**A Bayesian framework for Generalized Linear Mixed Models identified New Loci for Late-onset Alzheimer’s Disease**

Xulong Wang, Vivek Philip, Gareth Howell, 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 sample relatedness and population stratification, LMMs have been restricted to quantitative 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: (1) support of categorical, binary and quantitative variables; (2) cohesive integration of previous GWAS results for related traits; (3) correction for sample relatedness by mixed modeling; and (4) model estimation by both MCMC sampling and maximal likelihood estimation. We applied *Bayes-GLMM* to the whole-genome sequencing cohort of the Alzheimer's Disease Sequencing Project (ADSP). This study contains 570 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. *Bayes-GLMM* identified three new loci that were significantly associated with Alzheimer’s disease. In summary, this work provides the first implementation of a flexible, generalized mixed model approach in a Bayesian framework for association studies.

**Introduction**

Linking genomic variants to traits is central to discover the genomic mechanisms of complex diseases. To 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 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 genomic association will soon enable discoveries at base pair resolution. Meanwhile, statistical methods for GWAS have dramatically evolved from odds ratio tests to generalized linear regression models, and to more sophisticated multivariate linear mixed models (LMMs). LMM approaches have the capacity to correct population structures and sample 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, GEMMA, and GMMAT (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; Chen et al., 2016).

While LMMs are efficient in 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 stages. 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 to fit LMMs following the assumption that the trait has constant residual variance. However, the constant residual variance assumption is usually violated by categorical trait. Consequently, fitting categorical variables by linear models leads to biased effect estimates.

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). While such meta-analysis efficient in integrating the association results of multiple studies, it is often desirable to feed a new study with existing knowledge. Further, 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.

**Results**

**Alzheimer’s disease sequencing project**

Development of *Bayes-GWAS* was originally motivated by the analysis of the whole-genome sequencing cohort of Alzheimer’s disease sequencing project (ADSP). ADSP was initiated to discover novel genomic variants for late-onset Alzheimer’s disease (LOAD). The whole genome sequence (WGS) cohorts of ADSP contained 570 participants from 111 families. This family-based design generated profound sample relatedness and promoted a mixed model approach. Further, phenotypic traits were four levels of Alzheimer’s diagnoses: no (N = 78), possible (N = 81), probable (N = 356), and definite (N = 55), which necessitated a generalized model. Family pedigree, race, ethnicity, age, sex, and APOE e2/e3/e4 genotype were also reported for each participant. The population was 61% female. Interquartile range of sample ages was 67 to 80 years. In APOE genotypes, homozygous APOE/e3 took 56.7% (N = 323) of the population, followed by 35.1% (N = 200) of e3/e4, 6.84% (N = 39) of e2/e3, 1.05% (N = 6) of e2/e4, and 0.351% (N = 2) of e2/e2 (Figure 1).

Additive effects of age, sex (female), and APOE allele-types (e2, e4) were tested with *Bayes-GWAS* (Figure 2). To account for sample relatedness, kinship structure was computed from autosomal variants, and included as the variance-covariance matrix of a random effect in the model. All model parameters were estimated by MCMC sampling, which showed that the risky and protective effects of APOE/e4 and APOE/e2 were both significant. Sex was also a significant factor, with females at higher risk. Age had a small but significant risky effect, 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).

**GWAS of ADSP WGS cohort by *Bayes-GLMM***

ADSP consortium identified a total of 27.9 million SNP from the WGS cohort, of which 10.3 million had minor allele frequency larger than 0.01 (Supplementary Figure 2). Associations of the 10.3 million SNP to LOAD status were tested by *Bayes-GLMM* in two steps (Figure 3). In the first step, a generalized linear model (ordered categorical model) without the random term was applied on each of the 10.3 variants. Model parameters were estimated by the maximal likelihood estimation (MLE) method for the sake of efficiency. Variants with p-values less than 0.0001 were marked as potential significant variants (N = 9726, Figure 4a). In the second step, the marked variants were tested with the full GLMM, including the random term to correct sample relatedness. Model parameters were estimated by MCMC sampling to avoid the numerical difficulties in estimating GLMM by MLE method. Final *p*-values for every variant were obtained from their empirical posterior distributions (Figure 4b).

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

We identified 4 variants in 3 genetically unlinked loci that were genome-wide significant (*p* < 5 x 10-8). These 4 variants (rs10490263, rs74944275, rs149372995, rs140233081) were all intergenic, of which rs10490263 was 233,714 bp upstream of SLC8A1 and 337 bp upstream of lincRNA AC007317.1, rs74944275 was 111,711 downstream of C5orf30 and 18,568 bp downstream of lincRNA CTD-2154H6.1, rs140233081 and rs149372995 were 8,097 and 8,292 downstream of PRKAR1B, respectively. To obtain functional relevance of these 4 variants, we explored the Roadmap Epigenomics and ENCODE projects for their chromatin states and protein binding annotations. We found rs10490263 overlapped with promoter histone marks in blood T cells and brain hippocampus, rs74944275 overlapped with both promoter and enhancer histone marks in multiple brain regions. Further, rs74944275 changed the binding motifs of CCNT2, Evi-1, GATA, and HDAC2, while rs140233081 and rs149372995 changed the binding motifs of NERF1a, SMC3, and TCF12. We also explored the curated eQTL datasets by GTEx to explore the effects of these SNPs on gene expression, and found rs10490263 was significantly associated with the expression of AC007317.1 in skin and testis.

Furthermore, we identified 55 variants in 28 loci with *p*-values less than 1 x 10-6 (Table 1, Figure 5). The overwhelming majority of the top 55 variants were risky; *e.g*., 94.5% of the top 55 had positive effects. Further, variants with strong effects tended to occur at lower allele frequency, suggesting that these variants may be negatively selected. The top 55 variants led to 146 genetic consequences, in which 73 were intron-related, followed by 69 intergenic, and 4 regulatory region variants (Supplementary table 1).

The 73 intron-related variants mapped to 19 variants and 18 genes. 12 out of the 18 genes appeared in the NHGRI GWAS category (Welter et al., 2014). Top traits of the 12 genes were Obesity-related traits (PTPRD, SORCS2 and SLC24A4), Alzheimer’s disease (SLC24A4, GABRG3), acute lymphoblastic leukemia (ERC2 and ST6GALNAC3), adiponectin levels (CMIP and HIVEP2 in 3 studies), bipolar disorder and schizophrenia (ERC2), and type-2 diabetes (PTPRD).

**Integrating prior knowledge**

Prior knowledge integration is one of the prominent features of Bayesian modeling. In GWAS, prior information of a variant can be modeled in multiple methods with each method allowing posterior estimations to carry different weights of the priors. We proposed, and implemented in Bayes-GLMM, a method to configure priors that targeting the unique challenges of GWAS. Our method took the reported standardized effect sizes as the prior information, and integrated it into the hierarchical model of the variant effect. To demonstrate the performance of this method, we simulated a phenotypic trait (0/1) and genotype of a variant (0, 1, 2), and used a logistic regression model (LR) to test their association. Variant effect was the statistic of interest. A range of standardized effect size was fed as the priors. Results in figure 6 showed our method of configuring priors effectively modulating the information from the data.

**Discussion**

We proposed a new GWAS method, *Bayes-GLMM*, and applied it on ADSP’s whole-genome sequencing cohort. Our method efficiently addresses three challenges in GWAS: categorical phenotypes, population structure and sample relatedness, and prior knowledge integration. Our analysis identified 3 new LOAD-associated loci that are genome-wide significant. Our flexible approach can readily consider binary and quantitative in addition to ordered categorical phenotypes. We therefore consider Bayes-GLMM to be a powerful addition to existing GWAS methods.

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. Covariance matrix of the mvNormal was constrained by the kinship matrix of the samples. Model parameters can be estimated by either L-BFGS maximal likelihood estimation (MLE), or Hamilton Markov chain Monte Carlo (HMC) sampling, as implemented in Stan.

While the MLE method was efficient and reliable in estimating generalized linear models, it was not reliable in estimating generalized linear mixed models. We found MLE’s estimation of the random term was skewed toward its pre-fed initial values, which suggested the optimizer was trapped into local optima. Therefore, the current implementation of MLE in Stan is not reliable in estimating the GLMM. On the other hand, MCMC sampler 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 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*.

The drawback of MCMC sampling was efficiency. Testing at a 2.3G Hz Intel processor, MLE method took roughly 0.12 seconds to estimate the GLM model of the GWAS on ADSP (sample size 760, refer to Method for details). As a comparison, MCMC sampler took roughly 30 seconds to collect 1000 samples for the same GLM model, and 15 minutes to collect 1000 samples for the GLMM model. *Bayes-GWAS* method was optimized in multiple ways to minimize the computational expense: (1) parallel computing; (2) conjugate prior distributions; (3) vectorization of model statements to exploit Stan’s efficient matrix operations; and (4) parameterize multivariate normal distribution of the random term by Cholesky factoring.

To reduce the computational burden in fitting GLMMs, categorical variables are often transformed into continuous variables to accommodate the efficient LMM methods. However, this practice is prone to yield incorrect type I error rate because categorical studies do not satisfy LMM’s constant residual variance assumption, that is, linear models assume residual variances are constant with respect to different values of model predictors (Chen et al., 2016). This practice also yields incorrect effect estimates due to the biased sampling in different phenotypic categories, which is prominent in the ADSP study with “probable” category took account 62% of the total while the other three categories took account 10-14%. We also found the inferences results of LMM were sensitive to different coding rules (figure required). Take rs34827707 for example, 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 to separate the four categories.

Another way to reduce the computational burden is to collapse the categorical response variable into binary and fit a binary GLMM, for example to collapse the “No” or “Possible” categories into “control”, and the “Possible” or “Definite” categories into “case”. Time for Stan’s MCMC sampler to collect 1000 samples for a binary GLMM was roughly 10 minutes, as oppose to 15 minutes for the categorical GLMM. Further, Chen et al. recently released a method GMMAT (generalized linear mixed model association test) that utilized penalized quasi-likelihood method to fit the binary-GLMM, which was significantly faster than the MCMC sampling approach. However, this practice is also error-prone due to the loss of information by collapsing multiple categories into one. We tested this practice in the ADSP data, and found the association results by binary-GLMM and categorical-GLMM disagreed with each other dramatically (say little more, figure required).

Bayesian modeling naturally allows the integration of prior information by specifying model parameter’s prior distribution. However, how to specify a variant’s prior information is an open question. Association results of each variant in a GWAS are usually reported by effect size and p-value. While critical in describing the association strength, exact values of effect sizes are highly biased to the given study because of their dependencies on the statistical model, genotype coding strategies, and covariates. Therefore, it is risky to use the reported effect sizes to configure the priors. As oppose to effect sizes, p-values, as a quantification of the deviation from a null hypothesis, are less specific to the given study. Given a hypothetical distribution of variant effect, p-values are directly transformed from the standardized effect sizes, which are the normalization of the estimated effect sizes by their standard errors. Therefore, p-values and standardized effect size are equivalent in describing the confidence of true association. However, standardized effect sizes are highly influenced by the sample size, mainly through affecting the standard errors of the estimation. Bigger sample size leads to smaller standard error, bigger absolute standardized effect size, and smaller p-value. Priors as standardized effect sizes from a large-scale study would dominate the posterior estimation of a variant’s association, and thereafter mask the information of the current study. To tackle this problem, we proposed a strategy that models the variant effect by a hierarchical model, in which variant effect was firstly modeled by a normal distribution with expected mean represented as the multiplication of the standardized expected mean and the standard deviation. The standardized expected mean was further modeled by a standard normal with expected mean specified as the prior standardized effect. Simulation results showed our method in configuring the priors effective in allowing priors only modulating information of the data under study.

*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% (in which population? Source? Also, reported frequencies for genotypes, instead of alleles). 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* risky 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). However, p-values of the APOE variants, rs429358 and rs7412, were not genome-wide significant, which reflected the relatively small sample size of the ADSP WGS cohort (570, say little more, figure required).

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 (this ending is not strong).

**Deleted**

To demonstrate these issues, we built a LMM for the ADSP dataset by transforming the four categorical AD statuses into numerical probabilities (no to 0, possible to 0.25, probable to 0.5, definite to 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 P-values by LMM and GLMM was xxx. However, by comparing the top 0.1% variants in either model, the LMM increased LRT 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 (a figure required). This result showed LMM was incompetent to control type I error.

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 nonlinearly (This is the non-constant residual variance issue as described in the GMMAT paper). 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 unstable for rare variants.

Leaving the APOE locus, other top AD variants show weak association with LOAD in the ADSP dataset. P-values of the other top 20 AD variants from Alzheimer’s disease were xxx to xxx (table required). This is consistent by using both GLMM and LMM models (Supplementary Figure, say more).

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.