Association for ADSP whole-genome sequence data

Overview of the Model

We performed variant calls on 576 whole-genome sequences from ADSP (Figure 1). Each individual was diagnosed with one of four AD levels: no, possible, probable, and definite. We used a generalized linear mixed model (GLMM) for association, for the following reasons: (1) an ordered categorical model (rather than a linear model) best fit the data structure; (2) a mixed model corrects for population structure by genotype-based relatedness; and (3) a Bayesian inference was possible by MCMC sampling. In addition to the random effect for relatedness and a fixed effect for each SNP, we also included fixed effects for age, sex, and APOE allele type (e2/e3/e4) covariates. In total, we assessed 12.6 million SNPs and 1.5 million indels with MAF > 0.01 in this population.

Details of the Model

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 jth (j = 1, 2, 3) and any lower categories follows: $P(Y_i \leq j) = \pi_{i1} + ... + \pi_{ij}$, where π_{ij} denotes the probability that the ith individual Y_i falls in category j. $P(Y_i \leq j)$ was logit-transformed and denoted as: $\text{logit}\big(P(Y_i \leq j)\big) = \theta_j - X_i\beta + u$, where $\{\theta_j\}$ provides each AD level a unique intercept. X_i is a vector of explanatory variables. β is the corresponding effect sizes vector. u is a multivariate variable that follows: $N(0, K\sigma_g^2)$, with covariance matrix K the genotype-based relatedness between individual pairs (IBS, Figure 2).

$K_{ij} = \frac{1}{M}\sum_{m=1}^{M}(G_{i,m} \cdot G_{j,m} + (1 - G_{i,m}) \cdot (1 - G_{j,m}))$, where M is variant number, $G_{i,m}$ and $G_{i,m}$ 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 $elpd = \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i \mid \theta^s)\right) - \sum_{i=1}^{n} V_{s=1}^{S} (\log (p(y_i \mid \theta^s))$ (Vehtari and Gelman, 2014).

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): $\mathrm{N(AD)} = \mathrm{X_i}\beta + \mathrm{u} + \mathrm{e}$, where *N(AD)* was numerical by transforming the categorical AD status: 0/no, 0.25/possible, 0.5/probable, and 1/definite, $u \sim \mathrm{N}(0,\ \mathrm{K}\sigma_g^2)$, $e \sim \mathrm{N}(0,\ \mathrm{I}\sigma_e^2)$. The LMM model was estimated with QTLRel in R (Cheng et al., 2011). Results by LMM and GLMM are similar in general (Figure 3). Pearson's correlation coefficient of LOD by LMM and GLMM was 0.96. Meanwhile, by comparing the top 0.1% variant in either model (16,681), the LMM increased LOD by 1.97 on average. The interquartile range of LOD differences between LMM and GLMM for the top 0.1% variants were 0.88 to 2.86.

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

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 lack of statistical power, leaving 12.6 million SNPs and 1.5 million indels (Figure 4). 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. Results showed 51.8% of the total consequences were intron-related, followed by noncoding transcript variants (14.1%) and intergenic variants (9.78%).

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 5). 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.

A GWAS was set for the 14.1 million variants using the GLMM (Figure 6a, 6b). Both of the fixed and random effects were estimated for each variant independently by maximal likelihood estimation (MLE). A total of 244 genomic variants in 73 independent LD blocks passed the LOD threshold of 15 (Figure 7a). 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 0.99 (Figure 7b). 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 7c).

Figure Legends

Figure 1. Summary statistics of the ADSP whole genome sequencing project participants. **A.** AD diagnosis for 576 individuals across 111 families. **B.** Age distributions of individuals in each AD status. **C.** APOE allele-type composition. **D.** APOE allele-type composition in each AD status. **E.** Sex composition. **F.** Sex composition in each AD status.

Figure 2. IBS kinship relatedness of individual pairs. In and out mean in and out of the same family from the recorded pedigree.

Figure 3. Pairwise scatterplot of LOD by GLMM and LMM. GLMM: generalized linear mixed model. LMM: linear mixed model.

Figure 4. Summary statistics of the ADSP variants. 12.6 million bi-allelic SNP **(A)** and 1.5 million bi-allelic INDEL **(B)**, with MAF cutoff 0.01, were included in the GWAS.

Figure 5. Estimation of model covariates by sampling. Points and error bars are the first mode and and 95% CI of posterior distributions for each covariate.

Figure 6. (A) GWAS Manhattan. **(B)** QQ plot of GWAS LOD from true and randomized genotypes of each variant.

Figure 7. (A) 244 variants with LOD value at least 15 by the additive model. **(B)** Effect sizes of variants surrounding the first peak by point estimation (x) and sampling (y). **(C)** WAIC and LOC profiles in the region.

Figure 1

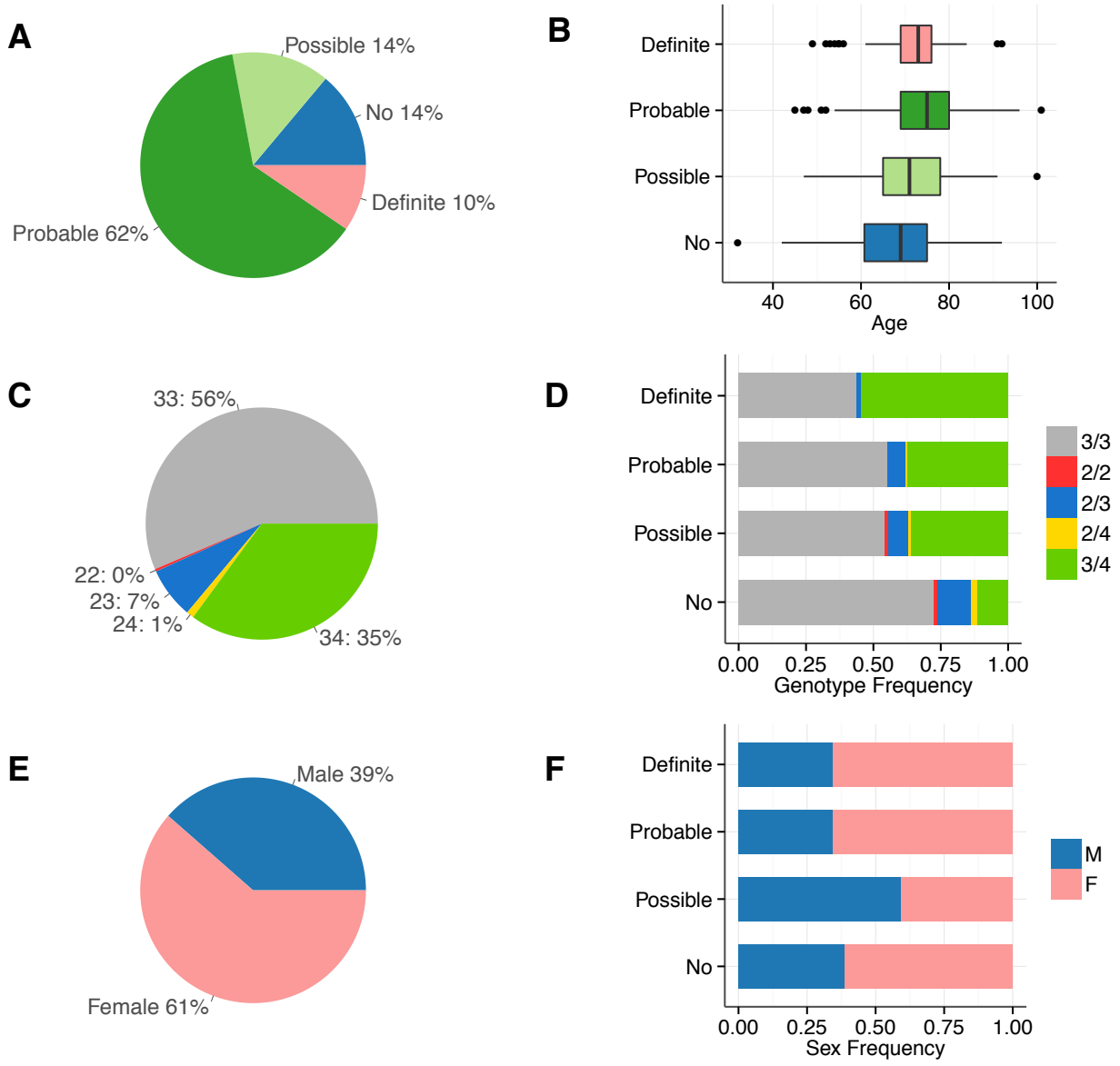

A
Possible 14%
No 14%
Definite 10%
Probable 62%
B
Definite
Probable
Possible
No
40
60
80
100
Age
C
33: 56%
22: 0%
23: 7%
24: 1%
34: 35%
D
Definite
Probable
Possible
No
0.00
0.25
0.50
0.75
1.00
Genotype Frequency
3/3
2/2
2/3
2/4
3/4
E
Male 39%
Female 61%
F
Definite
Probable
Possible
No
0.00
0.25
0.50
0.75
1.00
Sex Frequency
M
F

Figure 2

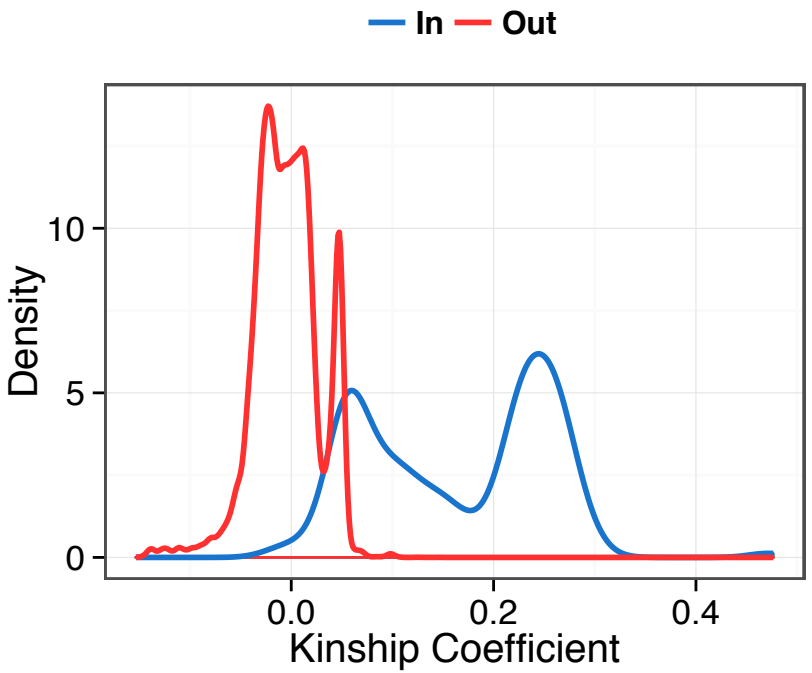


Figure 3

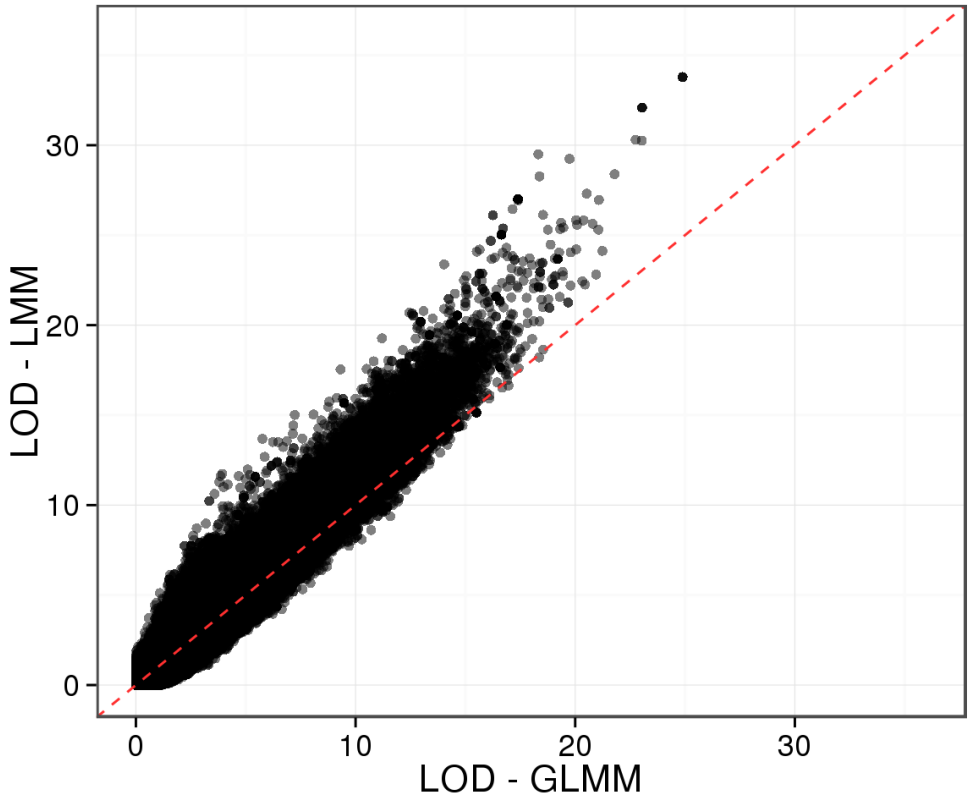


Figure 4

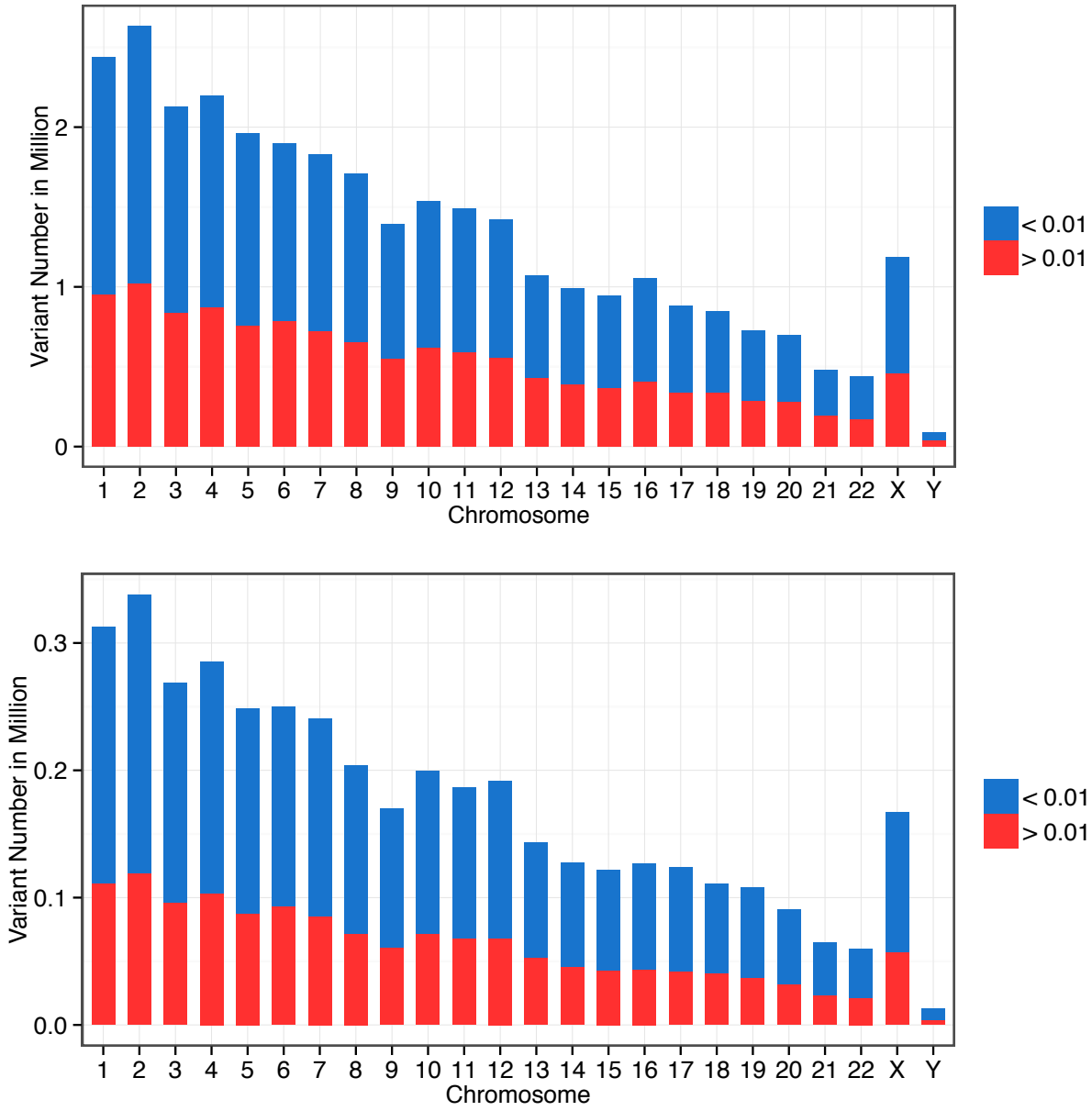

Variant Number in Million
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Chromosome
< 0.01
> 0.01
Variant Number in Million
0.3
0.2
0.1
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Chromosome
< 0.01
> 0.01

Figure 5

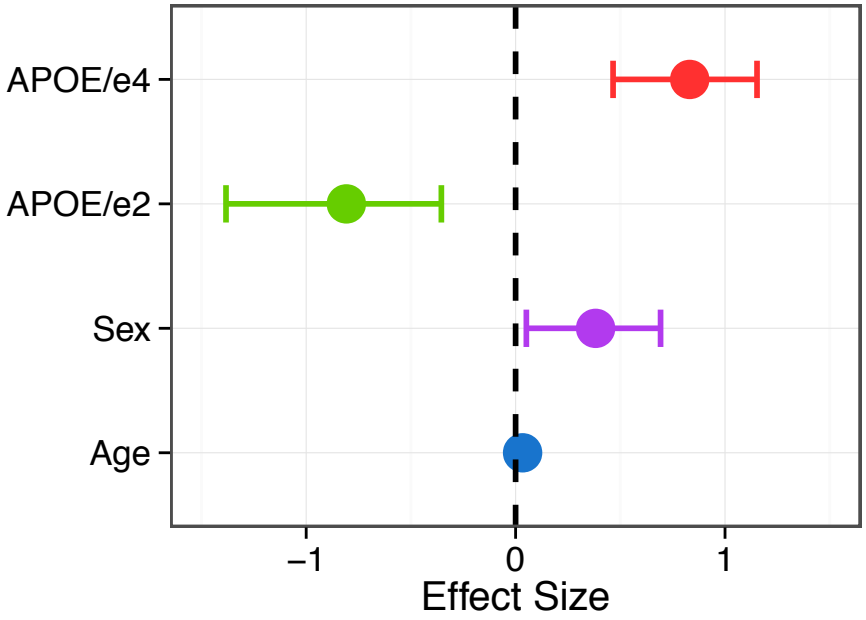


Figure 6

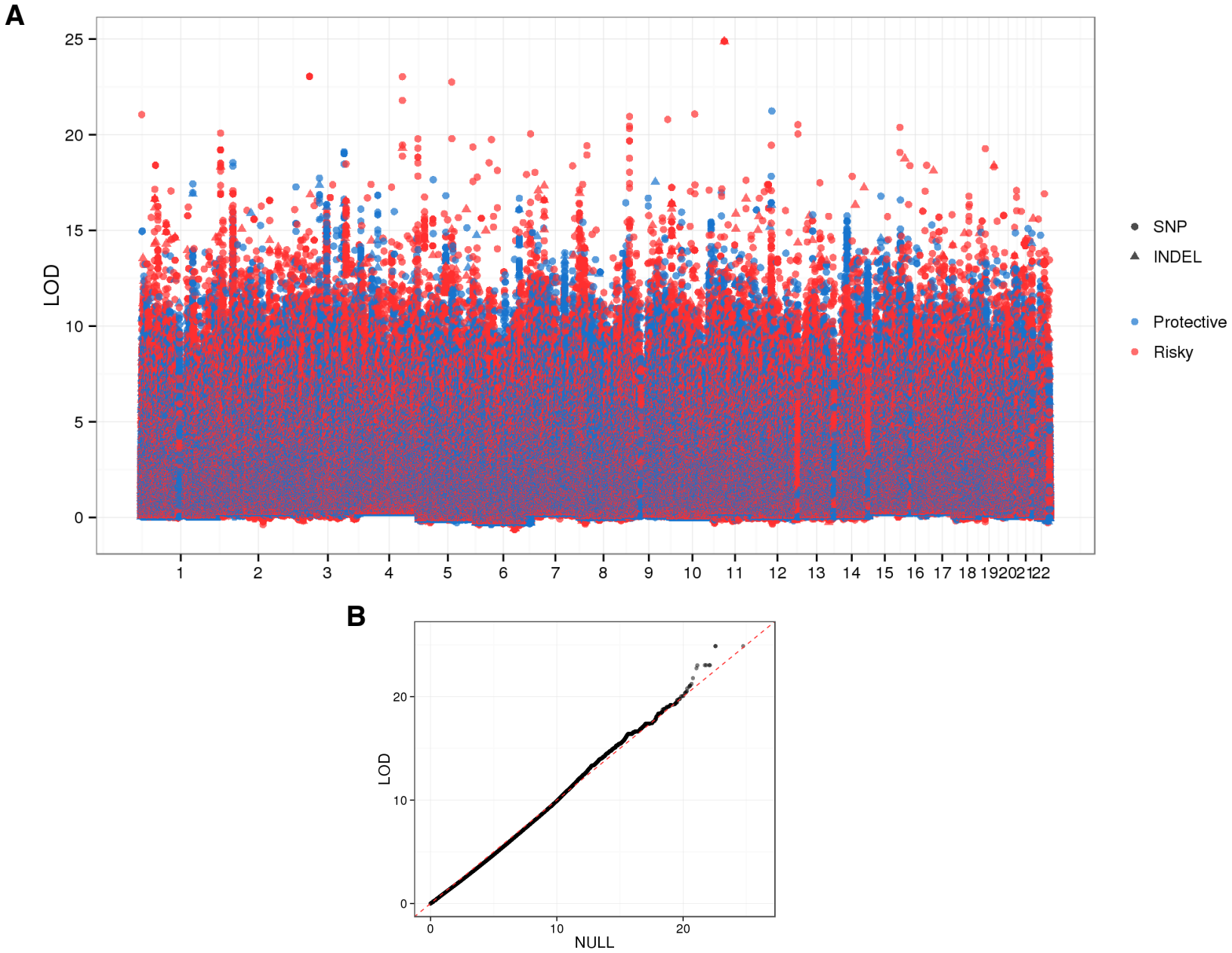

A
LOD
25
20
15
10
5
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
SNP
INDEL
Protective
Risky
B
LOD
20
10
0
0
10
20
NULL

Figure 7

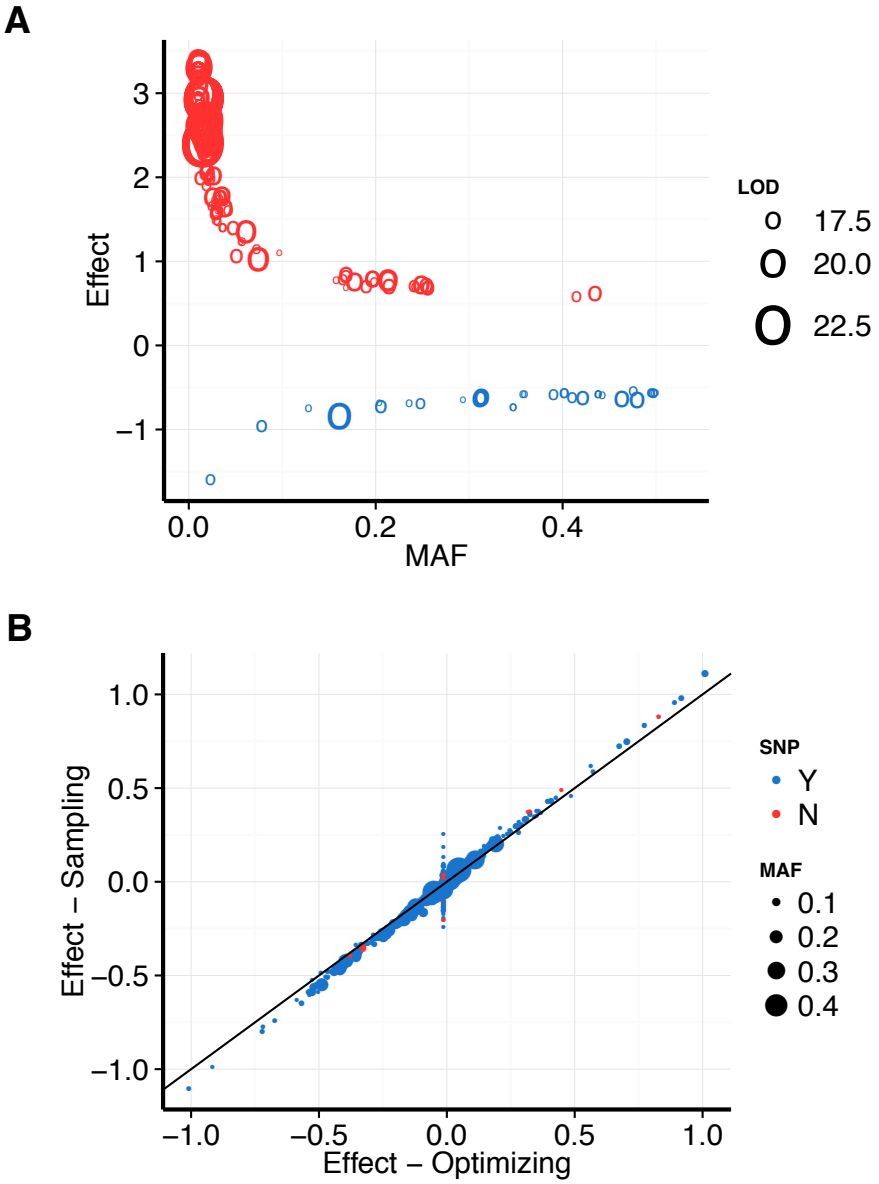
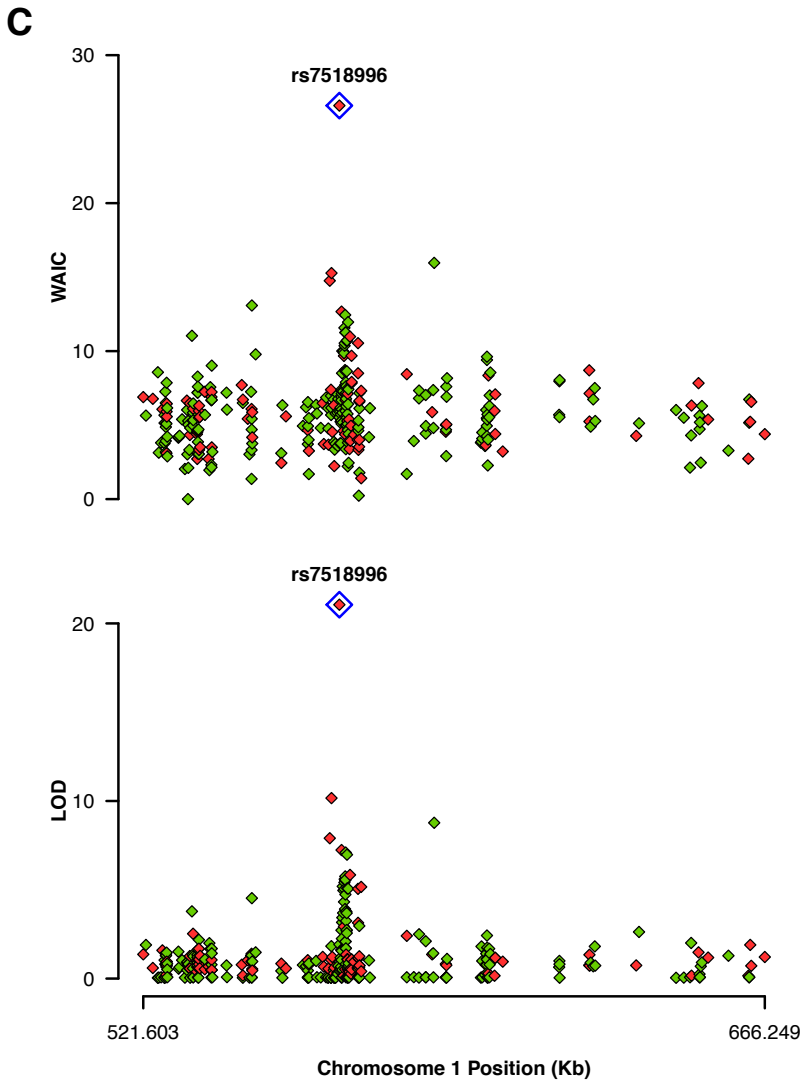

A
Effect
MAF
LOD
17.5
20.0
22.5
B
Effect – Sampling
Effect – Optimizing
SNP
Y
N
MAF
0.1
0.2
0.3
0.4

Figure 7


C
rs7518996
30
20
10
0
WAIC
rs7518996
20
10
0
LOD
521.603
666.249
Chromosome 1 Position (Kb)