**Title: A GWAS of late-onset Alzheimer’s disease using whole genome sequencing data**

**Abstract**

**Introduction**

**Methods**

Whole genome sequences of 576 individuals were measured by the NIH Alzheimer’s disease sequencing project (ADSP). Each individual was diagnosed with one of the four AD levels: no, possible, probable, definite. Additional sample information includes family relatedness, race, age, sex, and APOE allele types.

To detect significant variants for late-onset Alzheimer’s disease (LOAD), we built a generalized linear mixed model (GLMM). AD levels were modeled by an ordered categorical variable. Probability for an individual to fall in the th (j = 1, 2, 3) and any lower categories follows: , where denotes the probability that the th individual falls in category . was logit-transformed and denoted as: , where provides each AD level a unique intercept. is a vector of explanatory variables. is the corresponding effect sizes vector. is a multivariate variable that follows: , with covariance matrix the genotype-based relatedness between individual pairs (IBS). , where is variant number, and is the genotype of individual *i* and *j* on variant *m.* We used u as a random term to account for population structure in the GWAS. 23 kinship matrices were computed by the taking-one-off strategy, in that for any given variant of one chromosome, the corresponding kinship matrix was computed by taking off all variants of the given chromosome. KING was used for fast kinship estimation on the massive genotype data (Manichaikul et al., 2010).

The GLMM was built in Bayesian framework and implemented with Stan (<http://mc-stan.org>). A non-informative prior distribution, *cauchy(0, 1),* was assigned for each parameter. Point estimations of model parameters were obtained by maximizing model’s joint posterior likelihood. Full posterior distributions of model parameters were obtained by sampling using the No-U-Turn sampler (Hoffman and Gelman 2011).

To identify genomic variants that explain the most variance, log-likelihood ratio test (LRT) was computed by subtracting the posterior log-likelihood of the null model from that of the full model for each variant. A total of 22 null models were estimated using the same covariates (age, sex, APOE/e2, APOE/e4) together with one of the 22 taking-one-off kinship matrices. LOD score was approximated as two times the LRT. Significant LOD value was determined empirically by permutation as follow: (1) randomly choose 1 million SNP; (2) randomly permute the chosen SNP over samples; (3) estimate the 1 million models with each permutated SNP and compute the LRT; (4) save the maximal LRT value; (5) repeat steps 1-4 3000 times. This leads to 3000 maximal LRT values. Significant threshold was set as 0.95 quantile of the 3000 maximal LRT values.

A full Bayesian sampling was applied on 193,676 variants that fall in 500 kilo-bp ranges of the 73 significant peaks by maximal likelihood estimation. WAIC (Watanable-Akaike information criterion) was quantified as -2 \* elpd (expected log pointwise predictive density) for each variant to describe model fitness after Bayesian sampling, where (Vehtari and Gelman, 2014).

Three features can characterize our model: (1) the response was modeled by an ordered categorical variable, (2) population structure was modeled by a random multivariate variable with known covariance and estimated by each variant independently, (3) Bayesian framework and inference was applied. To validate the model results, we modified our model to fit established inference tools to compare the results. We constructed a linear mixed model (LMM): , where *N(AD)* was numerical by transforming the categorical AD status: 0/no, 0.25/possible, 0.5/probable, and 1/definite, , . The LMM model was estimated with QTLRel in R (Cheng et al., 2011).

R was used for most analysis except specified (www.r-project.org). Genomic variants were processed with Unix shell, Python, and Plink (Purcell et al., 2007). Variant effects were estimated with Ensembl variant effect predictor (VEP).

**Results**

Whole genome sequencing of 576 individuals over 111 families from ADSP was used in our GWAS. In ADSP, each individual’s AD status was clinically examined in one of four levels: no (80), possible (81), probable (360), and definite (55). Additional sample information includes sex, age, and APOE genotypes on three APOE allele types: e2/e3/e4. Frequencies of APOE e2, e3, and e4 are 4%, 78%, and 18%, respectively. Homozygous APOE e2 and APOE e4 are rare (2 and 0 samples). Interquartile range of sample ages is 67 to 80. This age range is typical for late-onset AD; hereby constrain our study in LOAD. Interestingly, while 61% of the samples are female, female proportions in the definite and probable AD groups are higher, 65.5% and 65.6%, respectively (Figure 1).

We defined on average 4.8 million short variants (SV) for each sample (4.4 to 5.6 million), and 45.4 million unique SV in total from all 576 samples. This accounts for 1.5% of the genome. Variants with MAF < 0.01 were excluded from GWAS for being lack of statistical power. This leaves 12.6 million SNPs and 1.5 million INDEL (Supplementary Figure 1). By stratifying variants according to genetic structures, we found intron regions have higher variant density than exon and intergenic regions. There are 5.6 variants per kilo-base in intron regions, but only 3.2 and 2.5 in the intergenic and exon regions. This suggests introns are more vulnerable to random mutation, hereby more relevant with disease development. Functional consequences of the 14.1 variants were predicted (Supplementary Figure 2). Results showed 51.8% of the total consequences were intron-related, followed by noncoding transcript variants (14.1%) and intergenic variants (9.78%).

To identify the most risky and protective LOAD variants, we developed a generalized linear mixed model (GLMM) in Bayesian framework. Population structure was modeled by including a random variable whose covariance follows the population’s kinship structure (Figure 2a, Supplementary Figure 3). APOE/e2/e4 genotypes, sex, and age were included as covariates. The protective and risky effects of APOE/e2 and APOE/e4 were both significant by estimating a model with the random term and covariates. The 95% CI for APOE/e2 and e4 were -1.45 to -0.26 and 0.42 to 1.2, respectively. Female showed as a significant risky factor of LOAD, with 95% CI of its effect 0.00 to 0.74 (Figure 2b). Surprisingly, while the risky effect of age is significant, the effect size is rather trivial (95% CI 0.02 to 0.05). This reflects that ADSP chose older peoples in general, with interquartile range 67 to 80. It is also likely that aging is not an AD causal factor. Further, no significant interactions between the covariates were observed by including an additional interaction terms for any covariates pairs (Supplementary Figure 4).

A GWAS was set for the 14.1 million variants using the GLMM (Figure 3a, 3b). Both of the fixed and random effects were estimated for each variant independently by maximal likelihood estimation (MLE). A LOD threshold for each chromosome was determined empirically by permutation (Figure 3c). A total of 244 genomic variants in 73 independent LD blocks passed the LOD threshold of 0.05 FDR (Figure 4a, 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. To obtain the full distributions of model parameters, MCMC sampling was applied on variants surrounding each peak of the 73 LD blocks (± 250 kilo-bp). Effect sizes estimated by MLE and sampling (first mode of the posterior distribution) are similar, with correlation coefficient 0.86 (Figure 4b). WAIC was used to quantify model fitness from Bayesian sampling. WAIC pattern by MCMC matched well with LOD by MLE in any of the 73 regions (Figure 4c).

The 244 variants lead to 1101 genetic variants, 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.

Genetic mechanisms of Alzheimer’s disease likely involve epistasis processes, in which interactions of different genomic loci play a role. Given the significant effect of APOE/e2, APOE/e4, and sex, we expanded our model to search for variants that interact with each of them. To identify sex-interacting variants, an interaction term between sex and variant was added for a GWAS. Significant interactions were obtained by comparing LOD of the models with and without the interaction term (Figure 6). APOE/e2 and APOE /e4-interacting variants were searched by the same way. 86 variants showed an additional increase of 15 or more in LOD by including the interaction term with APOE/e4. By comparison, 572 variants showed the same level of LOD increment by including the interaction term with APOE/e2 (Figure 6). The 86 APOE/e4-interacting variants lead to 148 genetic consequences, while the 572 APOE/e2-interacting variants lead to 1790 genetic consequences, including one missense case by rs966384 (Figure 7). The *G* to *A* mutation of rs966384 causes a *P* to *S* amino acid change in LRG1 protein. LRG1 was associated with normal pressure hydrocephalus.

**Discussion**

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

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

Linear mixed models (LMM) are commonly used in similar setup. To understand how our model performed differently from a LMM, we transformed the categorical AD status into numerical probabilities and estimated the LMM with QTLRel (Cheng et al., 2011). Results by LMM and GLMM are similar in general (Supplementary Figure 10a). Pearson’s correlation coefficient of LOD by LMM and GLMM was 0.96. Meanwhile, by comparing the top 0.1% variants in either model (16,681), the LMM increased LOD by 1.97 on average. Interquartile range of LOD differences between LMM and GLMM for the top 0.1% variants were 0.88 to 2.86. By inspecting the variants that returned the most different LOD values (top 76 with minimal LOD difference 8) by the two models, we found: (1) these variants are all rare (Supplementary Figure 10b). MAF of the top 76 variants were 0.010 to 0.036. (2) MAF of these variants across the four AD populations varied irregularly. Taking rs34827707 for example, while the minor allele appeared frequently in the definite AD population (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 (Supplementary Figure 10c). These two facts pointed to an issue of arbitrarily transforming the AD categories into numerical probabilities in LMM. Indeed, LMM is sensitive to different coding rules when (1) the variants are rare and (2) MAF across different AD categories varied irregularly. LOD value for rs34827707 dropped from 29 to 15 by changing the current coding rule into no/0, possible/0.33, probable/0.66, and definite/1. As comparison, three independent cut points were estimated in the GLMM for each variant. The smaller LOD values by GLMM reflect the fact that our statistical power for cut points estimation was much compromised for these variants.