Gene-Based Rare Allele Analysis Identified a Risk Gene of Alzheimer’s Disease

Jong Hun Kim¹, Pamela Song², Hyunsun Lim³, Jae-Hyung Lee⁴, Jun Hong Lee¹, Sun Ah Park⁵, for the Alzheimer’s Disease Neuroimaging Initiative

Introduction
Alzheimer’s disease (AD) is a leading cause of dementia and is known to have high heritability (as high as 60–80%) [1,2]. Genome-wide association studies (GWAS) have identified several known risk genes excluding APOE are not clinically useful. In various complex diseases, gene studies have targeted rare alleles for unsolved heritability. Our study aims to elucidate previously unknown risk genes for AD by targeting rare alleles. We used data from five publicly available genetic studies from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) and the database of Genotypes and Phenotypes (dbGap). A total of 4,171 cases and 9,358 controls were included. The genotype information of rare alleles was imputed using 1,000 genomes. We performed gene-based analysis of rare alleles (minor allele frequency≤3%). The genome-wide significance level was defined as meta \( P<1.8\times10^{-6} \) (0.05/number of genes in human genome = 0.05/28,517). ZNF628, which is located at chromosome 19q13.42, showed a genome-wide significant association with AD. The association of ZNF628 with AD was not dependent on APOE ε4. APOE and TREM2 were also significantly associated with AD, although not at genome-wide significance levels. Other genes identified by targeting common alleles could not be replicated in our gene-based rare allele analysis. We identified that rare variants in ZNF628 are associated with AD. The protein encoded by ZNF628 is known as a transcription factor. Furthermore, the associations of APOE and TREM2 with AD were highly significant, even in gene-based rare allele analysis, which implies that further deep sequencing of these genes is required in AD heritability studies.

Abstract
Alzheimer’s disease (AD) has a strong propensity to run in families. However, the known risk genes excluding APOE are not clinically useful. In various complex diseases, gene studies have targeted rare alleles for unsolved heritability. Our study aims to elucidate previously unknown risk genes for AD by targeting rare alleles. We used data from five publicly available genetic studies from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) and the database of Genotypes and Phenotypes (dbGap). A total of 4,171 cases and 9,358 controls were included. The genotype information of rare alleles was imputed using 1,000 genomes. We performed gene-based analysis of rare alleles (minor allele frequency≤3%). The genome-wide significance level was defined as meta \( P<1.8\times10^{-6} \) (0.05/number of genes in human genome = 0.05/28,517). ZNF628, which is located at chromosome 19q13.42, showed a genome-wide significant association with AD. The association of ZNF628 with AD was not dependent on APOE ε4. APOE and TREM2 were also significantly associated with AD, although not at genome-wide significance levels. Other genes identified by targeting common alleles could not be replicated in our gene-based rare allele analysis. We identified that rare variants in ZNF628 are associated with AD. The protein encoded by ZNF628 is known as a transcription factor. Furthermore, the associations of APOE and TREM2 with AD were highly significant, even in gene-based rare allele analysis, which implies that further deep sequencing of these genes is required in AD heritability studies.
regions are more difficult to obtain and interpret, than genotyping of a few loci.

Improvement of imputation methods has allowed accurate inference of rare alleles [26]. According to 1000 genomes study [20], the mean squared Pearson correlation coefficients ($R^2$) between rare SNPs (MAF 0.5%–5%) and imputed dosages were 0.7–0.9 in the European ancestry. Furthermore, mutational loads of rare alleles within genes obtained from imputation can confer high power [27]. In this study, we aimed to find risk genes for AD using gene-based analysis of rare alleles deduced from 1000 genomes and publicly available GWAS data.

Materials and Methods

Subjects

We used publicly available GWAS data from the Alzheimer’s Disease Neuroimaging Initiative (ADNI), Genetic Alzheimer’s Disease Associations (GenADA) study, Electronic Medical Records and Genomics (eMERGE), the National Institute on Aging Late Onset Alzheimer’s Disease (NIA-LOAD) family study, and the Framingham study. ADNI data were obtained from https://ida.loni.ucla.edu. GenADA (dbGaP accession number: phs000219.v1) [28,29], eMERGE (dbGaP accession number: phs000234.v1), NIA-LOAD (dbGaP accession number: phs000168.v1), and the Framingham study (dbGaP accession number: phs000007.v1) data were downloaded from dbGaP (http://www.ncbi.nlm.nih.gov/gap). Subjects with European ancestry were included. After genotypic quality control (QC), missing phenotypic data exclusion, and ethnic group selection, 4171 cases and 9358 control were included in this study. Summaries about the studies are shown in Table 1. Additional information for each study were detained in File S1. The institutional review board of Ilsan hospital approved our study. Written informed consent was given by participants. In addition patient records were anonymized prior to analysis.

Genotypic QC and imputation

We excluded alleles with low (<1%) MAF, low (<95%) call rate, and deviation of Hardy-Weinberg Equilibrium ($P < 10^{-6}$). The subjects with low (<95%) call rates, too high autosomal heterozygosity (false discovery rate, FDR > 1%) and too high relatedness (identical-by-state, IBS > 0.95) were excluded. For genotypic QC, we used the GenABEL package, v 1.69 [30]. After estimating haplotypes using SHAPEIT, v 1.0 [31], imputation with multi-population reference panels of 1000 genomes (phase I, release Mar 2012) was executed using IMPUTE2, v 2.2 with default parameters [32,33]. We discarded imputated SNPs with INFO < 0.4. The dosage data of imputation were used for further analyses. The dosage means the expected genotype score [34].

Statistical analyses

In the association study, we adjusted for age, sex, years of education, and significant principle components (PCs) of the genetic stratification (File S1). For consistency across studies, years of education were categorized as follows: 1, ≤ 4; 2, 4 < and ≤ 10; 3, 11 < and ≤ 15; 4, >15 years according to the established methods of stratifications in the GenADA study. We imputed

![Figure 1. The overall scheme of this study.](doi:10.1371/journal.pone.0107983.g001)
Table 2. The highly ranked seven genes in the first meta-analyses.

<table>
<thead>
<tr>
<th>Gene</th>
<th>CHR</th>
<th>Start</th>
<th>Meta P</th>
<th>Meta Z</th>
<th>Meta direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZNF628</td>
<td>19</td>
<td>55909286</td>
<td>6.0</td>
<td>5.2</td>
<td>+</td>
</tr>
<tr>
<td>APOE</td>
<td>19</td>
<td>45409038</td>
<td>4.8</td>
<td>4.1</td>
<td>+</td>
</tr>
<tr>
<td>TOMM40</td>
<td>19</td>
<td>45394477</td>
<td>4.3</td>
<td>4.1</td>
<td>+</td>
</tr>
<tr>
<td>NAPRT1</td>
<td>8</td>
<td>14656556</td>
<td>3.9</td>
<td>4.0</td>
<td>+</td>
</tr>
<tr>
<td>TREM2</td>
<td>6</td>
<td>41126245</td>
<td>3.9</td>
<td>4.1</td>
<td>+</td>
</tr>
<tr>
<td>3 CBLB</td>
<td>3</td>
<td>105438891</td>
<td>3.8</td>
<td>4.1</td>
<td>+</td>
</tr>
</tbody>
</table>

* Larger absolute score represents smaller Z P and the direction of the score represents the direction of risk [35].

The highly ranked seven genes in the first meta-analyses.

<table>
<thead>
<tr>
<th>Gene</th>
<th>CHR</th>
<th>Start</th>
<th>Meta P</th>
<th>Meta Z</th>
<th>Meta direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>APOE</td>
<td>19</td>
<td>45409038</td>
<td>4.3</td>
<td>4.1</td>
<td>+</td>
</tr>
<tr>
<td>TOMM40</td>
<td>19</td>
<td>45394477</td>
<td>4.0</td>
<td>3.7</td>
<td>+</td>
</tr>
<tr>
<td>NAPRT1</td>
<td>8</td>
<td>14656556</td>
<td>3.9</td>
<td>4.1</td>
<td>+</td>
</tr>
<tr>
<td>TREM2</td>
<td>6</td>
<td>41126245</td>
<td>3.9</td>
<td>4.1</td>
<td>+</td>
</tr>
<tr>
<td>3 CBLB</td>
<td>3</td>
<td>105438891</td>
<td>3.8</td>
<td>4.1</td>
<td>+</td>
</tr>
</tbody>
</table>

We show the highly ranked genes (meta P, 2.0 \times 10^{-4}) in the first meta-analysis in this table.

Protein names: zinc finger protein 628, ZNF628; apolipoprotein E, APOE; translocase of outer mitochondrial membrane 40, TOMM40; matrix metallopeptidase 1, MMP1; nicotinate phosphoribosyltransferase domain containing 1, NAPRT1; triggering receptor expressed on myeloid cells 2, TREM2; Cbl proto-oncogene B, E3 ubiquitin protein ligase, CBLB.

The forest plot was drawn using ‘rmeta’ R package.

APOE is the strongest risk gene among the known risk genes for AD. In several genome-wide association studies for AD [3], the top ranked genes could show false associations with AD, because they are within same LD block of APOE e4. In addition, the pathogenesis of AD patients might be different between carriers and noncarriers of APOE e4 [36]. Therefore, we examined the dependency on APOE e4 genotype status by two ways. First, the results were compared after adjustment for APOE e4 genotype status – the number of APOE e4 allele in each individual. eMERGE and the Framingham study did not include data on APOE e4 genotype status. Therefore, we used imputed dosages of APOE e4 for these two studies (Table S1 in File S1). Second, the collinearity between selected genes and APOE e4 genotype status was examined.

Gene-based rare allele analysis

In this study, gene-based rare allele analysis means accumulations of rare alleles within the same coding region implemented in GRANVIL [27]. The definition of gene boundaries was based on the UCSC genome browser (build 37). The Framingham study showed inflated type I error and skewed results (Figures S1 and S2 in File S1). Therefore, we need to adjust for genetic stratification of the Framingham study using another algorithm implemented in GenABEL v1.69 and ProbABEL v0.30 [30,37] (Figure S2 in File S1). For gene-based analysis of the Framingham study, we need to make computer program for ourselves. We made a dosage of a gene (D) similar to an allele’s dosage in the Framingham study, as follows [27].

\[
D = \frac{\sum_{i=1}^{n} G_i}{n}
\]

Where \( G_i \) is a dosage of the \( i \)th SNP and \( n \) is a number of rare alleles within a gene that were used in the analysis.

Analyses proceeded in two steps. The overall study scheme is shown in Figure 1. We performed the first meta-analysis to select genes with genome-wide significance. The genome-wide significance was defined as significance of \( P<1.8 \times 10^{-6} \) (0.05/number of genes in human genome in UCSC genome browser (build 37) = 0.05/28517). However, there are three shortcomings in the gene-based rare allele analysis using imputation. First, it is difficult to interpret if there are a lot of rare alleles in a gene. Second, by pooling risk and protective alleles, power can be decreased. However, considering such directions before selecting candidate genes, overinflation of type I error can be problematic. Third the accuracy of imputation can be decreased in rare alleles with very low MAF. We performed confirmatory analysis (the second meta-analysis) with selected SNPs. We did confirmatory analysis, according to two reasons. First, if we could test genetic risk factors with a small number of SNPs, it would be more convenient for genotyping and interpretation. Therefore, we selected several risk

missing years of education to a mean value. The years of education was regarded as a continuous variable.
SNPs in the finally selected gene according to meta $P$ and meta $Z$ ($P<0.05$ and $Z>0$) after performing classical SNP based GWAS and meta-analysis. Second, we excluded rare variants with MAF $<0.5\%$, because the imputation accuracy decreases in very low MAF [20].

Results

The first meta-analysis

In the meta-analysis, $ZNF628$ had genome-wide significance ($meta \ P=5.3 \times 10^{-7}$ [OR 1.5, 95% CI 1.3–1.8]) (Table 2 and Figure 2A). SNPs in $ZNF628$ used in this study are summarized in Table S2 in File S1. In addition, $APOE$ had also genome-wide significance ($meta \ P=1.4 \times 10^{-6}$). Other genes with high significances, but not with genome-wide significance were $TOMM40$, $MMP1$, $NAPRT1$, $TREM2$, and $CBLB$.

Dependency on $APOE$ $4$ genotype status

We examined the dependencies of the selected genes by adjusting for $APOE$ $e4$ (Table 2). The significance of $ZNF628$ was remained, even after adjustment. However, the significance of $APOE$ decreased after adjustment for $APOE$ $e4$ (after adjustment, $P$ value of $APOE$ increased to 0.025).

Additionally, the collinearity between $ZNF628$ and $APOE$ $e4$ genotype status were examined based on the variance inflation factor (VIF, Table S3 in File S1). The VIFs of all studies were approximately 1.

Meta-analysis with selected risk SNPs (the confirmatory second analysis)

For a more applicable clinical approach, we identified significant risk SNPs by meta $P$ and meta $Z$ scores. Furthermore, considering the imputation accuracy [20], we selected SNPs with $0.5\% \leq MAF \leq 3\%$. Two risk SNPs (dbSNP ID: rs112407198 and
rs112407198 (position: 19:55995401) and rs73057174 (position: 19:55995710) reached the criteria of genome-wide significance level (Table 4).

### Discussion

We performed meta-analysis with publicly available genetic studies of AD with imputed rare (MAF < 0.01%) alleles. ZNF628 was identified to have significant association with AD. Additionally, our rare allele analysis revealed the significant association of APOE and TREM2 with AD, which suggested that our results were valid and that these genes require further study [39,40].

ZNF628 is a C2H2-zinc finger protein, a type of transcription factors [41] consisting of three exons. C2H2-type zinc finger proteins are known to be essential for normal growth and development [41]. ZNF628 is found in mammals, but not Zebra fish or C. elegans [41]. ZNF628 is evenly expressed in various tissues including brain [42,43]. ZNF628 is conserved among mammals and seems to be functionally important [41]. The possible DNA binding site is the sequence motif – C/GA/TA/AGCTTGGTTGGTTGC [41]. As this time, the target proteins and related human disorders associated with ZNF628 have not been reported. It is possible that the rare alleles in ZNF628 change the expression levels of certain proteins related to AD pathogenesis.

In the selected allele analysis of ZNF628 (the second confirmatory analysis), P and Z values of two SNPs (rs112407198 and rs73057174) reached the criteria of P<0.05 and Z>0. These SNPs are located outside the C2H2-type zinc finger domains and synonymous SNPs (Figure 3). The synonymous mutations are known to change the protein expression level and conformation [44] by affecting mRNA structure [45] or changing the time of cotranslational folding [46]. The altered expression levels or structure of ZNF628 could affect the expression level of other proteins.

There were no dependencies between ZNF628 and APOE ε4 genotype status. ZNF628 is separated from APOE by more than 10^8 bp, although they are both located on chromosome 19. Therefore, ZNF628 is not included in same LD block with APOE ε4. ZNF628 did not lose its significance in meta-analysis even after adjustment for APOE ε4 genotype status. Therefore, ZNF628 appears to be related with AD independently from APOE ε4. In contrast, the significance of APOE was affected by APOE ε4. The association of the rare alleles in APOE with AD was highly significant (P = 1.4×10^{-6}) with AD, although this significance disappeared after adjusting for APOE ε4. This suggested that rare alleles in the same LD block with APOE ε4 conferred significant association with AD.

Other risk genes that have been found in GWAS targeting common alleles were not replicated in our gene-based rare allele...
Table 4. Results of gene-based rare allele analysis top ranking genes in the AlzGene database.

<table>
<thead>
<tr>
<th>Gene</th>
<th>CHR</th>
<th>start</th>
<th>Meta Z</th>
<th>Meta P</th>
<th>Directions*</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCA7</td>
<td>19</td>
<td>1040101</td>
<td>0.2</td>
<td>0.8539</td>
<td>++++</td>
</tr>
<tr>
<td>PICALM</td>
<td>11</td>
<td>8566845</td>
<td>-0.4</td>
<td>0.7068</td>
<td>-+++</td>
</tr>
<tr>
<td>CLU</td>
<td>8</td>
<td>27454450</td>
<td>-1.1</td>
<td>0.2554</td>
<td>+++</td>
</tr>
<tr>
<td>M54A6A</td>
<td>11</td>
<td>59939080</td>
<td>0.5</td>
<td>0.6377</td>
<td>+++</td>
</tr>
<tr>
<td>CD33</td>
<td>19</td>
<td>51728334</td>
<td>-0.4</td>
<td>0.7185</td>
<td>+++</td>
</tr>
<tr>
<td>BIN1</td>
<td>2</td>
<td>127805606</td>
<td>2.0</td>
<td>0.0460</td>
<td>+++++</td>
</tr>
<tr>
<td>M54A4E</td>
<td>11</td>
<td>59980567</td>
<td>0.5</td>
<td>0.6377</td>
<td>+++</td>
</tr>
<tr>
<td>CR1</td>
<td>1</td>
<td>207669472</td>
<td>0.7</td>
<td>0.5050</td>
<td>+++</td>
</tr>
<tr>
<td>CD2AP</td>
<td>6</td>
<td>47445524</td>
<td>-1.1</td>
<td>0.2593</td>
<td>+++</td>
</tr>
</tbody>
</table>

Protein names: ATP-binding cassette, sub-family A, ABCA7; phosphatidylinositol binding clathrin assembly protein, PICALM; clusterin, CLU; membrane-spanning 4-domains, subfamily A, member 6A, M54A6A; CD33 molecule, CD33; bridging integrator 1, BIN1; putative membrane-spanning 4-domains subfamily A member 4E, M54A4E; complement component (3b/4b) receptor 1, CR1; CD2-associated protein, CD2AP.

* The signs represent those of the Z score of each study. The question mark represents missing data in the study because of low INFO or high MAF. The order of the signs is ADNI, ADN2, GenADA, eMERGE, NIA, and Framingham study.

Acknowledgments

We used the supercomputing resource of the Korea Institute of Science and Technology Information (KIST). ADNI got a grant from the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering. Additionally, ADNI was supported by the following: Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Immugenetics, N.V.; IXICO Ltd.; Jansen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. ADNI clinical sites in Canada got funds from the Canadian Institutes of Health Research. The Foundation for the National Institutes of Health (www.fnih.org) facilitated private sector contributions. The science organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Disease Cooperative Study at the University of California, Rev October 16, 2012 San Diego. The Laboratory for Neuro Imaging at the University of California, Los Angeles disseminated ADNI data. ADNI was also supported by NIH grants (P30 AG010129 and K01 AG030514). The genotypic and associated phenotypic data used in the study, Multi-Site Collaborative Study for Genotype-Phenotype Associations in Alzheimer’s Disease (GenADA)” were provided by the GlaxoSmithKline, R & D Limited. Alzheimer’s Disease Patient Registry (ADPR); and Adult Changes in Thought (ACT) study was supported by a U01 from the National Institute on Aging (Eric B. Larson, PI, U01AG006781). The 3M Corporation gave a gift and it was used to expand the ACT cohort. DNA aliquots sufficient for GWAS from ADPR Probable AD cases, who had been enrolled in Genetic Differences in Alzheimer’s Cases and Controls (Walter Kukull, PI, R01 AG007584) and obtained under that grant, were made available to eMERGE without charge. Genotyping, which was performed at Johns Hopkins University, was supported by the NIH (U01HG004438); GWAS were supported through a Cooperative Agreement from the National Human Genome Research Institute, U01HG004610 (Eric B. Larson, PI). The eMERGE Administrative Coordinating Center (U01HG004605) and the National Center for Biotechnology Information (NCBI) helped phenotype harmonization and genotype data cleaning. The “Genetic Consortium for Late Onset Alzheimer’s Disease” was supported by the Division of Neuroscience, NIA. In NIA-LOAD study, Genetic Consortium for Late Onset Alzheimer’s Disease included a GWAS funded.
as part of the Division of Neuroscience, NIA. The Framingham Heart Study is performed and supported by the National Heart, Lung, and Blood Institute (NHBLI) in collaboration with Boston University (Contract No. N01-HC-25195). Funding for SHARe Affymetrix genotyping was provided by NHBLI Contract N02-HL-64278. The Framingham Dementia Mild Plus Incidence dataset was supported by NIH/NIA grants R01 AG08122 and R01 AG033193. The Framingham Dementia Moderate Plus Incidence dataset was supported by NIH/NIA grants R01 AG08122 and R01 AG033193.

However, we did not receive the commercial funds that are shown in this section. Although the data in our study can be publicly available, it was mandatory to show the funding sources for the studies. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

This manuscript was not written in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHBLI. Some of data used in preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report.

A complete listing of ADNI investigators can be found at http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_Last.pdf.

Author Contributions

Conceived and designed the experiments: JHK SAP. Performed the experiments: JHK SAP. Analyzed the data: JHK HSL J-HL SAP. Contributed reagents/materials/analysis tools: JHK J-HL SAP. Contributed to the writing of the manuscript: JHK PS HSL J-HL JHL SAP. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: JHK PS HSL J-HL JHL SAP.

References

25. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, et al. (2008) Rare functional variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320: 539–543.