**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 enhanced 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 quantitative, continuous 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. To date, there are 1,751 curated 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 now being characterized at unprecedented speed. The 1000 Genomes Project has characterized roughly 88 million variants by whole genome sequencing of 2504 individuals from 26 populations (1000 Genomes Project Consortium, 2015). Large scale sequencing approaches to gene association will soon enable discovery at basepair resolution. Meanwhile, statistical methods for GWAS have dramatically evolved from odds ratio tests to generalized linear regression models to more sophisticated multivariate linear mixed models (LMMs). LMM approaches have the capacity to correct population structures and relatedness (Henderson, 1953), thereby minimizing false positives due to allelic co-segregation. Consequently, the number of LMM-compatible computational tools for genetic studies is rapidly increasing, 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 an efficient approach to correcting sample relatedness, existing tools restrict users to numerical response variables. Meanwhile, phenotypic traits in GWAS are often categorical, such as binary variables in case-control studies or multi-level ordered categorical variables corresponding to disease stage. To model discrete response variables in the context of mixed models for population relatedness correction, generalized linear mixed models (GLMMs) are required. Current approaches commonly transform categorical variables into continuous variables using approximations that assume linearity or otherwise risk distorting the phenotypic information.

The proliferation of multiple GWAS for a single disease has also generated a need for principled methods to combine results from multiple studies. Such efforts, often pursued as meta-analyses, can dramatically boost statistical power through an increase in sample size (Kavvoura and Ioannidis, 2008). Association strengths of a given variant or a genetic locus typically fluctuate across studies, which may be due to different population compositions, environmental exposures, clinical reporting standards, and experimental platforms. As a result, it is often difficult or impossible to merge raw data of different studies into a single association model. Integration approaches with more flexibility are needed to address these issues.

To address these challenges, we created the *Bayes-GWAS* method that exploits the flexibility of a Bayesian modeling framework and the computing efficiency of the recently developed statistical programming language Stan (<http://mc-stan.org>; Carpenter et al., 2015). As a Bayesian strategy, model parameters are assumed to be stochastic rather than fixed as in the case in 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 and computing posterior probabilities given new data (Verzilli et al., 2008; Newcombe et al., 2009; Stephens and Balding, 2009). Bayes-GWAS priors are determined from reported effect sizes and corresponding *p*-values, thereby allowing integration of published studies based on summary statistics. Bayes-GWAS is available as an R package for public use.

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.

**Results**

**Alzheimer’s disease sequencing project**

We applied Bayes-GLMM to data from the Alzheimer’s Disease Sequencing Project (ADSP). ADSP was initiated to discover novel genomic variants for late-onset Alzheimer’s disease (LOAD). Here we consider whole genome sequence (WGS) data from 576 individuals from 111 families. Alzheimer’s diagnoses were categorized at four confidence levels: no (N = 80), possible (N = 81), probable (N = 360), and definite (N = 55), and family pedigree, race, ethnicity, age, sex, and APOE e2/e3/e4 genotype were reported for each individual. Frequencies of APOE e2, e3, and e4 are 4%, 78%, and 18%, respectively. The population was 61% female and the interquartile range of sample ages was 67 to 80 years. This age range constrain the ADSP study into LOAD.

All model parameters were estimated by MCMC sampling. To account for the substantial relatedness between samples, kinship structure was computed from autosomal variants (Figure 1g) and included as a random effect in the model. Additive effects of age, sex, and APOE genotype were tested with fixed effects. The increased and decreased risk of APOE/e4 and APOE/e2, respectively, were both significant. The 95% CI for APOE/e2 and e4 were -1.45 to -0.26 and 0.42 to 1.2, respectively. Sex was also a significant factor, with females at higher risk (95% CI was 0.00 to 0.74) (Figure 1h). Age had a small but significant effect (95% CI 0.02 to 0.05), with the weakness of the association likely due to the narrow age range and potential longevity of non-affected individuals. All covariate pairs were tested with fixed-effect interaction terms, but no significant interactions were observed (Supplementary Figure 1).

**Genomic variants from ADSP WGS**

We detected between 4.4 and 5.6 million short variants (SV) per sample, including both single-nucleotide polymorphisms (SNPs) and single-base insertions or deletions (indels), with a mean of 4.8 million. There were 45.4 million unique SV in total, corresponding to 1.5% of the genome. Variants with minor allele frequency (MAF) less than 0.01 were excluded from GWAS for uncertain reliability and lack of statistical power, leaving 12.6 million SNPs and 1.5 million indels for association (Supplementary Figure 2a-b). By stratifying variants according to genome annotation, we found that intron regions had higher variant density than exon and intergenic regions. Per kilobase, introns contained an average of 5.6 variants, while intergenic and exonic regions contained 3.2 and 2.5 variants, respectively. 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%).

**Top LOAD-associated variants from ADSP WGS**

Associations for LOAD were tested independently for the 14.1 million variants in two steps. To efficiently identify potential candidate associations, all variants were first tested by a generalized linear model implemented in Bayes-GLMM without the relatedness term. Model parameters were estimated by the maximal likelihood estimation method (Figure 2a). Variants with *p*-values less than 0.0001 (13,232, or 0.09% of the total) were then tested with the full GLMM, including the correction for relatedness to remove associations due to genotype colinearity. We followed a leave-one-out strategy in which the SNP’s host chromosome was withheld from the kinship correction to avoid the SNP effect being subsumed in the random effect. Parameters were estimated by MCMC sampling (Figure 2b). Final *p*-values for every variant were obtained from empirical posterior distributions.

We identified 1603 variants in 378 genetically unlinked loci with *p*-values less than 1 x 10-4, 246 variants in 72 loci with *p-*values less than 1 x 10-5, 26 variants in 16 loci with *p*-values less than 1 x 10-6, and 4 variants in 3 loci with *p*-values less than 1 x 10-7 (Supplementary table). The overwhelming majority of these variants increased LOAD risk; *e.g*., 97% of the top 246 had positive effects. Further, variants with strong effects tended to occur at lower allele frequency (Figure 3a), suggesting that these variants may be negatively selected. Pearson’s correlation between MAF and absolute effect size was -0.82 for the top 1603 variants, and -0.83 for the top 246 variants. This suggested extreme variants are less heritable.

Functional consequences of the top variants were predicted by the Ensembl variant effect predictor. Intron variants were predominantly the most abundant genetic consequences. 50% of the genetic consequences from the top 246 variants (p-value cutoff 1e-5) were intron-related (435 out of 876). The 435 intron-consequences mapped to 74 variants and 45 genes. 19 out of the 45 genes appeared in the NHGRI GWAS category (Welter et al., 2014). Top traits of the 19 genes were Alzheimer’s disease (APOC1 and GABRG3 in 3 studies), LDL/HDL cholesterol (APOC1 and CMIP in 12 studies), Adiponectin levels (CMIP and HIVEP2 in 3 studies), and type-2 diabetes (CMIP and PTPRD in 3 studies) (Supplementary Table).

Intron consequences were followed by intergenic and noncoding transcript consequences, which took 17% and 16% of the total, respectively. Majority of the noncoding transcript variants were also in the introns (137 out of 140). Additionally, we identified 61 upstream gene consequences and 44 downstream gene consequences in 5K bp range, 21 NMD transcript consequences, 11 regulatory region consequences, 6 three prime consequences, and 4 missense consequences (Figure 3). The four missense consequences came from a single *APOE* SNP, rs429358, that affected four APOE isoforms. Three more variants were significant at *p* < 1 x 10-4: rs139734410 (*p* = 4 x 10-5), rs191267549 (*p* = 7.8 x 10-5), and rs41268079 (*p* = 9.6 x 10-5). rs139734410 (C/A) caused a D to E amino acid change on LRRC8E, while rs191267549 (C/A) caused a P to T amino acid change on NANOS1, and rs41268079 (A/C) caused a E to A amino acid change in ZNF684.

**Integrating published GWAS as priors**

To build on previous studies in identifying true causal variants, we integrated results from a recent meta-GWAS of LOAD as priors using Bayes-GLMM (Lambert *et al*., 2013). Lambert *et al.* examined associations of roughly 7 million SNPs by pulling together 17,008 Alzheimer’s disease cases and 37,154 controls from multiple sources. A total of 6.76 of the 7 million SNPs in Lambert *et al.* appeared in our ADSP WGS dataset. A GWAS with informative priors were set on 15585 chosen variants, which showed suggestive significance in either Lambert *et al*. or ADSP data (*p* < 1 x 10-4 in either the Lambert *et al.* or ADSP variants as estimated by the GLM). Posterior effects were estimated by MCMC sampling. Prior information of a given variant was defined by the variant’s prior effect size divided by its standard error. We found this method of using priors appealing in three aspects: (1) it standardized the different interpretations of effect size from different statistical models; (2) it took information on both effect size and its standard error; (3) it softened the strong weight of priors from big sample. By simulation, we found (1) the posterior significances improved as long as prior and existing information agreed in signs, and (2) the priors only affect posterior estimation by modulating existing information in a sigmoid fashion.

By integrating prior information from Lambert *et al*., we identified 119 variants in 2 unlinked loci that were genome-wide significant (*p* < 5 x 10-8). The first locus had two variants at the intergenic region of chromosome 2 (*rs4663105, rs6733839*). The gene nearest these variants is*BIN1*. The second locus (chr19:45240584-45429543) was 189 Kbp in length and harbored 117 significant variants in or near the *APOE* gene. Once again, the most abundant genetic locations of the 117 variants were introns (Table?), which corresponded to 64 variants in 8 genes (BCL3, CBLC, BCAM, PVRL2, CTB-129P6, TOMM40, APOE, APOC1). Additionally, we identified 29 regulatory variants, 10 synonymous variants, and 5 missense variants. The five missense consequences corresponded to two variants that changed the amino acids of APOE (rs440446: N to K; rs429358: C to R).

**Discussion**

We proposed a new GWAS method, Bayes-GLMM, and applied it on whole-genome sequencing data from the Alzheimer’s Disease Sequencing Project. Our method efficiently addresses three challenges in GWAS: categorical phenotypes, population structure, and prior knowledge integration. Our analysis identified new LOAD-associated loci that are marginally significant. Our flexible approach can readily consider binary and quantitative phenotypes in addition to ordered categorical data. We therefore consider Bayes-GLMM to be a powerful addition to existing GWAS methods.

To reduce the computational burden in fitting GLMMs, categorical variables are often transformed into continuous variables and fitted with efficient LMM methods. 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) for the two models, we found: (1) these variants were rare, with. MAF ranging from 0.010 to 0.036; (2) MAF of these variants across the four AD diagnoses 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 that transformation of ordered categories into continuous 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. In contrast, the GLMM estimated three cut points for each variant independently. Smaller LOD values for these variants by GLMM reflect the fact that the statistical power for cut point estimation was compromised by the continuous approximation.

The *APOE* gene is the greatest known genetic factor for LOAD. The three alleles are defined by two SNP variants: rs429358 and rs7412. The *APOEε3* allele is considered neutral while APOEe2 and APOEe4 are decrease and increase LOAD risk, 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%, reflecting a selected enrichment of *APOE* risk alleles in the study population. The risk profile of the three *APOE* alleles is reproduced in our study when *APOE* genotype is explicitly included as an independent variable (Figure 2). When these genotype factors are not included, rs429358 was the most significantly associated variant at the locus (*p* = …, effect: ). This association disappeared when the *APOE* genotype was 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). 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.

Bayes-GWAS enables practical association of millions of variants in a mixed model approach. For the MLE method, the average estimation time per variant was xxx second for the full categorical model, and xxx second for the full binary model. In contrast, MCMC sampling took roughly xxx minutes to estimate a full categorical model and xxx minutes for the binary model. Therefore, we found it practical to apply MCMC sampling on selected variants filtered by MLE summary statistics. Results from MCMC and MLE were consistent, in both effect size and significance. Although it is computationally intensive, we see advantages in MCMC sampling. It allows an improved assessment of the robustness and stability of model inferences by reporting the full posterior distributions of model parameters and the convergence of multiple sampling chains. This information allows one to dissect how multiple factors contribute to a model estimation, including poorly defined prior distributions, collinearity of predictors, and inappropriate initial sampling values. These properties are especially appealing for estimating complex genetic models such as the Bayesian GLMM in Bayes-GWAS.

We minimized the computational expense of Bayes-GWAS in several ways: (1) parallel computing; (2) vectorization of model statements to take advantage of the efficient matrix operations in Stan [cite]; (3) feeding models with proper initial values; and (4) reparameterize mvNormal by Cholesky factoring. The mvNormal component for the random effect was the most computationally expensive part of our method. Cholesky factoring required that the number of normal distributionsestimated is equal to sample size. To alleviate this computing demand, we followed the practical strategy of only estimating the random effect in the null model, which was then applied to each full model as a fixed parameter. By doing so, we implicitly assume that the random effect is independent of the additional information of each of the variant in the full models. Using ADSP data, this strategy increased the computing efficiency by xx fold for the categorical model compared to a full estimation of the random repeatedly for each variant.

To summarize, here we have proposed a method for GWAS with three major features: (1)a generalized model to support multiple types of phenotypic data; (2) a Bayesian strategy to effectively integrate previous GWAS results for the same trait ; and (3)a mixed-model implementation to correct population structure. With genome-wide association transitioning to whole-genome and whole-exome platforms, statistical methods for large scale association studies are essential for uncovering the genetic basis of complex disease.

**Backup**

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.

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.

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.