Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer’s disease

The Alzheimer Disease Genetics Consortium (ADGC) performed a genome-wide association study of late-onset Alzheimer disease using a three-stage design consisting of a discovery stage (stage 1) and two replication stages (stages 2 and 3). Both joint analysis and meta-analysis approaches were used. We obtained genome-wide significant results at MS4A4 (rs4938933; stages 1 and 2, meta-analysis \( P (P_M) = 1.7 \times 10^{-9} \), joint analysis \( P (P) = 1.7 \times 10^{-9} \); stages 1, 2 and 3, \( P_M = 8.2 \times 10^{-12} \), CD2AP (rs9349407; stages 1, 2 and 3, \( P_M = 8.6 \times 10^{-9} \), EPHA1 (rs11767557; stages 1, 2 and 3, \( P_M = 6.0 \times 10^{-10} \) and CD33 (rs3865444; stages 1, 2 and 3, \( P_M = 1.6 \times 10^{-9} \)). We also replicated previous associations at CR1 (rs6701713; \( P_M = 4.6 \times 10^{-10} \), \( P_1 = 5.2 \times 10^{-11} \), CLU (rs1532278; \( P_M = 8.3 \times 10^{-8} \), \( P_1 = 1.9 \times 10^{-8} \), BIN1 (rs7561528; \( P_M = 4.0 \times 10^{-14} \), \( P_1 = 5.2 \times 10^{-14} \) and PICALM (rs561655; \( P_M = 7.0 \times 10^{-11} \), \( P_1 = 1.0 \times 10^{-10} \)), but not at EXOC3L2, to late-onset Alzheimer’s disease susceptibility.

Alzheimer’s disease is a neurodegenerative disorder affecting more than 13% of individuals aged 65 years and older and 30–50% of individuals aged 80 years and older. Early work identified mutations in APP, PSEN1 and PSEN2 that cause early-onset autosomal dominant Alzheimer’s disease and variants in APOE that affect late-onset Alzheimer’s disease (LOAD) susceptibility. Recent genome-wide association studies (GWAS) identified variants in CR1, CLU, PICALM and BIN1 as LOAD susceptibility loci. However, because LOAD heritability estimates are high (\( h^2 = 60–80\% \)), much of the genetic contribution to this condition remains unknown.

To identify genetic variants associated with risk for Alzheimer’s disease, the ADGC assembled a discovery dataset (stage 1, 8,309 individuals with LOAD cases) and 7,366 cognitively normal elders (CNEs) as controls) using data from eight cohorts and a ninth newly assembled cohort from the National Institute on Aging (NIA)-funded Alzheimer Disease Centers (ADCs) (Supplementary Tables 1.2 and Supplementary Note), with data coordinated by the National Alzheimer Coordinating Center (NACC) and samples coordinated by the National Cell Repository for Alzheimer Disease (NCRAD). For the stage 2 replication, we used four additional datasets and additional samples from the ADCs (3,531 LOAD cases and 3,565 CNEs). The stage 3 replication used the results of association analyses provided by three other consortia, including 6,992 LOAD cases and 24,666 mixed-age controls, reported in a companion manuscript. For stages 1 and 2, we used both a meta-analysis approach that integrated results from the association analyses of individual datasets and a joint analysis approach in which genotype data from each study were pooled. The latter method has improved power over the meta-analysis in the absence of between-study heterogeneity and has a more direct correction for confounding sampling bias. We were limited to meta-analysis for stage 3 analyses.

Because the cohorts were genotyped using different platforms, we used imputation to generate a common set of 2,324,889 SNPs. We applied uniform stringent quality control measures to all data sets to remove low-quality and redundant samples and problematic SNPs (Supplementary Tables 3.4 and Online Methods). We performed an association analysis assuming an additive model on the log odds ratio scale with adjustment for population substructure using logistic regression for case-control data and generalized estimating equations (GEE) with a logistic model for family data. We combined results from individual datasets in the meta-analysis using the inverse variance method, applying a genomic control to each dataset. We performed the joint analysis using GEE and incorporated terms to adjust for population substructure and site-specific effects (Online Methods). For both approaches, we also examined an extended model of covariate adjustment that adjusted for age (age at onset or death in cases and age at exam or death in controls), sex and number of APOE4 alleles (0, 1 or 2). Genomic inflation factors (\( \lambda \)) for both the discovery meta-analysis and the joint analysis and extended models were less than 1.05, indicating that there was not substantial inflation of the test statistics (Supplementary Table 3 and Supplementary Fig. 1). Association findings from the meta-analysis and joint analysis were comparable.

In stage 1, the strongest signal was from the APOE region (rs4420638; \( P_M = 1.1 \times 10^{-266} \), \( P_1 = 1.3 \times 10^{-253} \)) (Supplementary Table 5). Excluding the APOE region, SNPs at nine distinct loci yielded \( P_M \) or \( P_1 \leq 10^{-6} \) (Table 1; all SNPs with \( P < 10^{-8} \) are shown in Supplementary Table 5). SNPs from these nine loci were carried forward to stage 2. Five of these loci had not previously been associated with LOAD at a genome-wide significance level of \( P \leq 5.0 \times 10^{-8} \) (loci in MS4A4, EPHA1, CD33, ARID5B and CD2AP). Because the companion study identified SNPs at ABCA7 to be within a new LOAD locus, we included ABCA7-region SNPs in our stage 2 analysis and provided our results to researchers from that study. For all loci listed in Table 1, we did not detect evidence for effect heterogeneity (Supplementary Fig. 2). One newly associated locus (in MS4A4) was significant in the stage 1+2 analysis. Four other loci approached but
### Table 1 Genome-wide association results for LOAD in the ADGC stage 1 and stage 2 datasets

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chr.</th>
<th>Nearest gene</th>
<th>MA</th>
<th>MAF</th>
<th>fMAF</th>
<th>Combined analysis (stage 1+2)</th>
<th>ADGC replication (stage 2)</th>
<th>Combined analysis (stage 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs670173</td>
<td>2</td>
<td>BIN1</td>
<td>0.20</td>
<td>0.36</td>
<td>0.35</td>
<td>1.04–1.24</td>
<td></td>
<td>1.08–1.22</td>
</tr>
<tr>
<td>rs7561528</td>
<td>2</td>
<td>BIN7</td>
<td>0.17</td>
<td>0.45</td>
<td>0.47</td>
<td>1.02–1.21</td>
<td></td>
<td>1.04–1.21</td>
</tr>
<tr>
<td>rs4859407</td>
<td>6</td>
<td>CD33</td>
<td>2.6 × 10^-10</td>
<td>0.01</td>
<td>0.11</td>
<td>1.04–1.27</td>
<td></td>
<td>1.09–1.23</td>
</tr>
<tr>
<td>rs9349407</td>
<td>1</td>
<td>BIN1</td>
<td>1.04–1.24</td>
<td>0.01</td>
<td>0.11</td>
<td>1.04–1.27</td>
<td></td>
<td>1.08–1.22</td>
</tr>
</tbody>
</table>

SNPs in the CD2AP locus also met our stage 1 criteria for additional analysis (Fig. 1b). Stage 2 data modestly strengthened this association, but the results did not reach genome-wide significance. The stage 3 analysis yielded a genome-wide significant result for rs9349407 (P = 6.8 × 10^-10), which identified CD2AP as a new LOAD locus. The minor allele (MAF = 0.27) at this SNP increased the risk for LOAD (OR = 1.11, 95% CI 1.07–1.15) (Table 2 and Fig. 1b).

Another locus studied further in stages 2 and 3 centered on EPHA1. Previous work provided suggestive evidence that this is a LOAD risk locus, although the associations did not previously reach genome-wide significance (P = 1.7 × 10^-8). Here, results from stages 1 and 2 for rs1765755 located in the promoter region of EPHA1, reached genome-wide significance in the joint analysis. The addition of stage 3 results increased the evidence for association (P = 6.0 × 10^-10; Table 2 and Fig. 1c). The minor allele (MAF = 0.19) for this SNP is protective (OR = 0.90, 95% CI 0.86–0.93). We observed no evidence for heterogeneity at this locus (Supplementary Fig. 2d; heterogeneity P = 0.58).

In stages 1 and 2, we also obtained strong evidence for association for SNPs in CD33, a gene located approximately 6 Mb from APOE, but our results did not reach genome-wide significance. The addition of stage 3 data confirmed that CD33 is a LOAD risk locus (rs3865444; stages 1, 2 and 3, P = 1.6 × 10^-10). The minor allele (MAF = 0.30) for this SNP is protective (OR = 0.91, 95% CI 0.88–0.93; Tables 1 and 1d). A single SNP (rs3826656) in the 5 region of CD33 was previously reported as a genome-wide significant Alzheimer’s disease-related locus using a family-based approach (P = 6.6 × 10^-9) (ref. 15). We were unable to replicate this finding (P = 0.73, P = 0.39 in the stage 1 analysis for rs3826656). Though rs3826656 is only 1,348 bp from our top SNP (rs3865444), these two sites have only weak linkage disequilibrium (LD) (r^2 = 0.13).

Researchers in the accompanying study reported highly significant evidence for the association of an ABCA7 SNP, rs3764630, with LOAD (P = 4.5 × 10^-11), from a meta-analysis that included data from our study. In our stage 1+2 analysis, we obtained suggestive evidence for association with the ABCA7 SNP rs3752246 (P = 5.8 × 10^-7, P = 5.0 × 10^-7), which is a missense variant (p.Gly1527Ala) that may alter the function of the ABCA7 protein (see Supplementary Table 6 for functional SNPs in LD with SNPs yielding P > 10^-6).
Table 2 | Meta-analysis of stage 1+2 with stage 3 (CHARGE/GERAD/EDAD1 Consor{}tia) GWAS results

<table>
<thead>
<tr>
<th>Gene-SNP</th>
<th>Cases</th>
<th>Controls</th>
<th>Total</th>
<th>OR_{M} (95% CI)</th>
<th>P_{M}</th>
<th>OR_{J} (95% CI)</th>
<th>P_{J}</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD2AP: rs9349407</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADGC</td>
<td>11,840</td>
<td>10,931</td>
<td>22,771</td>
<td>1.12 (1.07–1.18)</td>
<td>1.0 × 10^{-6}</td>
<td>1.12 (1.07–1.17)</td>
<td>2.1 × 10^{-6}</td>
</tr>
<tr>
<td>External</td>
<td>6,922</td>
<td>18,896</td>
<td>25,818</td>
<td>1.09 (1.03–1.15)</td>
<td>0.002</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>ADGC + External</td>
<td>18,762</td>
<td>29,827</td>
<td>48,589</td>
<td>1.11 (1.07–1.15)</td>
<td>8.6 × 10^{-9}</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>EPHA1: rs11767557</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADGC</td>
<td>11,840</td>
<td>10,931</td>
<td>22,771</td>
<td>0.87 (0.83–0.92)</td>
<td>2.4 × 10^{-7}</td>
<td>0.87 (0.83–0.91)</td>
<td>4.9 × 10^{-8}</td>
</tr>
<tr>
<td>External</td>
<td>6,922</td>
<td>24,666</td>
<td>31,588</td>
<td>0.91 (0.87–0.96)</td>
<td>2.9 × 10^{-4}</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>ADGC + External</td>
<td>18,762</td>
<td>35,597</td>
<td>54,359</td>
<td>0.90 (0.86–0.93)</td>
<td>6.0 × 10^{-10}</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>ARID5B: rs2588969</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADGC</td>
<td>11,840</td>
<td>10,931</td>
<td>22,771</td>
<td>0.89 (0.86–0.93)</td>
<td>1.1 × 10^{-7}</td>
<td>0.89 (0.86–0.93)</td>
<td>2.0 × 10^{-7}</td>
</tr>
<tr>
<td>External</td>
<td>6,922</td>
<td>18,896</td>
<td>25,818</td>
<td>0.92 (0.88–0.97)</td>
<td>5.4 × 10^{-4}</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>ADGC + External</td>
<td>18,762</td>
<td>29,827</td>
<td>48,589</td>
<td>0.99 (0.95–1.02)</td>
<td>0.362</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>MS4A4A: rs4938933</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADGC</td>
<td>11,840</td>
<td>10,931</td>
<td>22,771</td>
<td>0.88 (0.85–0.92)</td>
<td>1.7 × 10^{-9}</td>
<td>0.88 (0.85–0.92)</td>
<td>1.7 × 10^{-9}</td>
</tr>
<tr>
<td>External</td>
<td>6,922</td>
<td>18,896</td>
<td>25,818</td>
<td>0.92 (0.88–0.97)</td>
<td>5.4 × 10^{-4}</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>ADGC + External</td>
<td>18,762</td>
<td>29,827</td>
<td>48,589</td>
<td>0.89 (0.87–0.92)</td>
<td>8.2 × 10^{-12}</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>CD33: rs3865444</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADGC</td>
<td>11,840</td>
<td>10,931</td>
<td>22,771</td>
<td>0.86 (0.83–0.91)</td>
<td>1.1 × 10^{-7}</td>
<td>0.86 (0.83–0.91)</td>
<td>2.0 × 10^{-7}</td>
</tr>
<tr>
<td>External</td>
<td>6,922</td>
<td>18,896</td>
<td>25,818</td>
<td>0.92 (0.88–0.97)</td>
<td>5.4 × 10^{-4}</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>ADGC + External</td>
<td>18,762</td>
<td>29,827</td>
<td>48,589</td>
<td>0.91 (0.88–0.93)</td>
<td>1.6 × 10^{-9}</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

We also examined SNPs with statistically significant GWAS results reported by others (GAB2 (ref. 16), PCDH11X, GOLM1 (ref. 18) and MTHFD1L; Supplementary Table 7). Stage 1 data were used, except for PCDH11X, for which stage 1+2 data were used because Affymetrix platforms do not contain the appropriate SNP. Only SNPs in the ADGC dataset and the external replication (stages 1, 2 and 3) datasets (highlighted with a red diamond). Computed estimates of linkage disequilibrium (r^2) with the most significant SNP at each locus are indicated at the bottom of each panel. The length and the direction of the arrows represent the scaled size and the direction of the genes, respectively.

Figure 1 | Regional association plots from the three-stage meta-analysis with LOAD. P_{M} values for association are shown for (a) the MS4A gene cluster, (b) CD2AP, (c) EPHA1 and (d) CD33. For each locus, the genomic position (NCBI Build 37.1) is plotted on the x axis against –log_{10} P on the y axis. For the SNP with the lowest P value at each locus in the stage 1 analyses, three P values for association are shown: P_{J} meta-analysis of the ADGC discovery (stage 1) dataset (highlighted with a black diamond), P_{1+2} meta-analysis of the combined ADGC discovery and replication (stages 1 and 2) datasets (highlighted with a blue diamond) and P_{1+2+3} meta-analysis of the combined ADGC dataset and the external replication (stages 1, 2 and 3) datasets (highlighted with a red diamond). Computed estimates of linkage disequilibrium (r^2) with the most significant SNP at each locus are shown as an orange diamond for r^2 ≥ 0.8, a yellow diamond for 0.5 < r^2 < 0.8, a grey diamond for 0.2 ≤ r^2 < 0.5 and a white diamond for r^2 < 0.2. Genes in each region are indicated at the bottom of each panel. The length and the direction of the arrows represent the scaled size and the direction of the genes, respectively.

© 2011 Nature America, Inc. All rights reserved.
dataset used here. However, using the stage 1+2 data, we did not replicate this finding or see evidence of APOE genotype interactions with the loci listed in Table 1 (data not shown).

Previous work reported an association between LOAD and the chromosome 19 SNP rs597668, which is located 7.2 kb proximal to EXOC3L2 and 296 kb distal of APOE2. Although we did observe a signal for this SNP (stage 1, PM = 1.5 × 10−9, P = 7.7 × 10−10) and other SNPs in the EXOC3L3-MARK4 region, the evidence was completely extinguished for all SNPs after adjustment for APOE (Online Methods and Supplementary Table 8), suggesting that signal in this region is from APOE.

Our observation of genome-wide significant associations at MS4A4A, CD2AP, EPFAH1 and CD33 extends our understanding of the genetic architecture of LOAD and confirms the emerging consensus that common genetic variation plays an important role in the etiology of LOAD. With our findings and those in the companion study12, there are now ten LOAD susceptibility loci (in APOE, CR1, CLU, PICALM, BIN1, EPFAH1, MS4A4A, CD33, CD2AP and ABCA7). Examining the amount of genetic effect attributable to these candidate genes, the most strongly associated SNPs at each locus other than that in APOE had population attributable fractions between 2.72% and 5.97% (Supplementary Table 9), with a cumulative population-attributable fraction for non-APOE loci estimated to be as much as 35%; however, these estimates may vary widely between studies12, and the actual effect sizes are likely to be much smaller than those estimated here because of the ‘winner’s curse’. Also, the results do not account for interaction among loci and are not derived from appropriate population-based samples.

A recent review of GWAS22 noted that risk alleles with small effect sizes (0.80 < OR < 1.2) likely exist for complex diseases such as LOAD but remain undetected, even with thousands of samples, because of insufficient power24. Our discovery dataset (stage 1, 8,309 cases and 7,366 controls) was well powered to detect associations exceeding the genome-wide significance threshold of P < 10−8 (Supplementary Table 9). If there are many loci of more modest effects, some, but not all, will likely be detected in any one study. This likely explains the genome-wide significance for the ABCA7 locus in the accompanying manuscript12, which reached only modest statistical significance in our dataset (rs3752246; PM = 1.0 × 10−3, P = 1.9 × 10−3). Finding additional LOAD loci will require larger studies with increased depth of genotyping to test for the effects of both common and rare variants.

ACKNOWLEDGMENTS
The National Institutes of Health, National Institute on Aging (NIA-NIA) supported this work through the following grants: ADGC, U01 AG032984, R22 AG030632, NACC; U01 AG016996; NCRAD, U24 AG18186; NIA LOAD, U24 AG26395; U24 AG26390; Boston University, P30 AG13846, R01 HG02213, K24 AG027841, U01 AG10483, R01 CA129769, R01 MH082925, R01 AG09092, R01 AG17137, R01 AG25299; Columbia University, P50 AG06870, R37 AG051473; Duke University, P30 AG028737; Emory University, AG025688; Indiana University, P30 AG10133; Johns Hopkins University, P50 AG05146, R01 AG20688; Massachusetts General Hospital, P50 AG05134; Mount Sinai School of Medicine, P50 AG011828, R01 AG02219; New York University, P30 AG08051, MO1RR00996, and UL1 RR02993; Northwestern University, P30 AG14843; Oregon Health & Science University, P30 AG08017, R01 AG26916; Rush University, P30 AG01061, R01 AG19085, R01 AG15819, R01 AG17917, R01 AG30146; University of Alabama at Birmingham, P50 AG01682, U1LR02777; University of Arizona/TGEN, P30 AG019610, R01 AG031581, R01 NS059873; University of California, Davis, P30 P5010129; University of California, Irvine, P50 AG016573, P50, P50 AG016575, P50 AG016576, P50 AG016577; University of California, Los Angeles, P50 AG016570; University of California, San Diego, P50 AG05131; University of California, San Francisco, P50 AG023501, P01 AG016724; University of Kentucky, P30 AG02833; University of Michigan, P50 AG08671; University of Pennsylvania, P30 AG010124; University of Pittsburgh, P50 AG05133, AG05146; University of Southern California, P50 AG05142; University of Texas Southwestern, P30 AG012300; University of Miami, R01 AG27944, P5010499, P01 AG27944, AG021547, AG019757; University of Washington, P50 AG05136, U01 AG066781, U01 AG049610; Vanderbilt University, P01 AG19085; and Washington University, P50 AG005681, P01 AG03991. ADNI Funding for ADNI is through the Northern California Institute for Research and Education by grants from Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Eli Lilly Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innoventics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc., E. Hoffmann-La Roche, Schering-Plough, Synarc, Inc., Alzheimer’s Association, Alzheimer’s Drug Discovery Foundation, the Dana Foundation, and by the National Institute of Biomedical Imaging and Bioengineering and NIA grants U01 AG024944, R01 AG036555, R50 AG05314. We thank C. Phelps, M. Morrison-Bograd and M. Miller from NIA who are ex-officio ADGC members. Support was also from the Alzheimer’s Association (LAIF, IRRG-08-8972; MP-V, IRRG-05-11417) and the Veterans Affairs Administration. P.-H. is supported by Wellcome Trust, Howard Hughes Medical Institute, and the Canadian Institute of Health Research.

AUTHOR CONTRIBUTIONS


These authors contributed equally to this work. Correspondence should be addressed to G.D.S. (gerardsc@mail.med.upenn.edu).
ONLINE METHODS

Subjects. A full description of study cohorts is provided in the Supplementary Note and Supplementary Tables 1 and 2.

Covariate data. Age of onset data was available from some cohorts (Alzheimer’s Disease Center (ADC), Translational Genomics Research Institute series 2 (TGEN2), National Institute on Aging Late-onset AD (NIA-LOAD), Multi-Institutional Research on Alzheimer’s Genetic Epidemiology (MIRAGE), Adult Changes in Thought (ACT), Multi-Site Collaborative Study for Genotype-Phenotype Associations in Alzheimer’s Disease (GenADA), University of Pittsburgh (UP) and the Rush University Religious Orders Study/Memory and Aging Project (ROS/MAP), whereas for others, only age at ascertainment (Washington University (WU) and ADNI), age at diagnosis (Mayo Clinic (MAYO)), or a combination of both age at ascertainment and age at death was available (a subset of autopsy-confirmed samples in the University of Miami/Vanderbilt University/ Mt. Sinai School of Medicine (UM/VU/MSSM) cohort). For subjects with autopsy-confirmed diagnosis and no clinical diagnosis, the age at diagnosis was equated to the age at death. For all studies, the age used for CNEs was the age of last exam or age at death. Case and CNE subjects with age at symptom onset or age at death less than 60 were excluded from the analysis. We restricted our association analyses to individuals of European ancestry because there were insufficient subjects from non-European ancestry groups to obtain meaningful results.

Genotyping, data cleaning and imputation. Genotypes were from either Illumina or Affymetrix high-density SNP microarrays (Supplementary Table 3). Genotype data were cleaned by applying minimum call rates (95% and 98%) and minimum minor allele frequencies (0.02 and 0.01) for cohorts genotyped on Affymetrix and Illumina chips, respectively. SNPs not in Hardy-Weinberg equilibrium (P < 10^-6) were excluded. Subjects where the gender was misspecified were identified by analysis of X-chromosome SNPs using PLINK. For cohorts genotyped on multiple chips (MIRAGE and UM/VU/MSSM), genotype and sample quality thresholds were applied within subsets of individuals genotyped on each chip. For all other cohorts, quality thresholds were applied per cohort. Relationships among individuals in the family-based cohorts (MIRAGE and NIALOAD) were confirmed by pairwise genome-wide estimates of proportion identity-by-descent (IBD) using PREST software. Any discrepancies identified were reviewed in light of available clinical and pedigree data to determine the most likely relationship consistent with a proportion of IBD, and any remaining scenarios were excluded from analysis. Latent relatedness in the case-control cohorts was identified by proportion IBD using PLINK software. Both of each pair of identical samples by IBD (π ≥ 0.99) were dropped, and one subject was selected from each related pair (0.4 < π < 0.90) prioritizing non-missing case or non-missing control status and then higher call rate in selection. Duplicate enrollments among studies (Supplementary Table 4) were identified using proportion of IBD in a genotyped dataset including all cohorts where pairs with π > 0.95 were considered duplicate enrollments. Duplicates with discordant case or control status by study were dropped from both studies, and those with concordant status were included in only one cohort and selected according to a predetermined priority list of cohorts which considered genotype data, phenotype data and the type of cohort. Genome-wide imputation was performed per cohort using MACH software with HapMap phase 2 (release 22) CEPH Utah pedigree (CEU) reference haplotypes and genotype data passing quality control as inference. Imputation quality was determined as R^2 and only SNPs imputed with R^2 ≥ 0.50 were included in the analysis.

APOE genotyping. APOE genotypes were determined for the ADC, ACT, NIALOAD, UM/VU/MSSM, MAYO and GenADA cohorts using SNPs rs442241 and rs429358; for the MIRAGE cohort using the Roche Diagnostics LightCycler 480 instrument (Roche Diagnostics) LightCycler Kit APOE C1/C2R158 (TIB MOLBIOL); for TGEN2, ADNI, UP and WU cohorts by pyrosequencing or restriction fragment length polymorphism analysis and for ROSMAP by high-throughput sequencing of codons 112 and 158 in APOE by Agencourt Bioscience Corporation.

Meta-analysis. Presence of intra-study population substructure was evaluated separately by cohort in a two-step process that first removed outliers before estimating population substructure within the remaining population. For the first step, either the STRUCTURE software package [39] (UM/VU/MSSM and MIRAGE) or the ‘smartpca’ script in EIGENSTRAT (remaining cohorts) was used to remove outliers and/or confirm self-reported ethnicity after filtering to remove SNPs in pairwise LD. In the second step, we used EIGENSTRAT, often a second time, to estimate principal component loadings for inclusion in association analysis. For each study, the first two, three or four estimated principal components were identified for inclusion as covariates in association analysis (Supplementary Table 3). Outlier detection for the ADC, TGEN2, GenADA, ACT, ADNI, ROS/MAP, OHSU, UP, WU and MAYO cohorts was evaluated by comparison to the HapMap 3 CEPH (CEU) population. EIGENSTRAT analyses of family cohort data (NIA-LOAD and MIRAGE) used a sample of unrelated individuals to fit principal components after outliers with respect to European-American ancestry were removed.

Genotyped and imputed SNP data passing quality control were tested for association with Alzheimer’s disease in each dataset using logistic generalized linear model (GLM) for case-control analysis and logistic generalized estimating equations (GEE) for family-based cohorts in R0-34. All analyses assumed an additive genetic model, coding genotyped SNPs by the number of minor alleles (0, 1 or 2) and imputed SNPs by the posterior probability of the minor allele (range 0–2). Primary association analyses were adjusted for population substructure (baseline model).

SNP association results for each dataset were meta-analyzed using the inverse variance method implemented in the software package METAL. The meta-analysis P value was estimated by the summarized test statistic after applying a genomic control within each individual study. Heterogeneity among odds ratios in the meta-analysis was assessed using the Cochrane’s Q and P statistics. Regional association plots were prepared for the most strongly associated SNPs in CR1, BIN1, CD2AP, EPHA1, CLU, MS4A4A/MS4A6A, PICALM, ABCA7 and CD33 using the gene locations from UCSC Genome browser (hg19, GRCh37, Feb 2009 release) and SNP locations from the corresponding dbSNP build 131. Estimates of LD were calculated with the FUGUE software using HapMap phase 2 (release 24, CEU) genotype data and build 131 SNP positions. Forest plots of study-specific effects and analysis results are presented for the same set of SNPs using the ‘meta’ package in R.

Joint analysis. Testing for population substructure across studies was performed in a combined dataset using the set of SNPs genotyped in all study cohorts. After filtering SNPs with pairwise LD (r^2 < 0.20, 31,310 SNPs were evaluated using EIGENSTRAT. The top three principal components from EIGENSTRAT were used as covariates in the joint analysis for association in addition to an adjustment for site-specific effects using dummy variables for each cohort. SNP associations with Alzheimer’s disease affection status were examined in a pooled analysis of subjects from all cohorts, excluding SNPs missing from one or more individual dataset or with genotypes available on fewer than 98% of individuals overall. In total, 2,312,972 directly genotyped or imputed SNPs common to all datasets were tested for association in 8,509 cases and 7,366 CNEs, including 3,489 individuals in family datasets using GEE analyses in R. Joint analyses of the baseline model, full model and models evaluating robustness to APOE included as covariates the principal components from inter-study and intra-study population substructure and a dummy covariate for cohort-specific effects. Genomic inflation factors for the discovery joint analysis in the basic and extended models of covariate adjustment were 1.05 and 1.04, respectively (Supplementary Table 3), which were similar to those from meta-analysis.

Secondary analysis. Association results in regions yielding at least one SNP with P < 10^-8 (follow-up SNPs) were further evaluated for robustness to APOE e4 carrier status in analyses stratified according to presence or absence of APOE e4 and an interaction analysis including effects for SNP, APOE e4 and their interaction. In addition, we examined the EXOC3L2 region in chromosomes 19 previously reported as independent of APOE genotype in a full model including covariates for age at onset or age at last exam, gender and the dosage of APOE e4 alleles.
Internal and external replication analyses. SNPs attaining a $P \leq 1 \times 10^{-6}$ for association with LOAD in the discovery cohort were evaluated in five independent datasets (ADC3, OHSU, MAYO, ROS/MAP and UP) consisting of 3,531 cases and 3,565 CNEs using the same analytical approaches as described above. Replication was performed using both meta-analysis and joint analysis. The datasets included in discovery and replication analyses are summarized in Supplementary Tables 1 and 2. Following internal replication, an external replication cohort was sought to evaluate the most strongly associated SNP in each of four newly identified genes (CD2AP, EPHA1 (rs11767557), ARID5B (rs2588969) and CD33 (rs3865444)) for which results did not meet genome-wide significance ($P_{	ext{GWA}} > 5 \times 10^{-8}$, $P_1 > 5 \times 10^{-8}$) in the combined discovery and replication datasets (stage 1+2). We obtained summarized results from five independent external datasets generously provided by the Genetic and Environmental Risk in Alzheimer’s Disease (GERAD) Consortium, the European Alzheimer’s Disease Initiative (EADI) Consortium and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, and these were used for combined stage 1, 2 and 3 meta analysis. After removing subjects recognized as part of the ADGC cohorts, the sample included 6,922 Alzheimer’s disease cases and 24,666 controls. These datasets were analyzed using meta-analysis as described above for the stage 1 and 2 datasets. Results from stages 1, 2 and 3 were likewise assessed by meta-analysis as described above.