Reviewer #1 (Comments for the Authors (Required)):   
  
In this paper the authors develop a Bayesian generalized linear mixed model for categorical trait GWAS, and propose a novel way to incorporate the prior knowledge about SNP effect into the analysis to increase power. Their model is overall statistically valid and the parameter inference takes advantages of the most advanced MCMC techniques in Stan, the platform they used to implement the model. This method is applied to the Alzheimer's disease (AD) using a whole-genome sequencing cohort. They showed by a mouse experiment that two putative genes identified from their analysis are functionally relevant to vascular dysfunction, which provides supportive evidence to their findings. Although the sample size is relatively small (n=570), the ability to better model the categorical trait for GWAS is shown. Both their modelling and findings could potentially be valuable contributions to the methodology of GWAS and the understanding of AD etiology. Overall this paper is well written, but clarifications are needed to better present the method. Please find my specific comments below.   
  
The presentation of linear mixed model is confusing (lines 125-140), with several typos. In line 130, "e~N(0,1)", but in line 139, it is said "σe followed inverse gamma distribution". If e~N(0,1) is assumed, does it result from the assumption that y is standardized to have variance one? If so, it needs to be stated. Another typo is that beta\_0 ~ 𝑁(0,1) but later they have sigma\_0 ~ inv\_gamma(2,1). Also, the authors assign a standard normal to beta, but contemporary methods commonly use a flat prior to avoid introducing shrinkage on the fixed effects. These need to be clarified.

Thanks for pointing out both of the typos. We have corrected them in the revised manuscript. Specifically, was corrected to and was corrected to .

It is an interesting topic to study how flat priors would perform differently than the standard normal priors for fixed effect sizes. There are three reasons for us to take standard normal as priors of variant effect when additional knowledges are not taken into account. First, it fits our initial assumption that the variant under consideration does not contribute to the phenotype. Second, it constrains variants of large effect sizes that are less likely to be true. Third, it allows modeling the mean and variance parameters of priors in Gaussian models, providing a framework for incorporating external knowledge.

We consider the effective shrinkage of a standard normal prior to be advantageous in genetic association analysis. Unlike a standard normal, a flat prior is practically a Gaussian model of very large variance. From the mathematical point of view, it is conceivable that the standard normal prior tends to shrink large fixed effects because it is practically pulling the posterior effect size back to 0 with a stronger weight (1 versus 1/infinite). With this, the standard normal priors would effectively decrease false positive findings, which is the top concern in genome-wide association studies.

To investigate how a flat prior performs differently than the standard normal prior, we built two ordered categorical models that implement either standard normal (M1) or uniform distribution (M2) as priors of variant effect, and tested their performance on the top 100 variants we found in the ADSP dataset in terms of significance. We found (1) fixed effect estimation by M1 was on average 5.7% smaller than that of M2; (2) magnitude of the decrease was proportional to the effect size.

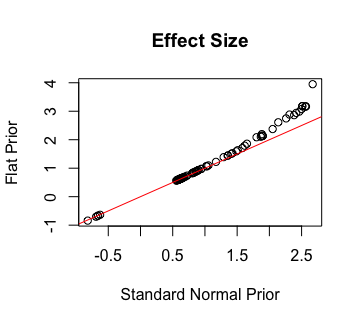


Figure 1. Effect size estimation of the top 100 variants by ordered categorical models with standard normal (X-axis) or uniform distribution (Y-axis) as priors of variant effect.

We argue it is desirable to use standard normal priors in GWAS setups. On the other hand, flat priors would be more desirable in certain situations. To support this, we have implemented the flat prior option for fixed effects in the updated Bayes-GLMM package.

The authors use logistic link in the GLMM. I wonder if they have investigated the effect of using a different link function, such as probit link function, which is equivalent to the well-known liability threshold model in animal breeding.

We chose to use logistic link because coefficients of logistic models represent odds ratios in the log scale. On the other hand, it is hard to interpret coefficients of probit models. However, we agree it is interesting to investigate how probit link would perform differently than the logistic link. To do so, we built two ordered categorical models using either logistic or probit as the link function, and tested their performances using the ADSP datasets. Because values of parameter estimations are not comparable under different link functions, we use posterior likelihood probabilities, as a metrics of model fitness, for the comparison purpose. Again, the top 100 variants as identified by our method were used. We found posterior likelihoods of the two models are similar. We also found there are more cases where logistic link function performed slightly better than the probit link (Figure 2). Further, we have implemented probit link option in the updated Bayes-GLMM method to meet the needs of certain practices.

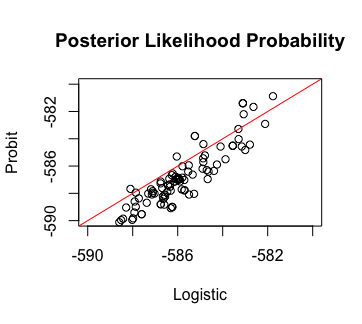


Figure 2: Posterior likelihood probability (log scale) by fitting the top 100 variants using two models. The two models using logistic (X-axis) or probit (Y-axix) as link function.

There is miscommunication in the ordered-GLMM part. Line 165, the minus sign should be plus. They said "theta = 10\*cumsum(theta\_0)". Why is ten? Is this coefficient data dependent? This needs to be explained. I wonder if it is equivalent to move the coefficient to be the parameter of the dirichlet prior for theta\_0.

We intentionally specify minus signs for model covariates in line 165. By doing so, the explanation of fixed effects estimations (*β* and ) are more natural, in that positive values of the fixed effects mean increased disease risk, while negative fixed effect values represent protective effects. The minus signs do not affect model fit in other ways.

It is clear that the original manuscript did not explain well the rationales of modeling the cut point parameters. We have added more detailed descriptions in the revised manuscript as the following:

*The cut point parameters in ordered categorical models is a vector of monotonically increasing real numbers. In our method, the increasing cut point vector was specified by the cumulative sum of a primitive parameter , which itself is a random sample of Dirichlet distribution taking advantage of the fact that Dirichlet distribution samples are a vector of positive real numbers that always sum 1*.

Per the question on where the scale factor 10 comes from and whether it is data-dependent, short answer is the scale factor 10 comes from the model structure, and it does not depend on data. In detail, cut point parameters always range between 0 to 10 under our current implementation, as shown by simulation runs, where either the range of the predictor variables or the levels of the categorical response variable were modified. In addition, the primitive parameter as Dirichlet samples are always below 1, and this is why the scale factor 10 was specified. We also tested different scale factors, and found scale factor of a larger values (>10) are also equally legitimate. The parameter of the Dirichlet prior on does not help in taking off the scale factor, because the Dirichlet parameter is only affecting the relative differences of positive real numbers (shape), but the sum will always be 1.

To incorporate the prior information of SNP effects, the authors model beta\_0 ~N(t\*sigma\_0, sigma\_0) but do not explain their choice to model the mean of beta\_0 in this way. I think they multiply t by sigma\_0 in order to standardize the prior knowledge (t) of the SNP effect to be in the same scale (i.e. sd) of beta\_0. If so, should it be t\*sqrt(sigma\_0)? Because for beta\_0 / t = sd(beta\_0)/sd(t) it follows beta\_0 = t\*sd(beta\_0) = t\*sqrt(sigma\_0) and E(beta\_0) = beta\_0 for just one beta\_0. Moreover, from a Bayesian perspective, sigma\_0 measures the uncertainty of one's belief on the SNP effect at the value of t\*sigma\_0. So linking the mean with variance could potentially be problematic, as it implies that the SNP with higher uncertainty on the effect tends to have larger effect size. Although it might not be a big issue in practice as sigma\_0 is predominated by the prior (which has mean 1) in their model, it should worth a clarification/discussion. Xulong

In modeling the variant effect by ~ , was defined as the standard deviation of the Gaussian model, instead of the variance. This is consistent with the reviewer’s understanding. The confusion came from our insufficient description of the model, which was fixed in the revised manuscript. Specifically, the following change was made in the “Modeling the prior information of variant effects” subsection:

*Old: In this method, prior distribution of variant effect was modeled by a hierarchical model,*  ~ , *in which t represented prior information of the given variant.*

*New: In this method, prior distribution of variant effect was modeled by a hierarchical model,*  ~ , *in which t represented prior information of the given variant and the standard deviation of the Gaussian model.*

We understand the reviewer’s concern in configuring expected mean of the Gaussian distribution as the multiplication of the standardized expected mean *t* and the standard derivation . The fact that *t* was further modeled as an independent parameter by a normal distribution would counteract the influence of on the estimation of variant effect. Further, our method does not take differential prior knowledge on the standard deviation parameter , which is highly influenced by the sample size of the prior work.   
  
Line 194, L-BFGS is not defined. In the equation below phi is not defined.

We have added a brief description of the L-BFGS method in the revised manuscript, together with a reference (below). The represents derivative operation. We have also changed to *d* in the revised manuscript because is not a usual way to represent derivative operation.

*L-BFGS stands for the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm for optimization problems. L-BFGS is in the family of quasi-Newton methods that approximates the original BFGS algorithm using a limited amount of computer memory.*

*Nocedal J. Wright SJ (2006). Numerical Optimization. 2nd edition. Springer-Verlag.*

Line 200, "In MCMC sampling, SE(β0) was computed directly from the samples. A standardized z value was computed as β0 / SE(β0), which led to a P-value that quantified the probability of obtaining the β0 by chance." Is the standard deviation of MCMC samples used as the SE of the estimate? Note that the standard deviation of posterior samples may not have the Frequentist property of sampling variance with repeated data. So, the Frequentist interpretation for the p-value calculated from the posterior SE may not hold. Although the asymptotic normality of the posterior mean is approximated by invoking central limit theory, the consistency between the variance of posterior samples and the sampling variance needs to be justified by simulations based on their data. Xulong

We fully agree with this comment. It is confusing to report P-values and standard errors of estimate in Bayesian models. We understand that both metrics are derived from the fixed-model-random-data philosophy of frequentist statistics by assuming repetitive data generating experiment (random data) under a fixed model. This contradicts with the random-model-fixed-data philosophy of the Bayesian statistics. With this, P-value in the Bayes-GLMM method does not mean the same as P-value in the frequentist framework, as the probability of observing more extreme test statistics under the null hypothesis.

However, we note that P-values are useful metrics in hypothesis testing and significance evaluation, which are the core objectives in the genome-wide association studies that we are approaching. Furthermore, methods for hypothesis tests and significance evaluation in the Bayesian framework are not as established as the frequentist framework. To our knowledge, Bayes factor is the most accepted metrics for hypothesis testing under the Bayesian framework. But to compute Bayes factor is computationally challenging when the model or parameter under consideration is continuous because it usually involves multiple layers of integral operations. For the generalized linear mixed model that we are building, we could not find a way to efficiently compute Bayes factor, either within Stan or using Stan outputs.

To address this issue, we borrowed the P-value idea from the frequentist regime to test the significance of variant effect by computing tail probability of its posterior probability distributions as:

= 2 \* , when mean() > 0

= 2 \* , when mean() < 0

This quantity can be interpreted under the frequentist regime as the probability of observing the estimated mean effect size under the null hypothesis that the effect size is 0. The standard error of the test statistics (estimated mean effect size) is assumed as the standard deviation of the posterior distribution.

We adopt this P-value because it is a useful in assessing significance of variant effect by summarizing its posterior distribution while taking account both of its mean and variance. To minimize confusions of this metrics, we will denote P-value as P(tail) with clear definition in the revised manuscript. We will also provide 95% credible interval of variant effect estimations ahead of the P(tail) result for all model estimations in the revised manuscript.

Line 220, any reference to this equation? Without term (1− 𝑔𝑚,𝑖)∗(1− 𝑔𝑚,𝑗), it is the VanRaden et al. (2008)'s G matrix if g is centred. Xulong

The formula in line 200 reflects an earlier implementation. We eventually settled to PLINK in computing the kinship matrix. The PLINK formula is not clear from the manual. Xulong is searching for the it.

Line 312. The statement about negative selection is not convincing. It is not clear whether the larger effect sizes of rare variants are simply due to sampling, since larger sampling variance for rare variants is expected in GWAS. The authors could overlay a power curve as in Figure 1 of Marouli et al. (2017) to show the excess of the effects of rare variants in contrast to sampling. In addition, if there is negative selection, the alleles that are deleterious and therefore kept at low frequency must be risk-increasing alleles. The authors should plot the effects of the minor alleles against their frequencies, and examine if most (or the mean effect) of the minor alleles with low frequencies have positive effects on the AD risk (see Figure 4 in Yang et al. 2015 for an example). Greg

Are the identified genes overlap with any approved drug targets (e.g. those in Drugbank and Therapeutic Targets Database)? This might be useful for drug repositioning. Greg  
  
Line 356, "In GWAS, prior information of a variant can be implemented with multiple strategies, each allowing posterior estimations to carry different weights of the priors." What are the strategies? Any reference? Xulong Greg

Prior information in the GWAS setup usually refers to studies of big sample samples. As an example, 74,046 people was included in the widely recognized GWAS study of Alzheimer’s disease by the IGAP consortium (International Genomics of Alzheimer’s Project, Lambert et al., 2013). Therefore, it is important to avoid the prior information from dominating the new information from the current study, which is of more interests. In seeking for a reasonable method to specify priors of variant effect () in GWAS problems, we tested multiple strategies as listed in table 1 (below).

|  |  |  |
| --- | --- | --- |
| Method | Formula | Description |
| 1 |  | * Plug in published values of expected mean () and its standard error () |
| 2 |  | * Model by a Gaussian distribution with expected mean the published value * Model by a Gaussian distribution with expected mean the published value |
| 3 |  | * Plug in published values of standardized expected mean (*t*) * Model by an Inverse-Gamma distribution |
| 4 |  | * Model by a Gaussian distribution with expected mean the published value * Model by an Inverse-Gamma distribution |

*PS: XULONG is putting together results for each of the 4 strategies. The IGAP results by Lambert et al. was taken for priors. This will also answer the next comment.*   
  
It is not clear what data is used as prior information in the study. To demonstrate any advantage, it would be good to show the GWAS results with and without incorporating the prior information. Xulong Greg  
*XULONG will add more words. This is closely related with the last comment.*

For a rigorous analysis, more details about the MCMC implementation should be shown, including the length of chain, burn-in, and assessment of convergence. Xulong

We have added more details of our MCMC implementation in the method section of the revised manuscript, as follow:

*In MCMC sampling, we drew 400 samples (200 as burn-in, 200 as effective) for each of 3 randomly initiated Markov chains, which results in 600 effective samples in total. We used Gelman-Robin criteria ( in Stan) to assess convergence of multiple chains (Gelman and Robin, 1992).*

The statistics, which quantifies ratio of variability between samples across the chains and samples within a chain, was added in Figure 2 legend.

*Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. Statistical Science, 7(4):457–472. 26, 355*

Line 437, the second "possible" should be "probably". Greg  
  
The authors mention GMMAT is much faster than their methods but can only fit binary data. I wonder how different the results would be between their binary-GLMM and GMMAT. Xulong – have you done this or is this easy? This is basically the same as using Bayes-GLMM with a binarization, which I think maybe you did? We could refer to that instead.

*XULONG will pull out results from CADILLAC.*

Line 478, should be Figure 7. Xulong

The reviewer was correct. We have corrected this typo in the revised manuscript. Thanks for pointing this out.

In Figure 2, are the shown the 95% HPD of the posterior distribution? It is also not clear how they estimate the cut points. Xulong

The reviewer was correct that Figure 2 shows 95% HPD (highest posterior density) of model parameters. We have added this information in the figure legend.

It is clear that the original manuscript did not explain well the rationales of modeling the cut point parameters. We have added more detailed descriptions in the revised manuscript, as detailed in our response to the reviewer’s third question (above). Model parameters in Figure 2, including the cut point parameters, were estimated by MCMC sampling.

The authors report the significant SNPs at a genome-wide significant threshold of 5\*10^-8. However, given the sequence data (~10M SNPs), rigorously speaking, the genome-wide significant level should be at 5\*10^-9. I suggest the authors to report both. Greg  
  
  
  
  
Reviewer #2 (Comments for the Authors (Required)):   
  
This manuscript applies a novel genetic association analysis approach to a whole-genome sequencing Alzheimer's disease cohort. The manuscript is very well written and clear. For this review I have been asked to review the Alzheimer's disease aspects.   
  
The Alzheimer's Disease Sequencing Project (ADSP) data is used as a test cohort, where the association method must overcome categorical disease variables, sample relatedness, population substructure and prior knowledge integration. The categorical disease variables of AD diagnosis (including no, possible, probable and definite) are generally collapsed into case-control status, and this work shows the importance of taking these groupings into account. This work looks to overcome major challenges relevant to GWAS analysis as a whole, therefore the study is of great general interest in the wider genetics community.   
The authors give a short but comprehensive overview of the genetics of LOAD including referencing all recent large scale GWAS studies and the exome sequencing studies which identified rare variants in TREM2.   
Using this whole genome sequencing data, four novel non-coding variants, in three loci associated with AD were identified using the Bayes-GLMM methods (P<5x10-8). 28 loci were associated with p<1x10-6. Many variants mapped to genes in biologically plausible disease associated pathways, including two previously AD associated genes SLC24A4 and GABRG3.   
The associated PRKAT1B and PDGFA gene regions were followed up by investigating brain expression in mice using immunoflurescence, and in humans using post-mortem RNA sequence data. Expression was correlated with plaque burden in a key brain region. This work adds to the increasing evidence that vascular dysfunction is a critical component of AD pathogenesis. They highlight the potential for the identification of novel disease mechanisms and therapeutic targets.   
  
In terms of novel findings for AD, this result highlights potential novel pathways in AD pathogenesis, but this small sample size requires replication in larger datasets. This would be possible though available whole genome sequencing data from ADNI.

Greg I’ll have another look into the ADNI data, but snp-based GWAS usually don’t report on this region. We’ll need to wait until there is more whole-genome seq data on AD.