A Bayesian framework for generalized linear mixed modeling identifies new candidate loci for late-onset Alzheimer's disease

Xulong Wang[*], Vivek M. Philip[*], Guruprasad Ananda[†], Charles C. White[‡], Ankit Malhotra[†], Paul J. Michalski[†], Krishna R. Murthy Karuturi[†], Sumana R. Chintalapudi[*], Casey Acklin[*], Michael Sasner[*], David A. Bennett[§], Philip L. De Jager[‡,**], Gareth R. Howell[*], Gregory W. Carter[*]

[*]The Jackson Laboratory Mammalian Genetics, Bar Harbor, ME, USA 04609

[†]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA 06032

[‡]Broad Institute, Cambridge, MA, USA 02142

[§]Rush Alzheimer Disease Center, Rush University Medical Center, Chicago, IL, USA 60612

[**]Center for Translational & Computational Neuroimmunology, Department of Neurology, Columbia University Medical Center, New York, NY, USA 10027

Public data source used: dbGaP, Study Accession phs000572.v7.p4.

Running Title

Generalized Mixed Model for GWAS

Keywords

genome-wide association, whole genome sequencing, Alzheimer's disease

Corresponding authors

Gregory W. Carter

The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609

207-288-6025, greg.carter@jax.org

Gareth R. Howell

The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609

207-288-6572, gareth.howell@jax.org

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 commonly 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. 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 Markov chain Monte Carlo (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 from 111 families, each with Alzheimer's disease diagnosed at one of four confidence levels. With *Bayes-GLMM* we identified four variants in three loci significantly associated with Alzheimer’s disease. Two variants, rs140233081 and rs149372995 lie between *PRKAR1B* and *PDGFA*. The coded proteins are localized to the glial-vascular unit, and *PDGFA* transcript levels are associated with AD-related neuropathology. In summary, this work provides implementation of a flexible, generalized mixed model approach in a Bayesian framework for association studies.

Introduction

Linking genomic variants to traits is central to discovering the mechanisms of genetic diseases. To date, NHGRI has curated over 1,750 publications of genome-wide association studies (GWAS) that considered at least 100,000 single nucleotide polymorphisms (SNP) [1, 2]. The adoption of high throughput sequencing technology has facilitated the rapid identification of potentially causal variants. The 1000 Genomes Project has characterized roughly 88 million variants by whole genome sequencing of 2504 individuals from 26 populations [3]. Such sequencing approaches to genomic association will soon enable discovery at base pair resolution. Meanwhile, statistical methods for GWAS have 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 sample relatedness [4], 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 [5-12].

While LMMs are efficient in correcting sample relatedness, response variables are restricted as numerical. 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. Chen et al. published a method that handles a binary response variable in the context of a mixed model [5]. However, multiple-level categorical variables are not supported. 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 often violated by categorical trait, which can bias effect estimates.

The proliferation of multiple GWAS for a single disease has also generated a need for methods to systematically combine results from multiple studies. Such efforts, often pursued as meta-analyses, can dramatically boost statistical power through an increase in sample size [13]. However, 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. Furthermore, a more general integration of prior information is often desirable, such as co-expression or other correlations between genes. Integration approaches with more flexibility are needed to address these issues.

To address these challenges, we created the *Bayes-GLMM* 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; [14]). As a Bayesian strategy, model parameters are assumed to be stochastic rather than fixed as in the case in frequentist approaches [15]. 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 [16-18]. *Bayes-GLMM* priors are determined from reported effect sizes and corresponding *p*-values, thereby allowing

integration of published studies based on summary statistics. *Bayes-GLMM* is available as an R package for public use.

We applied Bayes-*GLMM* to the analysis of whole genome sequencing association studies using resources made available by the Alzheimer's disease sequencing project (ADSP). AD is the most common form of dementia, predicted to affect 50 million people worldwide by 2020. Unfortunately, there is no known cure. AD is commonly divided into early-onset (EOAD) and late-onset (LOAD) disease. The known genetic causes of EOAD are relatively simple with mutations in amyloid precursor protein (*APP*) and APP processing enzymes such as the presenilins (e.g. *PSEN1*, *PSEN2*). However, the genetics of LOAD are poorly understood. Variations in apoliprotein E (*APOE*) are the greatest genetic risk factor, with the *ε4* allele conferring 30-50% increased risk for AD [19]. Recently, rare variants in triggering receptor expressed on myeloid cells 2 (*TREM2*) were identified that increase risk for AD [20, 21]. However, few other specific causative variants have been confirmed for AD, although numerous loci have associated by GWAS [22-28]. The lack of causative variants severely hampers diagnosis, animal model creation and the development of new therapies for LOAD. Here, we report four novel non-coding variants, identified through applying *Bayes-GLMM* to the ADSP whole genome sequence dataset. Highlighting the potential of *Bayes-GLMM*, these putative causative variants provide new avenues for testing the role of novel genes/pathways in LOAD.

Materials and Methods

Data

Data were obtained from the Alzheimer's Disease Sequencing Project via dbGaP, Study Accession phs000572.v7.p4.

Overview of the statistical models

Bayes-GLMM implemented generalized linear mixed models in a Bayesian framework. Bayesian models are defined by two parts: (1) a likelihood function that describes the data-generating process, and (2) the prior distributions of model parameters. *Bayes-GLMM* took linear regression model (LM), logistic regression model (logit-LM), and ordered logistic regression model (ordered-logit-LM) as likelihoods functions of numerical, binary, and categorical traits respectively.

Linear mixed models

In linear modeling, the numerical response variable Y_i was modeled in the linear mixed model scheme.

$$Y = X\beta + g\beta_0 + u + e$$

$$u \sim mvN(0, \sigma K)$$

$$e \sim N(0, 1)$$

$$\beta \sim N(0, 1)$$

$$\beta_0 \sim N(0, 1)$$

$$\sigma \sim inv_{gamma(2,1)}$$

$$\sigma_0 \sim inv_{gamma(2,1)}$$

In the above equations, *X* was an *n* by *m* covariate matrix with *n* the sample size and *m* the number of conditional variables. β was the corresponding parameter vector in length *m*. *g* was the numerical genotype of a variant coded as 0, 1 or 2 representing homozygous reference allele-type, heterozygous, and homozygous alternative allele-type. β_0 was the variant's effect size. A standard normal, *N(0, 1)*, was used for β_0 of variants with no known effects. Further, β followed *N(0, 1)* in prior, and σ and σ_e followed inverse gamma distribution in priors.

To model the sample relatedness, *u* was included as a random term that followed a multivariate normal distribution, with prior distribution *mvN(0, σ K)* with expected mean vector 0 and covariance matrix *σ K*. *σ* was the variance component. *K* was the kinship matrix of the samples. *mvN(0, σ K)* was parameterized by the Cholesky factoring of *K* and *n* independent standard normal distributions.

$$u = L * z$$
$$L = Chol(K)$$
$$z \sim mvN(0, \sigma I)$$

Generalized linear mixed models for binary variables

In logit-LM, the 0/1 response variable Y_i followed a binomial distribution with a scalar parameter π representing the probability that Y_i equaled 1. *π* was further transformed by the logit function and modeled in the linear model scheme.

$$\pi = P(Y_i = 1)$$
$$logit(\pi) = X\beta + g\beta_0 + u$$
$$\beta \sim N(0, 1)$$

$$\beta_0 \sim N(0,1)$$
$$u \sim mvN(0, \sigma K)$$
$$\sigma \sim inv_{gamma(2,1)}$$

Generalized linear mixed models for ordered-categorical variables

In ordered-logit-LM, the ordered categorical response variable Y_i with J levels followed a multinomial distribution with a vector of parameters π, where π_{ij} represents the probability that the ith observation falls in response category j. Cumulative distribution of π was logit-transformed and modeled in the linear model scheme.

$$P(Y_i \leq j) = \pi_{i1} + \cdots + \pi_{ij}$$
$$logit\big(P(Y_i \leq j)\big) = \theta_j - X\beta - g\beta_0 + u \quad j = 1, \ldots, J-1$$
$$\theta = 10 * cumsum(\theta_0)$$
$$\theta_0 = dirichlet(1)$$
$$\beta \sim N(0,1)$$
$$\beta_0 \sim N(0,1)$$
$$u \sim mvN(0, \sigma K)$$
$$\sigma \sim inv_{gamma(2,1)}$$

θ_j modeled the distances between the categories by providing each category a unique intercept. θ was defined as ten times the cumulative sum of a multivariate variable θ_0, where θ_0 followed a *J-1* dimension *Dirichlet* distribution in prior.

Modeling the prior information of variant effects

To integrate prior information of variant effects, *Bayes-GLMM* implemented an approach that allowed priors to only modulate information of the data under study. In this method, prior distribution of variant effect was modeled by a hierarchical model, $\beta_0 \sim N(t * \sigma_0, \sigma_0)$, in which t represented prior information of the given variant. t was further modeled by a normal distribution with expected mean the standardized effect size *prior* and unit deviation. The variable *prior* was defined by the variant's prior effect size divided by its standard error, which was often reported in published GWAS summary statistics. A standard normal, *N(0, 1)*, was used for β_0 of variants with no known effects.

$$\beta_0 \sim N(t * \sigma_0, \sigma_0)$$
$$t \sim N(prior, 1)$$
$$\sigma_0 \sim inv_{gamma(2,1)}$$

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 used information on both effect size and its standard error; and (3) it softened the strong weight of priors from studies with unbalanced sample sizes.

Model estimations

Our models were built under Stan, which provides a flexible and efficient programming environment for statistical modeling. Inherited from Stan, *Bayes-GLMM* supported two methods for parameter estimation, L-BFGS maximal likelihood estimation (MLE) and Hamilton Markov chain Monte Carlo (HMC) sampling. The MLE method made a point estimation for each parameter that maximized the joint posterior of model parameters, whereas the MCMC sampling method captured a full posterior distribution for each parameter by iterative sampling. Significance of the estimated effect size β_0 can be accessed by combing β_0 and its standard error $SE(\beta_0)$. Standard errors of MLE were computed as the inverse of the square root of the diagonal elements of the observed Fisher information matrix [44]. In MCMC sampling, $SE(\beta_0)$ was computed directly from the samples. A standardized *z* value was computed as $\beta_0 / SE(\beta_0)$, which led to a P-value that quantified the probability of obtaining the β_0 by chance.

$$SE\left(\hat{\theta}_{ML}\right) = \frac{1}{\sqrt{I\left(\hat{\theta}_{ML}\right)}}$$

$$I(\theta) = -\frac{\varphi^2}{\varphi\theta_i\theta_j} l(\theta) \quad 1 \le i, j \le p$$

$\hat{\theta}_{ML}$ was MLE of model parameters, $I(\theta)$ was the Fisher information matrix, and *p* was the number of parameters.

In genetic association studies, comparing the two nested null and full models was a widely used method to estimate the significance of a variant. The full models were the same as described above whereas the null models ignored the variant, *g*, as a linear predictor. In MLE, the null-to-full model improvements was quantified by LRT, which equals two times the log likelihood difference between the full and null models using the MLE estimation of model parameters.

$$LRT = -2 * (\log\left(P(data|\theta_p^n)\right) - \log(P(data|\theta_p^f)))$$

θ_p^n and θ_p^f were the MLE of the parameter spaces under the null and full models, respectively.

Kinship matrix

We used *u* as a random term to account for the sample relatedness. *u* follows the normal distribution *mvNormal(0, σK)*, where *K* was the kinship matrix of the samples. For each *K* entry, genotype-based relatedness for the sample pair, or IBS (identical by state) coefficient, was computed using the full spectrum of genomic variants in the ADSP samples. PLINK was used for fast kinship estimation on the massive genotype data.

$$k_{i,j} = \frac{1}{M}\sum_{m=1}^{M}(g_{m,i} * g_{m,j} + (1 - g_{m,i}) * (1 - g_{m,j}))$$

$k_{i,j}$ is the IBS relatedness between sample *i* and *j*. *M* is the variant number. $g_{m,i}$ and $g_{m,j}$ is the genotype of variant *m* in sample *i* and *j*, respectively.

Linear mixed models in the frequentist scheme

To compare the performances of our method to that of a LMM in the frequentist scheme in analyzing the ADSP dataset, we built a LMM as follow:

$$y_i = X_i\beta + u + e$$
$$u \sim mvN(0, \delta_g^2 K)$$
$$e \sim N(0, \delta_e^2 I)$$

y_i was the numerical mapping of the AD categories: no = 0, possible = 0.25, probable = 0.5, and definite = 1. *X* was the covariate matrix including age and sex, *u* was the random term, and *e* was the model residual. The LMM model was estimated with QTLRel in R [6].

Code availability

We deployed *Bayes-GLMM* as a GitHub repository for public use (https://github.com/xulong82/bayes.glmm).

Mouse strains, tissue harvesting and sectioning

All experiments involving mice were conducted in accordance with policies and procedures described in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health, and were approved by the Institutional Animal Care and Use Committee (IACUC) at The Jackson Laboratory. All mice were bred and housed in a 12/12 hours light/dark cycle. Six months old male C57BL/6J mice were injected intraperitoneally with a lethal quantity of ketamine/xylazine according to IACUC approved procedures. They were perfused with 1X PBS

(phosphate buffered saline) and whole brains were removed and fixed in 4% paraformaldehyde for two hours at 4˚C. Following fixation, the tissue was rinsed in 1X PBS, incubated in 10% sucrose for eight hours at 4˚C, then incubated in 30% sucrose overnight at 4˚C. Brains were then frozen in optimal cutting temperature (OCT) compound and stored at -80˚C until sectioning. Frozen brains were sectioned at 25µm and mounted on glass slides, which were stored at -80˚C until required for immunofluorescence staining.

Immunofluorescence

Brain sections were incubated overnight at 4˚C in the following primary antibodies: rabbit polyclonal anti-PDGFA (1:50, Bioss antibodies), sheep polyclonal anti-PRKAR1B (1:50, R&D Systems), goat anti-COL-IV (1:50, EMD Millipore), goat anti-CD31 (1:50, R&D Systems). Sections were immersed in deionized water for 3 minutes at 37˚C and then treated with 0.5mg/ml pepsin in 0.2N HCL for 15 minutes at 37˚C. Slides were then washed in 1X PBS twice for 10 minutes at room temperature. With the exception of anti-Col-IV, antibodies were diluted in 0.5% PBTB (1X PBS, .0.5% TritonX-100, 0.5% BSA (bovine serum albumin)) containing 10% normal donkey serum. Anti-Col-IV was diluted in 0.5% PBT. Sections were washed three times in 0.5% PBT then incubated for two hours at room temperature with their respective secondary antibodies (donkey anti-rabbit Alexa Fluor 594, donkey anti-goat Alexa Fluor 488, and donkey anti-sheep Alexa Fluor 594, 1:1000 dilution, Life Technologies). All sections were then counterstained with DAPI (1:1000 in 1X PBS) and then washed with 1X PBS prior to mounting with Aqua PolyMount. Images were taken using a Leica SP5 confocal microscope located within the Imaging facility at The Jackson Laboratory.

Results

Alzheimer's disease sequencing project

Development of *Bayes-GLMM* was motivated by the advent of the whole-genome sequencing association studies, such as the Alzheimer's disease sequencing project (ADSP) (www.niagads.org/adsp; Methods). ADSP was initiated to discover novel genomic variants for late-onset Alzheimer's disease. The whole genome sequence (WGS) cohorts of ADSP contained 570 participants from 111 families. This family-based design generated profound sample relatedness that warranted a mixed model approach. Furthermore, 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 categorical model. Family pedigree, race, ethnicity, age, sex, and *APOE ε2/ε3/ε4* 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 *APOEε3* comprised 56.7% (N = 323) of the population, followed by 35.1% (N = 200) of *APOEε3*/*APOEε4*, 6.84% (N = 39) of *APOEε2*/*APOEε3*, 1.05% (N = 6) of *APOEε2*/*APOEε4*, and 0.351% (N = 2) of *APOEε2*/*APOEε2* (Figure 1). Individuals homozygous for *APOEε4* were excluded from the study.

Additive effects of age, sex (female), and *APOE* allele-types (*ε2*, *ε3*, *ε4*), together with the cut points parameters of the ordered categorical model, were tested with *Bayes-GLMM* (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 that followed a multivariate normal distribution (Methods). Model parameters were estimated by MCMC sampling. As expected, we observed that the *APOEε4* allele significantly increased risk of Alzheimer's (p = 0.00014) while the *APOEε2* allele reduced risk (p = 0.0033) relative to the baseline *APOEε3* allele. Sex was also a significant factor, with females at higher risk (p = 0.032). Increasing age corresponded to a small but significant risk increase (p = 0.00036). The small effect size of age was a result of multiple factors: (1) the relatively large values for age as a model predictor (67-80); (2) a narrow age range; and (3) the possible 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*

The ADSP consortium identified a total of 27.9 million SNP from the WGS cohort, of which 10.3 million passed their quality check and had minor allele frequency greater than 0.01 (Supplementary Figure 2). Associations of the 10.3 million SNP to AD status were tested by *Bayes-GLMM* in two steps (Figure 3). In the first step, a generalized linear model (ordered categorical model) was applied to each of the 10.3 million variants without the random term. The purpose of this step was a preliminary screen for potential candidate variants. Model parameters were estimated by the maximal likelihood estimation (MLE) method for computational efficiency. Variants with $p < 0.0001$ were identified as potential candidate variants (N = 9726, Figure 4A). In the second step, candidate variants from the first step were tested with the full GLMM, including the random term to address sample relatedness. Model parameters were estimated by MCMC sampling to avoid the instability in estimating GLMM by MLE. Final p-values for every variant were obtained from their empirical posterior distributions (Figure 4B).

Top LOAD-associated variants from ADSP WGS

We identified four variants in three independent loci with $p < 5 \times 10^{-8}$ and 55 variants in 28 loci with $p < 1 \times 10^{-6}$ (Table 1). 52 out of the top 55 variants increased LOAD risk. Furthermore, variants with strong effects tended to occur at lower allele frequency, suggesting that these variants might be under negative selection (Figure 5). The top 55 variants mapped to 146 genomic annotations using Ensembl Variant Effect Predictor (variants commonly mapped to multiple annotations): 73 were in introns, 31 were in intergenic regions, 27 were upstream of genes (within 5 kb upstream from the 5' end), 11 were downstream of genes (within 5 kb downstream from the 3' end), and four were regulatory regions (Supplementary Table 1). The 73 intronic annotations mapped to 19 variants and 18 unique genes. Twelve out of the 18 genes appeared in the NHGRI GWAS catalog as disease-associated [2] (Supplementary Table 2). Associated traits of the 12 genes included obesity-related traits (*PTPRD*, *SORCS2*, and*SLC24A4*), Alzheimer's disease (*SLC24A4* and *GABRG3*), acute lymphoblastic leukemia (*ERC2* and *ST6GALNAC3*), adiponectin levels (*CMIP* and *HIVEP2*), bipolar disorder and schizophrenia (*ERC2*), and type-2 diabetes (*PTPRD*).

The four genome-wide significant variants ($p < 5 \times 10^{-8}$) were all intergenic: rs10490263, rs74944275, rs149372995, rs140233081. These SNPs are located as follows: rs10490263 is 233,714 bp upstream of *SLC8A1* and 337 bp upstream of lincRNA *AC007317.1*; rs74944275 is 111,711 downstream of *C5orf30* and 18,568 bp downstream of lincRNA *CTD-2154H6.1*; rs140233081 and rs149372995 are in LD and locate in between *PRKAR1B* and *PDGFA*. Additionally, these final two SNPs are 8,097 and 8,292 bp downstream of *PRKAR1B*, and 21,254 and 21,059 bp upstream of *PDGFA*, respectively. To assess the functional relevance of the four variants, we queried the Roadmap Epigenomics [29] and ENCODE [30] resources using HaploReg [31] for chromatin state and protein binding annotations. We found rs10490263 lies in promoter-associated histone marks in the hippocampus and circulating T cells, and rs74944275 lies in both promoter- and enhancer-associated histone marks in multiple brain regions. Furthermore, rs149372995 resides in a candidate-binding site of CTCF, rs74944275 resides in a candidate-binding site of CCNT2, Evi-1, GATA, and HDAC2, rs140233081 and rs149372995 lie in candidate bindings sites of NERF1a, SMC3, and TCF12.

Given the role of CTCF in genome organization and possible gene regulation, we further examined the flanking genes *PRKAR1B* and *PDGFA*. We localized the expression of protein products of these two genes using immunofluorescence. Both PRKAR1B and PDGFA have widespread expression in the mouse brain, but are particularly localized to glia-vascular structures (Figure 6). This could be significant given the recent data suggesting glia-vascular

alterations may predispose individuals to or occur very early in LOAD [32-34]. Furthermore, we evaluated RNA sequence data from the dorsolateral prefrontal cortex of participants in the Religious Order Study (ROS) and Rush Memory and Aging Project (MAP) studies, two longitudinal cohort studies of aging with prospective brain autopsy [35-37]. In these human data, we found that higher *PDGFA* transcript level is moderately correlated with greater neuritic plaque burden ($P = 0.005$, transcriptome-wide FDR = 0.03; $\beta > 0$) [38], suggesting that the *PDGFA* association with AD may relate to a role in the accumulation of one of the two key pathologic features of AD.

Table 1: Top 55 variants with $P < 1 \times 10^{-6}$. Variants in italics met standard genome-wide significance of $P < 5 \times 10^{-8}$.

RSID	CHR	POSITION	REF	ALT	MAF	Effect Size	Std. Dev	*P*
rs74944275	*5*	*102726073*	*C*	*T*	*0.019*	*2.371*	*0.394*	*1.76×10^{-9}*
rs10490263	*2*	*40973289*	*C*	*T*	*0.469*	*0.697*	*0.116*	*2.15×10^{-9}*
rs149372995	*7*	*580540*	*A*	*G*	*0.051*	*1.532*	*0.269*	*1.18×10^{-8}*
rs140233081	*7*	*580735*	*C*	*A*	*0.056*	*1.396*	*0.255*	*4.26×10^{-8}*
rs139258867	11	33422464	C	T	0.017	2.487	0.461	6.89×10^{-8}
rs11709639	3	94975203	T	A	0.325	0.645	0.120	8.61×10^{-8}
rs75841969	12	127335883	G	A	0.052	1.464	0.275	1.05×10^{-7}
rs72720587	4	137323780	C	T	0.019	2.328	0.441	1.27×10^{-7}
rs2018116	14	92831272	T	C	0.768	0.636	0.121	1.41×10^{-7}
rs141404567	11	33393524	G	A	0.017	2.514	0.481	1.70×10^{-7}
rs2010568	2	118395972	G	C	0.429	0.646	0.125	2.38×10^{-7}
rs144152209	7	585369	C	A	0.060	1.349	0.262	2.49×10^{-7}
rs74917009	5	169915787	G	A	0.027	1.971	0.382	2.54×10^{-7}
rs144990130	7	582328	G	A	0.061	1.324	0.257	2.65×10^{-7}
rs12685122	9	9206006	T	G	0.211	0.867	0.169	2.95×10^{-7}
rs73046027	3	19950385	C	T	0.132	0.963	0.188	3.18×10^{-7}
rs7463321	8	20523821	T	C	0.167	0.860	0.168	3.23×10^{-7}
rs148758667	9	130665077	G	T	0.060	1.456	0.285	3.30×10^{-7}
rs72618491	3	94938828	G	A	0.335	0.606	0.119	3.31×10^{-7}
rs117662279	7	155362626	G	A	0.018	2.214	0.434	3.34×10^{-7}
rs1280103	4	187526002	C	A	0.396	-0.627	0.123	3.63×10^{-7}
rs7856285	9	18973653	G	A	0.609	0.612	0.120	3.65×10^{-7}
rs17383917	3	94984650	T	C	0.330	0.640	0.126	4.08×10^{-7}
rs11124760	2	41001812	C	T	0.437	0.639	0.126	4.15×10^{-7}
rs61768273	1	44509818	A	T	0.034	1.790	0.354	4.25×10^{-7}

RSID	CHR	POSITION	REF	ALT	MAF	Effect Size	Std. Dev	*P*
rs17383687	3	94963466	C	T	0.329	0.653	0.129	4.33×10^{-7}
rs12497549	3	20072654	C	T	0.157	0.898	0.178	4.52×10^{-7}
rs36147593	15	27587764	A	G	0.110	1.119	0.222	4.86×10^{-7}
rs10933941	3	94965589	G	A	0.329	0.635	0.126	4.98×10^{-7}
rs116407196	5	102973337	A	G	0.035	1.804	0.360	5.25×10^{-7}
rs7122488	11	21874253	T	C	0.646	0.633	0.126	5.26×10^{-7}
rs12639003	3	94966599	A	G	0.329	0.637	0.127	5.35×10^{-7}
rs12549162	8	20547331	C	G	0.167	0.872	0.174	5.50×10^{-7}
rs62483581	7	106726214	G	A	0.451	0.587	0.117	5.52×10^{-7}
rs12485639	3	94940998	C	A	0.320	0.604	0.121	5.55×10^{-7}
rs67822265	2	53715939	C	T	0.289	0.665	0.133	5.70×10^{-7}
rs17383861	3	94983399	G	A	0.331	0.635	0.127	5.91×10^{-7}
rs9826288	3	95044652	C	T	0.665	-0.618	0.124	6.27×10^{-7}
rs2478319	13	48111575	A	G	0.689	0.596	0.120	6.33×10^{-7}
rs72720573	4	137257730	T	C	0.018	2.416	0.485	6.46×10^{-7}
rs140419591	11	33327476	A	G	0.017	2.516	0.506	6.54×10^{-7}
rs78491489	7	44335828	C	T	0.113	0.934	0.188	6.80×10^{-7}
rs6689933	1	76837471	C	T	0.684	0.633	0.128	7.47×10^{-7}
rs17263248	3	55574820	A	G	0.200	0.804	0.163	7.54×10^{-7}
rs10435819	9	9197298	G	A	0.194	0.849	0.172	7.58×10^{-7}
rs61446477	3	94964689	A	G	0.329	0.639	0.129	7.73×10^{-7}
rs72720589	4	137333269	T	A	0.017	2.233	0.452	7.80×10^{-7}
rs7978950	12	47361547	C	T	0.412	0.611	0.124	7.92×10^{-7}
rs2176276	3	94989440	C	A	0.329	0.613	0.124	8.13×10^{-7}
rs1359665	13	48097289	G	A	0.684	0.642	0.130	8.16×10^{-7}
rs4849593	2	118369787	G	A	0.407	0.652	0.132	8.20×10^{-7}
rs72618501	3	94974822	T	C	0.330	0.633	0.129	8.66×10^{-7}
rs61768270	1	44498974	C	T	0.034	1.788	0.364	8.88×10^{-7}
rs1861305	2	40950582	A	G	0.468	0.589	0.120	8.96×10^{-7}
rs4367173	4	7383470	C	G	0.195	-0.664	0.136	9.58×10^{-7}

Integrating prior knowledge

Prior knowledge integration is a prominent feature of Bayesian modeling. In GWAS, prior information of a variant can be implemented with multiple strategies, each allowing posterior estimations to carry different weights of the priors. In *Bayes-GLMM*, we implemented a method to configure priors that targeting the unique challenges of GWAS, such as the different meanings of effect sizes from studies with different statistical models, and the particularly small

p-values from published large-scale studies. Our method took the reported standardized effect sizes as the prior information and integrated them into the hierarchical model of each variant effect (Methods). To demonstrate the performance of this method, we generated a binary phenotypic trait (coded as 0 or 1) and genotypic trait of a variant (coded as 0, 1, or 2) by Monte Carlo, and used a logistic regression model (LR) to test their associations. To illustrate the ability of *Bayes-GLMM* to integrate this information, we assessed the effect of prior information on the estimated variant effect by testing a range of prior standardized effect sizes. This method of prior configuration effectively modulates the information from the data (Figure 7), regardless of the differences between the prior information and the data in hand.

Discussion

We created a new GWAS method, *Bayes-GLMM*, and applied it on ADSP's whole-genome sequencing cohort. This method efficiently addresses three major challenges in GWAS: categorical phenotypes, population structure and sample relatedness, and prior knowledge integration. Furthermore, our generalized approach has the flexibility to operate on binary and quantitative traits in addition to ordered categorical phenotypes. These features enabled our identification of four new candidate variants in three loci that significantly increased the risk of Alzheimer's disease.

Out of the four new genome-wide significant candidate variants, rs140233081 and rs149372995 are in LD and locate in between *PRKAR1B* and *PDGFA* that are potentially relevant to vascular dysfunction. Recent evidence suggests that vascular dysfunction is a critical component of AD pathology [32-34] and potentially a necessary predisposing feature [39]. Further, vascular dysfunction has been shown to be necessary for the development of Alzheimer's-like phenotypes in a mouse model of amyloid pathology [40]. We have localized PDGFA and PRKAR1B to specific components of vascular anatomy. Our immunofluorescence shows PDGFA expression between the collagen-rich tunica external and the endothelium of the tunica intima, supporting the presence of PDGFA in vascular smooth muscle cells (VSMCs). Previous studies have shown PDGF to affect VSMC proliferation by inducing a phenotypic switch from a contractile state to a proliferative one [41]. Insufficient PDGFA expression, then, may impair vascular regeneration following plaque-related insults, thereby exacerbating AD. This potential mechanism paired with increased *PDGFA* under amyloid burden expression suggests the two candidate variants could reduce necessary PDGFA expression when plaques are present, thereby attenuating the increase in *PDGFA* we observed with amyloid burden. PRKAR1B was

seen in a punctate fashion suggesting the presence of cytoplasmic clusters of the protein, and we hypothesize that the PRKAR1B puncta represent accumulation of protein kinase A (PKA) at either the endoplasmic reticulum or the insulin receptor. Calcium release from the endoplasmic reticulum is typically suppressed by phospholamban (PLN), however such suppression is lifted following PLN phosphorylation by PKA. Changes in the regulation of calcium release due to altered *PRKAR1B* expression may very well have important consequences for AD, including but not limited to changes in vascular smooth muscle contraction that limit circulation to plaque-burdened brain regions. In addition to its calcium-related role, PKA is essential for signal transduction following activation of the insulin receptor, a process that has been shown to be the mechanism by which PDGF induces phenotypic switching in VSMCs [42]. In this way, changes in PRKAR1B may yield corresponding changes in circulation through suppressed arterial muscle contractility or through a direct influence on vascular growth and maintenance.

We consider our method *Bayes-GLMM* to be an important addition to the existing GWAS toolkit. The flexibility of Bayesian modeling allows the convenient configuration of sophisticated models, such as our GLMM. In *Bayes-GLMM*, logistic and ordered logistic regression likelihoods were used to model binary and ordered categorical variables, respectively. Conditional factors were included as model covariates and, although our study was underpowered for epistasis analysis, interaction terms can be straightforwardly included. Sample relatedness was modeled by a random term that followed a multivariate normal distribution. Model parameters can be estimated by either L-BFGS maximal likelihood estimation (MLE), or Hamilton Markov chain Monte Carlo sampling, as implemented in Stan.

Although the MLE implementation in *Bayes-GLMM* was efficient and reliable in estimating generalized linear models, it was unreliable in estimating generalized linear mixed models. We found that MLE of the random term was skewed toward initial values, suggesting the optimizer was trapped into local optima and limiting reliability in estimating the GLMM. On the other hand, the 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.

Bayes-GLMM method was optimized in multiple ways to minimize the computational expense: (1) support parallel computing; (2) conjugate prior distributions; (3) vectorization of model statements to exploit efficient matrix operations in Stan; and (4) parameterization of multivariate

normal distribution for the random effect by Cholesky factoring. Nevertheless, efficiency was still the primary drawback of MCMC sampling. Testing at a 2.3G Hz Intel processor, MLE took roughly 0.12 seconds to estimate the GLM model per variant of the ADSP dataset (Methods, Figure 3). In comparison, the MCMC sampler took roughly 30 seconds to generate 1000 samples for the same GLM model, and 15 minutes to process 1000 samples for the GLMM model. Our pre-scan with MLE followed by more precise estimation by MCMC proved a practical approach to overcome these processing limitations when applying *Bayes-GLMM* in GWAS.

To reduce the computational burden in fitting GLMMs, we suggest that categorical diagnoses could be collapsed into binary variables. For the ADSP data, the "no" and "possible" diagnoses become "control", while the "possible" and "definite" diagnoses are "case". Logistic mixed models, or binary mixed models, were implemented in *Bayes-GLMM* to accommodate binary variables. The MCMC sampler implemented in Stan took approximately ten minutes to collect 1000 samples for parameters of such a binary mixed model, as opposed to 15 minutes for the four-level categorical mixed model. Alternatively, the recently released GMMAT (generalized linear mixed model association test) method that utilized penalized quasi-likelihood method to fit a binary mixed model was significantly faster than the MCMC sampling approach [5]. However, this practice of collapsing the categorical variable reduced precision due to the information loss in simplifying multiple categories. We tested this practice in the ADSP data, and found the association results by binary-GLMM and categorical-GLMM showed substantial disagreement (Supplementary Figure 3).

Another strategy to reduce computational requirements is to transform categorical variables into continuous variables to accommodate efficient LMM methods [5, 43]. 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. This practice also yields incorrect effect estimates due to the unbalanced sampling in different phenotypic categories, which is prominent in the ADSP study in which the "probable" diagnosis accounted for 62% of the total while the other three categories accounted for only 10-14%. We also found the inferences results of LMM by QTLRel were sensitive to different quantitative coding of categorical variables (Supplementary Figure 4; Methods). Taking rs34827707 as an example, the likelihood ratio test (LRT) value for rs34827707 dropped from 29 to 15 by changing the coding from

no/possible/probably/definite as 0/0.25/0.5/1 to 0/0.33/0.66/1. In contrast, the GLMM robustly estimated three cut points to separate the four categories.

Bayesian modeling naturally allows the integration of prior information by specifying model parameter's prior distribution. However, how to best specify a variant's prior information is an open question when the prior study does not precisely match the experiment design in hand. Association results of each variant in a GWAS are commonly reported by effect size and *p*-value. While critical in describing the association strength, exact values of effect sizes are often specific to the given study because of dependencies on the statistical model, genotype coding strategies, and covariates. Therefore, it can be misleading to use the reported effect sizes to configure the priors. As opposed to effect sizes, *p*-values that quantify deviation from a null hypothesis can be less specific to the given study. However, *p*-values are strongly influenced by the sample size, and *p*-values from a large-scale study as priors would dominate the posterior estimation of a variant's association, thereby masking 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 (Figure 6).

While powerful, *Bayes-GLMM* has several drawbacks. First, the quantitative meaning of parameter values is not readily interpretable in terms of fractional effects. Second, heritability estimation is elusive due to a difficulty in estimating residual variance. Third, as implemented, only one variance component is supported. Although Bayesian modeling can readily encompass multiple variance components, this becomes impractical for GWAS due to computational limitations for most researchers. Fourth, sampling-based estimations remain computationally intensive and may not be suitable for larger data sets (*e.g.* the full set of ADSP variants). We expect that advances in model estimation techniques, improved algorithms, and broad application of cloud-based computational resources will alleviate these problems in the near future.

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. The ability to integrate existing GWAS as prior information can further power these studies to prioritize specific variants at known loci.

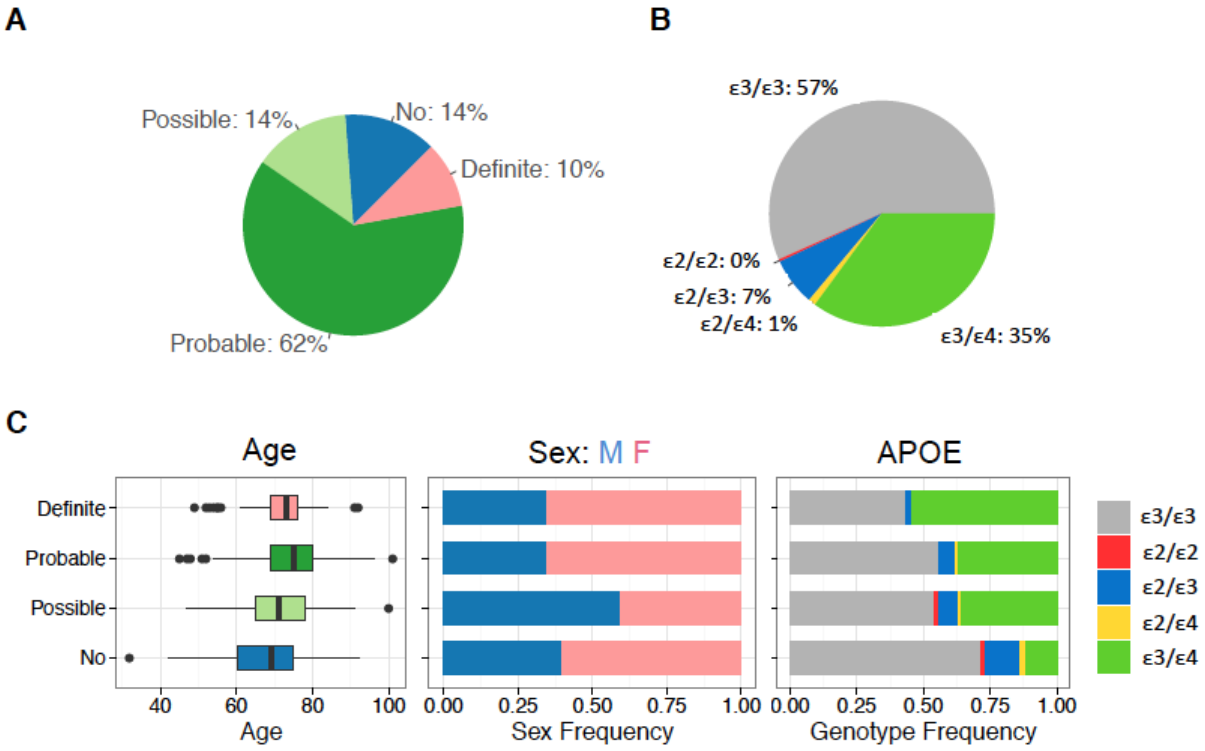
Acknowledgments

We thank B. Carpenter and A. Gelman of the Stan Development Team for assistance, G. Churchill for helpful discussions, and M. Miller and the ADSP Consortium for assisting with data resources. This work was funded by National Institute for Aging AG054345 (GWC, GRH); National Institute for Aging P30AG10161, R01AG15819. R01AG179917, R01AG36836, and U01AG46152 (DAB); The Pyewacket Foundation (GWC); and The Jackson Laboratory Director's Innovation Fund (GWC, GRH).

Author Contributions

GWC, GRH, XW, and MS designed the study. XW, VMP, GA, AM, CAS, PJM, GWC, and KRMK imported and analyzed data. SRC, CA, and GRH performed mouse experiments. XW, GWC, and GRH primarily wrote the manuscript, with additional contributions from all other authors.

FIGURES



515

Figure 1. Summary statistics of the ADSP whole genome sequencing (WGS) cohort. **A.** AD diagnosis for 570 individuals across 111 families. **B.** APOE allele-type composition. **C.** Age distributions of individuals in each AD diagnostic category (left), sex composition in each category (middle), and *APOE* allele-type composition in each category (right).

520

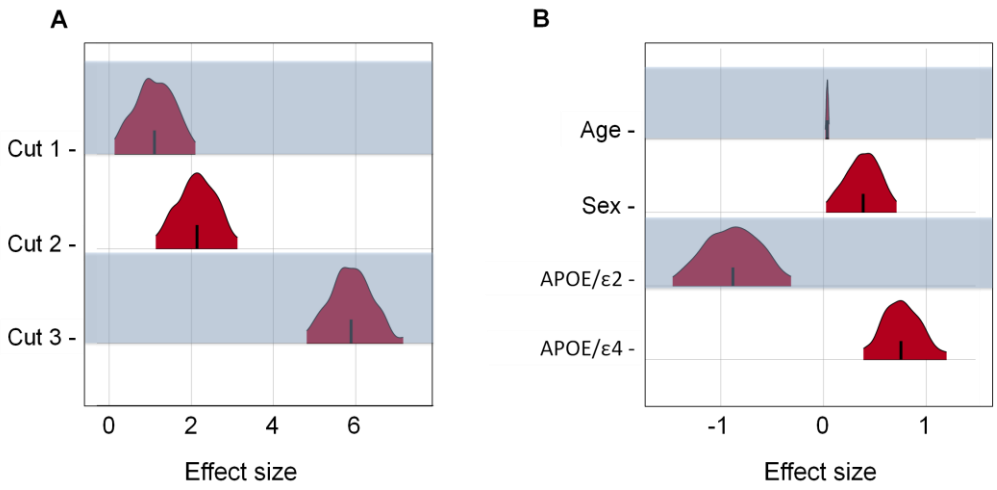


Figure 2. Bayes-GLMM estimation of model parameters by MCMC sampling of GLMM. **(A)** Posterior distributions of the ordered categorical model's cut points. **(B)** Posterior distributions of the model covariate's effect sizes: age, sex, *APOEε2* and *APOEε4*.

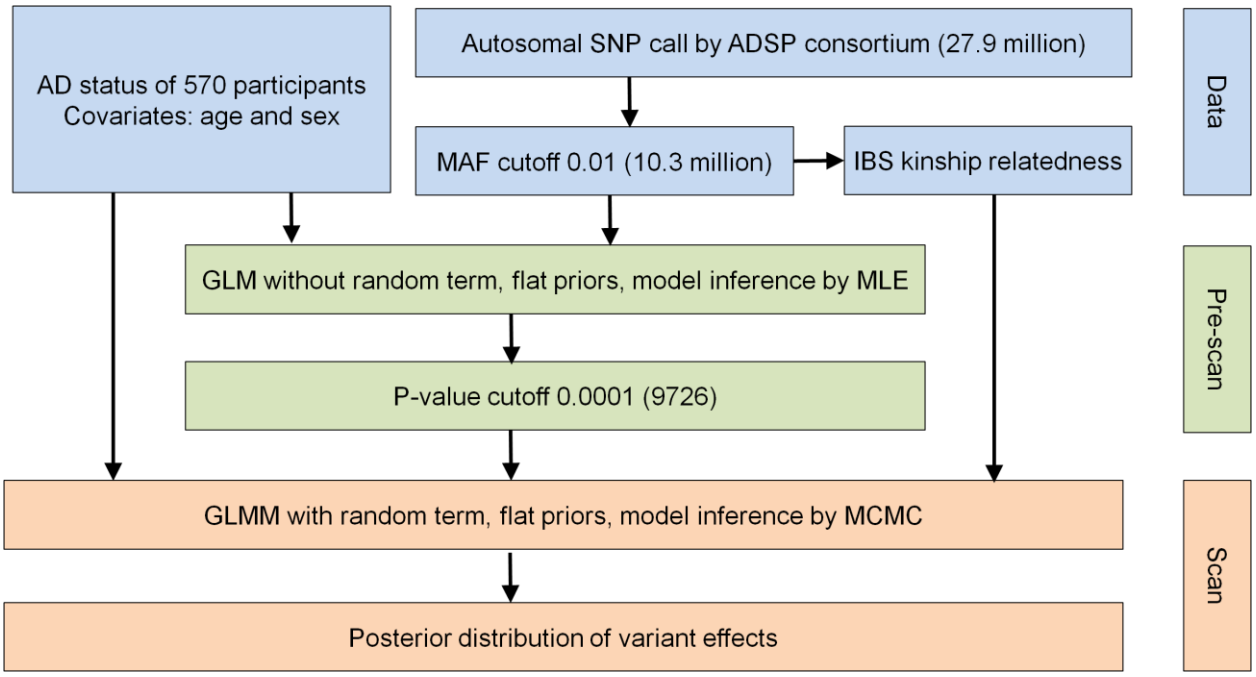
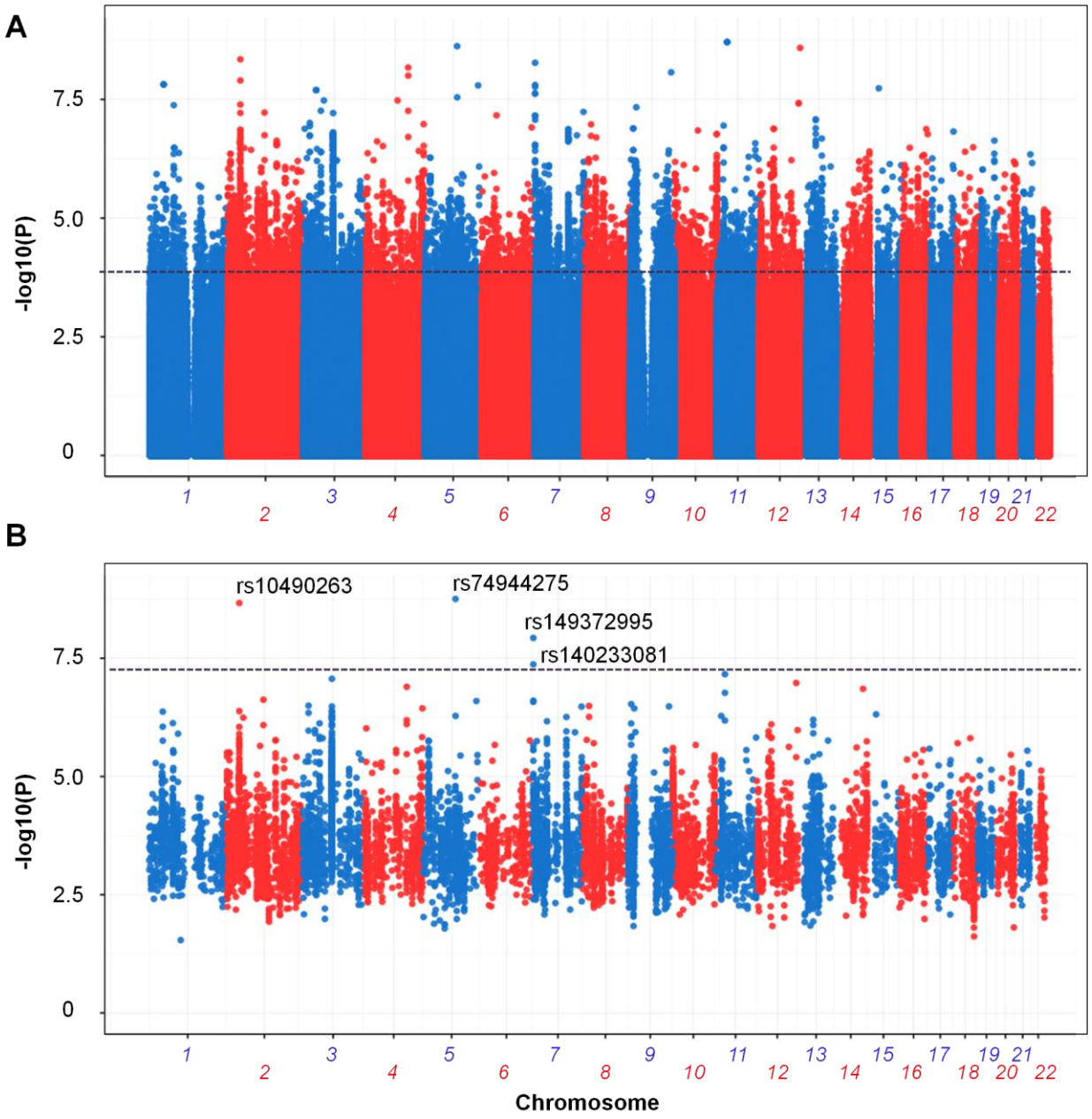


Figure 3. Analysis overview of two-step GWAS analysis using *Bayes-GLMM*. Initial data (blue) were filtered and pre-scanned with a fixed linear model (green). Results were filtered by significance and scanned using the full GLMM (orange).



530

Figure 4. Association results for ADSP WGS cohort by *Bayes-GLMM*. **(A)** Results for 10.3 million genomic variants by *Bayes-GLMM* without kinship correction. Model parameters were estimated by MLE. Variants with $p < 0.0001$, above the dashed line, were chosen for the full scan (9726 variants). **(B)** GWAS on filtered variants by GLMM with kinship correction. Model
535 parameters were estimated by MCMC sampling. Dashed line was the cutoff of genome wide significance ($p < 5 \times 10^{-8}$).

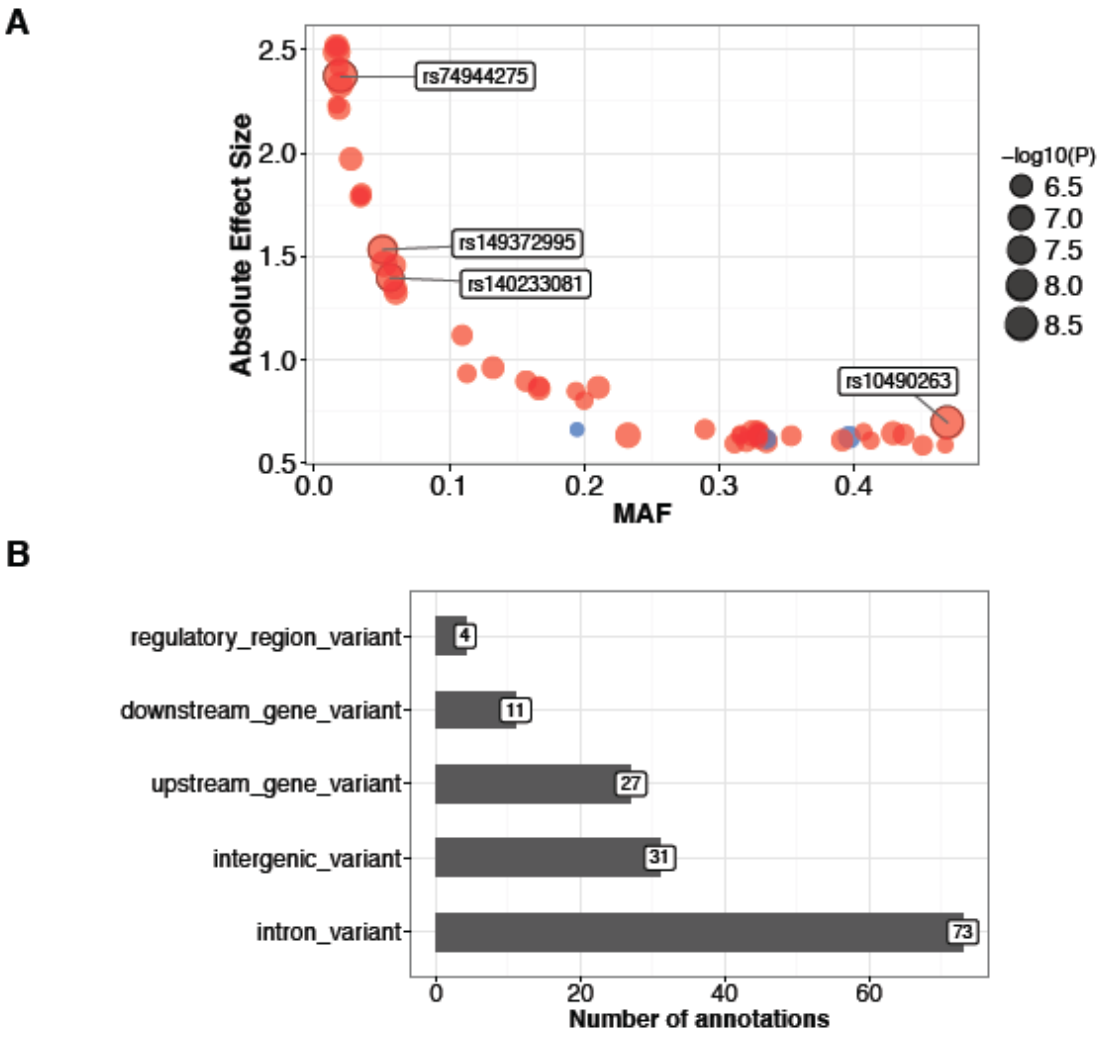


Figure 5. Effect sizes and consequences of top variants. **(A)** Allele frequencies and effect sizes for all variants with *Bayes-GLMM* dervied $p < 1 \times 10^{-6}$. Positive-effect (*i.e.* risk-increasing) loci are in red and negative-effect loci (*i.e.* protective) in blue. **(B)** Functional consequences of the top variants.

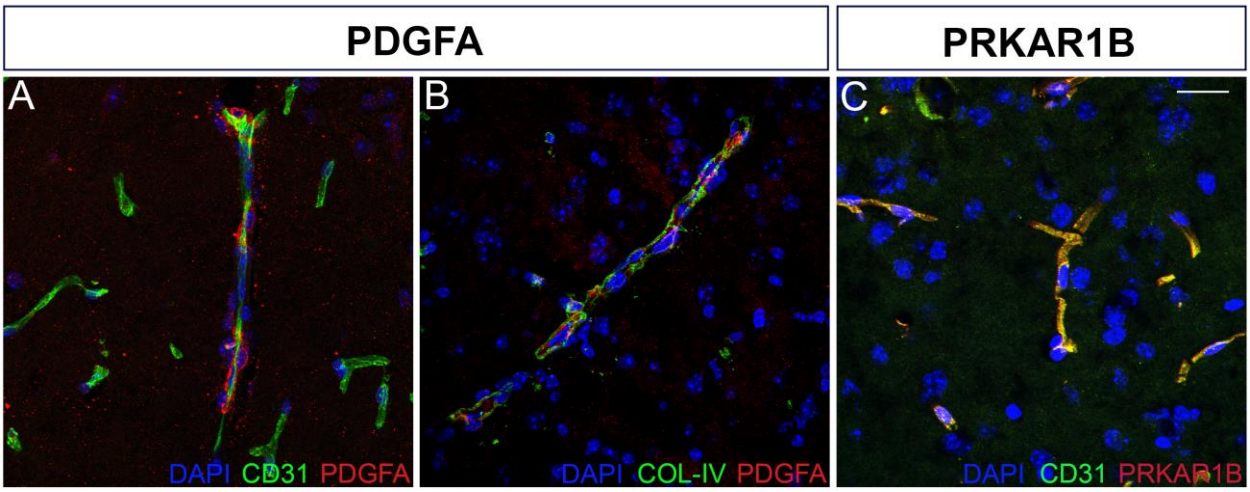


Figure 6. PDGFA and PRKAR1B localize to vascular structures in the mouse brain.
(A & B) PDGFA (red) shows close localization to endothelial cells (CD31, A) and basement membrane (COL-IV, B), components of the vascular substructure. **(C)** PRKAR1B (red) shows punctate expression in the region of blood vessels (CD31, green). See materials and methods for antibody details. Scale bar = 20 μm.

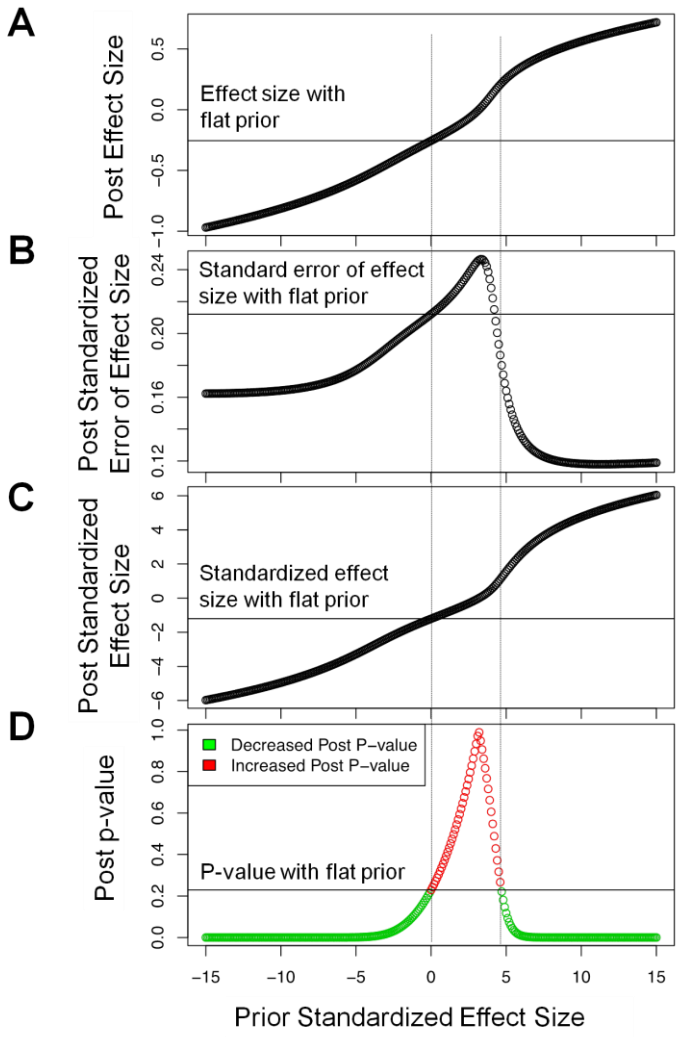


Figure 7. Effects of priors on **(A)** posterior effect size; **(B)** posterior standardized error of the effect size; **(C)** posterior standardized effect size; and **(D)** posterior p-values. X-axis denotes prior standardized effect size. The grey horizontal line in each graph is the respective posterior estimation when the prior standardized effect size is equal to 0. The two vertical dashed lines define a range of prior standardized effect size that increased the posterior *p*-value compared to a flat prior.

References

1. Manolio, T.A., *Genomewide association studies and assessment of the risk of disease.* N Engl J Med, 2010. **363**(2): p. 166-76.

2. Welter, D., et al., *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1001-6.
3. The Genomes Project, C., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
4. Henderson, C.R., *Estimation of Variance and Covariance Components.* Biometrics, 1953. **9**(2): p. 226-252.
5. Chen, H., et al., *Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models.* Am J Hum Genet, 2016. **98**(4): p. 653-66.
6. Cheng, R., et al., *QTLRel: an R package for genome-wide association studies in which relatedness is a concern.* BMC Genet, 2011. **12**: p. 66.
7. Gatti, D.M., et al., *Quantitative trait locus mapping methods for diversity outbred mice.* G3 (Bethesda), 2014. **4**(9): p. 1623-33.
8. Gilmour, A.R., R. Thompson, and B.R. Cullis, *Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models.* Biometrics, 1995. **51**(4): p. 1440-1450.
9. Kang, H.M., et al., *Efficient control of population structure in model organism association mapping.* Genetics, 2008. **178**(3): p. 1709-23.
10. Lippert, C., et al., *FaST linear mixed models for genome-wide association studies.* Nat Methods, 2011. **8**(10): p. 833-5.
11. Zhang, Z., et al., *Mixed linear model approach adapted for genome-wide association studies.* Nat Genet, 2010. **42**(4): p. 355-60.
12. Zhou, X. and M. Stephens, *Efficient multivariate linear mixed model algorithms for genome-wide association studies.* Nat Methods, 2014. **11**(4): p. 407-9.
13. Kavvoura, F.K. and J.P. Ioannidis, *Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls.* Hum Genet, 2008. **123**(1): p. 1-14.
14. Carpenter, B., et al., *Stan: A Probabilistic Programming Language.* 2017, 2017. **76**(1): p. 32.
15. Gelman, A., et al., *Bayesian Data Analysis*. Third Edition ed. CRC Texts in Statistical Science. 2013: Chapman and Hall/CRC. 675.
16. Newcombe, P.J., et al., *Multilocus Bayesian meta-analysis of gene-disease associations.* Am J Hum Genet, 2009. **84**(5): p. 567-80.
17. Stephens, M. and D.J. Balding, *Bayesian statistical methods for genetic association studies.* Nat Rev Genet, 2009. **10**(10): p. 681-90.

18. Verzilli, C., et al., *Bayesian meta-analysis of genetic association studies with different sets of markers.* Am J Hum Genet, 2008. **82**(4): p. 859-72.
19. Bertram, L. and R.E. Tanzi, *Thirty years of Alzheimer's disease genetics: the implications of systematic meta-analyses.* Nat Rev Neurosci, 2008. **9**(10): p. 768-78.
20. Guerreiro, R., et al., *TREM2 variants in Alzheimer's disease.* N Engl J Med, 2013. **368**(2): p. 117-27.
21. Jonsson, T., et al., *Variant of TREM2 associated with the risk of Alzheimer's disease.* N Engl J Med, 2013. **368**(2): p. 107-16.
22. Harold, D., et al., *Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease.* Nat Genet, 2009. **41**(10): p. 1088-93.
23. Hollingworth, P., et al., *Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease.* Nat Genet, 2011. **43**(5): p. 429-35.
24. Jones, L., et al., *Genetic evidence implicates the immune system and cholesterol metabolism in the aetiology of Alzheimer's disease.* PLoS One, 2010. **5**(11): p. e13950.
25. Jun, G., et al., *Meta-analysis confirms CR1, CLU, and PICALM as alzheimer disease risk loci and reveals interactions with APOE genotypes.* Arch Neurol, 2010. **67**(12): p. 1473-84.
26. Lambert, J.C., et al., *Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease.* Nat Genet, 2009. **41**(10): p. 1094-9.
27. Lambert, J.C., et al., *Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease.* Nat Genet, 2013. **45**(12): p. 1452-8.
28. Naj, A.C., et al., *Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease.* Nat Genet, 2011. **43**(5): p. 436-41.
29. Bernstein, B.E., et al., *The NIH Roadmap Epigenomics Mapping Consortium.* Nat Biotechnol, 2010. **28**(10): p. 1045-8.
30. *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
31. Ward, L.D. and M. Kellis, *HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants.* Nucleic Acids Res, 2012. **40**(Database issue): p. D930-4.
32. Bell, R.D., *The imbalance of vascular molecules in Alzheimer's disease.* J Alzheimers Dis, 2012. **32**(3): p. 699-709.
33. Montagne, A., et al., *Brain imaging of neurovascular dysfunction in Alzheimer's disease.* Acta Neuropathol, 2016. **131**(5): p. 687-707.

34. Zhao, Z., et al., *Establishment and Dysfunction of the Blood-Brain Barrier.* Cell, 2015. **163**(5): p. 1064-78.
35. Bennett, D.A., et al., *Overview and findings from the religious orders study.* Curr Alzheimer Res, 2012. **9**(6): p. 628-45.
36. Bennett, D.A., et al., *Overview and findings from the rush Memory and Aging Project.* Curr Alzheimer Res, 2012. **9**(6): p. 646-63.
37. De Jager, P.L., et al., *Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci.* Nat Neurosci, 2014. **17**(9): p. 1156-63.
38. Lim, A.S., et al., *Diurnal and seasonal molecular rhythms in human neocortex and their relation to Alzheimer's disease.* Nat Commun, 2017. **8**: p. 14931.
39. Iturria-Medina, Y., et al., *Early role of vascular dysregulation on late-onset Alzheimer's disease based on multifactorial data-driven analysis.* Nat Commun, 2016. **7**: p. 11934.
40. Soto, I., et al., *Meox2 haploinsufficiency increases neuronal cell loss in a mouse model of Alzheimer's disease.* Neurobiol Aging, 2016. **42**: p. 50-60.
41. Owens, G.K., M.S. Kumar, and B.R. Wamhoff, *Molecular regulation of vascular smooth muscle cell differentiation in development and disease.* Physiol Rev, 2004. **84**(3): p. 767-801.
42. Zhao, Y., et al., *PDGF-induced vascular smooth muscle cell proliferation is associated with dysregulation of insulin receptor substrates.* Am J Physiol Cell Physiol, 2011. **300**(6): p. C1375-85.
43. Kang, H.M., et al., *Variance component model to account for sample structure in genome-wide association studies.* Nat Genet, 2010. **42**(4): p. 348-54.
44. Pawitan, Y., *In All Likelihood: Statistical modeling and inference using likelihood*. 2001, Oxford: Oxford University Press. 525.