD of the top variants in each LD block matched well. Interestingly, 85% of the 244 variants are risky. Rare variants show bigger effect size (correlation coefficient: -0.785), suggesting extreme protective and risky variants are less heritable. The 244 variants lead to 1101 genetic consequences, in which 51.8% were intron-related (570 out of 1101). This ratio is the same as that of all intron-related consequences relative to consequences by all variants (29.2 million out of 56.5 million). This suggests associations between intron variants and LOAD is not stronger than variants in other genetic regions. The 570 intron-consequences mapped to 153 unique variants and 53 genes. These genes are significantly enriched in metabolic process (14) and calcium ion binding activity (6, Supplementary Table). Further, 23 out of the 53 genes are included in the NHGRI GWAS category (Welter et al., 2014) including PDE7B, a gene that was associated with

Alzheimer’s disease (Sherva et al., 2013). Other top traits of the 23 GWAS category genes are rheumatoid arthritis, IgG glycosylation, obesity-related traits, and type-2 diabetes (Supplementary Table). Interestingly, variants of the 23 overlapping GWAS genes are either intron (65) or intergenic (19).

Noncoding transcript variants are the second most abundant (15.3% and 168 incidences). This corresponds to 76 unique variants and 37 genes. Majority of the noncoding transcript consequences are also in the introns (158 out of 168). The other 10 exon-consequences mapped to 8 variants affecting 6 genes: hsa-mir-6723, AC144450, LINC00870, LINC00700, FDPSP3, RP11-560L11. LINC00870 and LINC00700, as lincRNA, were associated with migraine and metabolite levels, respectively (Yu et al., 2013; Anttila et al., 2013). Additionally, we have 79 intergenic, 23 significant regulatory region, and 1 missense variant consequences (Supplementary Table).

The one missense variant is rs191267549 (chr10: 120789413). The *C* to *A* mutation causes a *P* to *T* amino acid change in NANOS1 protein. 18 out of the 19 samples who carry the rs191267549 minor allele (A) are either probable or definite LOAD, whereas all 3 homozygous rs191267549 A/A cases are definite LOAD (Figure 5). The 18 minor allele carriers spread in 10 families. NANOS1 is widely expressed in the neural and immune system and functionally rich. It affects both gene transcription as a transcription factor (TF) and translation as a RNA-binding protein (RBP). As a TF, the 11 targeting genes are enriched in multiple KEGG and GO terms, including neurotrophin signaling, focal adhesion, cell cycle, axon guidance, Jak-STAT signaling, regulation of immune response, regulation of cellular metabolic process, and regulation of apoptotic process (Supplementary Table). As a RBP, NANOS1 contains a zinc-finger motif, which regulates translation of specific mRNAs by forming a complex with PUM2 that associates with the 3’UTR of mRNA targets. Interestingly, both NANOS1 and APOE are associated atherosclerosis, suggesting the two proteins might function together in a shared pathway.

To guide setting up prior distribution of variant effects, published results from a recent meta-analysis of LOAD was taken (Lambert et al., 2013). This study examined associations of roughly 7 million SNPs by pulling together 17,008 Alzheimer's disease cases and 37,154 controls from multiple sources. Out of the 7 million SNPs, 6.76 million appeared in ADSP WGS dataset. Effect sizes and p-values of the 6.76 million SNPs as reported were taken as prior information of association. Non-informative priors were set for the remaining variants that only appeared in the ADSP dataset, which includes 5.84 million SNPs and 1.5 million INDELs. Figure 4 on performance of priors, and new manhattan.

By MLE method in Bayes-GWAS, average estimation time per variant was xxx second for the full categorical model, and xxx second for the full binary model. MCMC sampling took roughly xxx minutes to estimate a full categorical model and xxx minutes for the binary model. Therefore, it is only practical to apply MCMC sampling on selected variants in GWAS settings with millions of variants under study. We applied MCMC sampling on variants within the top associated loci as returned by the MLE method. Results by MCMC and MLE were consistent, in both effect sizes and significances. Although computationally slow, the MCMC sampling method is a useful tool to inspect the robustness and stability of model inferences by inspecting the shape of posterior distributions for model parameters and the convergence of multiple sampling chains. This property is especially appealing for estimating complex models such as the Bayesian GLMM in Bayes-GWAS. Model estimations can be broken by multiple reasons, such as a illy defined prior distribution, collinearity of predictors, and inappropriate initial values. Figure 5 on bayes-glmm speed, and a suggested GWAS pipeline.

Discussion

We proposed a new GWAS method, bayes-glmm, and applied it on a new GWAS dataset by ADSP on late-onset Alzheimer's disease. Our method addresses three challenges in GWAS analysis: categorical phenotypes, population structure correction, and prior knowledge integration. Our analysis identified multiple new LOAD-associated loci that are marginally significant. Besides categorical phenotypes, our method also takes binary and quantitative phenotypes. Binary phenotype is a special case of a categorical variable, where the category number is two. A linear mixed model was used to model quantitative phenotypes. Hereby, bayes-glmm provides a powerful alternative for the existing GWAS arsenal.

To avoid the computational burden in fitting GLMMs, categorical variable can be transformed into numerical in order to fit the efficient methods for LMM inference. To test this practice, we built a LMM for the ADSP dataset by transforming the four categorical AD status into numerical probabilities (no/0, possible/0.25, probable/0.5, definite/1). The LMM realization was estimated with QTLRel (Cheng et al., 2011). Results by LMM and GLMM were similar in general. Pearson's correlation coefficient of LRT by LMM and GLMM was 0.96. However, by comparing the top 0.1% variants in either model (16,681), the LMM increased LOD by 1.97 on average. Interquartile range of LRT differences between LMM and GLMM for the top 0.1% variants were 0.88 to 2.86. By inspecting the variants that returned the most different LOD values (top 76 with minimal LOD difference 8) by the two models, we found: (1) these variants are all rare. MAF of the top 76 variants were 0.010 to 0.036. (2) MAF of these variants across the four AD populations varied irregularly. Taking rs34827707 for example, while the minor allele appeared frequently in the definite AD population with MAF 0.18, it was rare in all the other AD populations. MAF in no, possible, and probable populations were 0.01, 0, and 0.01, respectively. This suggested arbitrarily transformation of the categories into numerical probabilities is problematic for rare variants. Indeed, inference results was sensitive to different coding rules in situations described above. LOD value for rs34827707 dropped from 29 to 15 by changing the coding rule from no/0, possible/0.25, probable/0.5, and definite/1 to no/0, possible/0.33, probable/0.66, and definite/1. As comparison, three independent cut points were estimated in the GLMM for each variant. Smaller LOD values for these variants by GLMM reflect the fact that the statistical power for cut points estimation was much compromised.

APOE, with three allele-types APOE/e2/e3/e4, is so far the largest genetic factor of LOAD. The three alleles are defined by two variants: rs429358 and rs7412. APOE/e3 allele is neutral to LOAD while APOE/e2 and APOE/e4 are protective and risky, respectively. General frequencies of the three alleles are 8.4%, 77.9%, and 13.7%. As comparison, frequencies of the three APOE alleles in the ADSP dataset are 4.4%, 77.5%, and 18.1%. This indicates the ADSP population is more disease-prone. The protective and risky effects of APOE/e2 and APOE/e4 are both significant in our analysis by including APOE/e2 and APOE/e4 as two independent covariates (Figure 2). To check how APOE allele types affect the GWAS results of variants in the APOE locus, we took off APOE allele types from the model for another association test. A LOD peak in APOE locus is clear by the new model, with rs429358 the strongest hit (LOD: 21.3, effect size: 0.93). As comparison, the APOE locus was completely flat by the model with APOE allele types included (Supplementary Figure 5). Interestingly, two additional variants close to APOE locus but independent of APOE allele types also marginally significant: rs201897835 and rs34827707.

Leaving the APOE locus, other top AD variants show weak association with LOAD in the ADSP dataset. LOD of the other top 20 AD variants from Alzheimer’s disease are 0.01-2.38 (Supplementary Table). This is consistent by using both GLMM and LMM models (Supplementary Figure 11a). This reflects AD a complex neurodegenerative disease with different subtypes and mechanisms. While ADSP was specifically targeting late-onset AD, comparing our results with other top AD variants gives us opportunity to identify AD-subtype-specific genes and pathways. Noting worthy, rs3764650 in the intron of ABCA7 showed a weak association in our dataset (Supplementary Figure 11b/c). However, LOD was 2.37. This indicates we might have lost many interesting associations from our top list because of a lack of statistical power. More samples are required to detect the weak but potentially important associations.

Our method comes with several drawbacks. Firstly, model parameters are hard to explain. Secondly, heritability estimation is elusive because of the hidden residual variances. Thirdly, only one variance component is supported. Although Bayesian modeling has no difficulty handling multiple variance components, this becomes impractical for the GWAS settings given the available computing resources nowadays. Fourthly, the speed is not satisfactory. We expect superior algorithms to be developed for generalized linear mixed model.

To summarize, we proposed a method for GWAS with three major features. 1, support multiple phenotypic data type; 2, integrate previous GWAS results on the same trait cohesively by Bayesian modeling; 3, implement population relatedness correction. With biology coming into the big data era, statistical methods for large scale association studies are undoubtedly pivotal to uncover the genetic basis of complex diseases, which ultimately lead to the solutions of precision medicine.

Backup

To improve computing efficiency, Bayes-GWAS method was optimized in several ways: (1) parallel computing, (2) vectorization of model statements to take advantage of the efficient matrix operations in Stan, (3) feeding models with proper initial values, (4) reparameterize mvNormal by Cholesky factoring. The mvNormal component for the random effect was the most computationally expensive part of our method. With Cholesky factoring, the model was still estimating N normal distributions, with N the sample size. To alleviate this computing demand, a practical strategy was to only estimating the random effect in the null model, and plug the estimates into the full model as a fixed effect for each full model with an additional variant. By doing so, we implicitly assume that the random effect is independent of the additional information by each of the variant in the full models. Although theoretically flawed, this strategy is practically appealing in large scale GWAS settings. Using ADSP data, this strategy increased the computing efficiency by xx folds for the categorical model, comparing with a full estimation of the random repeatedly for each variant. ** Note: how similar are the estimation results? Would this only make false positive problem by increasing the LOD? If so, we would propose to first fit all variants without the random term, followed by another fit for only high loci with random term included.

Disease phenotypes as defined by ADSP fit an ordered categorical variable. We also collapse the four AD categories into two to simulate a case-control study (control: no, possible; case: probable, definite). Further, the family-wise study design and multiple-races sample pool suggested the necessity of sample relatedness correction. Age and sex was included as conditional covariates for their potential confounding effects to the variants under study.

From the NHGRI GWAS category, obesity-related traits has been studied in 958 publications, Type 2 diabetes in 320 studies, Schizophrenia in 256 studies, and so on.