Alzheimer’s disease: analyzing the missing heritability

Perry G. Ridgea,b, Shubhabrata Mukherjeee, Paul Cranec, John S.K. Kauwea*, Alzheimer’s Disease Genetics Consortium**

aDepartment of Biology, Brigham Young University, Provo, UT, USA, 84602
bARUP Institute for Clinical and Experimental Pathology, Salt Lake City, UT, USA, 84108
cDepartment of Medicine, University of Washington, Seattle, WA, USA, 98122

*To whom correspondence should be addressed
**Coauthors from the Alzheimer’s Disease Genetics Consortium are listed below

Perry G. Ridge: perry.ridge@byu.edu
Shubhabrata Mukherjee: smukherj@u.washington.edu
Paul Crane: pcrane@uw.edu
John S.K. Kauwe: kauwe@byu.edu, 801-422-2993, 801-422-0090 (fax)

ADGC Authors and Affiliations:

1Department of Neurology, Johns Hopkins University, Baltimore, Maryland, 2Department of Neurology, University of Michigan, Ann Arbor, Michigan, 3Geriatric Research, Education and Clinical Center (GRECC), VA Ann Arbor Healthcare System (VAAAHS), Ann Arbor, Michigan, 4Michigan Alzheimer Disease Center, Ann Arbor, Michigan, 5Department of Neurology, University of California Los Angeles, Los Angeles, California, 6Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, 7Department of Medicine (Genetics Program), Boston University, Boston, Massachusetts, 8Department of Pharmacology and Neuroscience, University of North Texas Health Science Center, Fort Worth, Texas, 9Department of Human Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania, 10Department of Neurological Sciences, Rush University Medical Center, Chicago, Illinois, 11Department of Behavioral Sciences, Rush University Medical Center, Chicago, Illinois, 12Civin Laboratory for Neuropathology, Banner Sun Health Research Institute, Phoenix, Arizona, 13The John P. Hussman Institute for Human Genomics, University of Miami, Miami, Florida, 14Dr. John T. Macdonald Foundation Department of Human Genetics, University of Miami, Miami, Florida, 15National Alzheimer's Coordinating Center, University of Washington, Seattle, Washington, 16Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, Illinois, 17Department of Pathology,
Washington University, St. Louis, Missouri, 102Department of Biostatistics, Mayo Clinic, Rochester, Minnesota, 103Department of Anatomic Pathology, Mayo Clinic, Rochester, Minnesota, 104Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, 105Arizona Alzheimer’s Consortium, Phoenix, Arizona, 106Department of Psychiatry, University of Arizona, Phoenix, Arizona, 107Banner Alzheimer's Institute, Phoenix, Arizona, 108Alzheimer's Disease Center, New York University, New York, New York, 109Gertrude H. Sergievsky Center, Columbia University, New York, New York, 110Department of Neurology, Columbia University, New York, New York, 111Tanz Centre for Research in Neurodegenerative Disease, University of Toronto, Toronto, Ontario, 112Department of Neurology, University of Texas Southwestern, Dallas, Texas, 113Department of Radiology and Imaging Sciences, Indiana University, Indianapolis, Indiana, 114Department of Pathology (Neuropathology), Rush University Medical Center, Chicago, Illinois, 115Department of Psychiatry, University of Southern California, Los Angeles, California, 116Cambridge Institute for Medical Research and Department of Clinical Neurosciences, University of Cambridge, Cambridge, 117Department of Pathology, Johns Hopkins University, Baltimore, Maryland, 118Sanders-Brown Center on Aging, Department of Anatomy and Neurobiology, University of Kentucky, Lexington, Kentucky, 119Department of Pathology & Laboratory Medicine, University of California Los Angeles, Los Angeles, California, 120Taub Institute on Alzheimer's Disease and the Aging Brain, Department of Pathology, Columbia University, New York, New York, 121Department of Psychiatry & Behavioral Sciences, Duke University, Durham, North Carolina, 122Department of Pathology, Oregon Health & Science University, Portland, Oregon, 123Evelyn F. McKnight Brain Institute, Department of Neurology, Miller School of Medicine, University of Miami, Miami, Florida
Abstract

Alzheimer’s disease (AD) is a complex disorder influenced by environmental and genetic factors. Recent work has identified 11 AD markers in 10 loci. We used Genome-wide Complex Trait Analysis to analyze >2 million SNPs for 10,922 individuals from the Alzheimer’s Disease Genetics Consortium to assess the phenotypic variance explained first by known late-onset AD loci, and then by all SNPs in the Alzheimer’s Disease Genetics Consortium dataset. In all, 33% of total phenotypic variance is explained by all common SNPs. APOE alone explained 6% and other known markers 2%, meaning more than 25% of phenotypic variance remains unexplained by known markers, but is tagged by common SNPs included on genotyping arrays or imputed with HapMap genotypes. Novel AD markers that explain large amounts of phenotypic variance are likely to be rare and unidentifiable using genome-wide association studies. Based on our findings and the current direction of human genetics research, we suggest specific study designs for future studies to identify the remaining heritability of Alzheimer’s disease.

Key Words

Alzheimer’s disease genetics; APOE; heritability; phenotypic variance; Alzheimer’s disease markers; next-generation sequencing; rare variants; study design
1. Introduction

Alzheimer’s disease (AD) is the most common form of dementia. Worldwide estimates of prevalence vary, with estimates of 24 to 35 million people affected [1-3]. Combined with an aging population, prevalence is expected to increase to 1 in 85 people by 2050 [2].

AD is a heterogeneous disease caused by a combination of environmental and genetic factors. The most important risk factor for Alzheimer’s disease is age [1,4]. Environmental risk factors include hypertension, estrogen supplements [5], smoking [6,7], stroke, heart disease, depression, arthritis, and diabetes [8]. In addition, certain lifestyle choices appear to decrease the risk of AD: exercise [9], intellectual stimulation [10], and maintaining a Mediterranean diet (including fish)[11,12].

The genetics of AD are complex. Several genes are known to harbor either causative or risk variants for AD. There are two primary types of AD as defined by age. The first is early-onset, or familial AD, and the second type is late-onset AD (LOAD), sometimes termed sporadic AD. Three genes, \textit{APP} [13], \textit{PSEN1} [14], and \textit{PSEN2} [15] are known to harbor many highly penetrant, autosomal dominantly-inherited variants, which lead to early-onset AD but account for only a small fraction of total AD cases.

LOAD accounts for 99% of AD cases and is caused by a more complex underlying genetic architecture. Genome-wide association studies (GWAS) have identified 10 different loci associated with AD (Table 1). Recent applications of next-generation sequencing (NGS) have suggested rare variants play important role and have large effects
in the etiology of AD [16-18]. Identifying additional variants will provide information that is integral to the development, evaluation and application of effective therapeutic strategies for AD. Lee et al. [19] used 3,333 cases and 3,924 controls, including 2,699 population-based controls to estimate that common genetic variants account for 24% of variance in AD. They also estimated the contribution of APOE using several proxy SNPs, with varying degrees of LD, with the APOE ε4 allele to estimate the APOE effect at approximately 4%. Here we evaluate the variance in AD status explained by common SNPs and along with all recently identified AD genes, including direct genotyping of the APOE ε2 and ε4 alleles, in 5,708 AD cases and 5,214 clinically ascertained controls. We also suggest strategies for identifying the remaining AD genes.
2. Methods

2.1 Dataset

We used the Alzheimer’s Disease Genetics Consortium (ADGC) dataset described in Naj et al. [20] for our analyses. Samples were genotyped using Affymetrix and Illumina SNP chips. Quality control of the imputed data was performed as described by Naj et al. 2011 [20]. Briefly, markers with a minor allele frequency of less than 1% and deviation from HWE where $P<10^{-6}$ were removed. To have a common set of SNPs across all samples, imputation to HapMap phase 2 (release 22)[21] was performed using MaCH [22] and strand ambiguous SNPs were removed, resulting in a rectangular dataset with 2,042,114 SNPs. Only SNPs imputed with $R^2 \geq 0.50$ were included in the dataset. We added an additional two SNPs, rs7412 and rs429358, corresponding to $APOE \varepsilon2$ and $\varepsilon4$, respectively.

We used a compiled dataset of directly genotyped SNPs common to all 15 studies to assess cryptic relatedness and calculate principal components to account for population-specific variations in allele distribution. We excluded strand ambiguous SNPs, resulting in a rectangular dataset with 21,880 directly observed (not imputed) SNPs in common across all the studies. We filtered SNPs with pairwise LD ($r^2$) $< 0.20$, resulting in a dataset with 17,054 SNPs. We used both PLINK [23] and KING-ROBUST [24] for relatedness analysis. KING-ROBUST provided unbiased kinship coefficient estimates for related individuals in our dataset. We excluded up to 3rd degree relatives (kinship $\geq 0.0442$) for a final dataset containing 19,692 individuals.
Of the 19,692 individuals in the original dataset we analyzed a subset of 10,922 individuals who had complete data for the 11 markers listed in Table 1, AD case-control status, age, sex, and 10 principal components from the population stratification analysis (missingness rates for each of the covariates and case-control status are reported in Table S1). Basic demographic information for the 10,922 individuals in the subset of the dataset used in this study is presented in Table 2. We collected chromosome length and number of genes per chromosome from the Vega database [25].

2.2 Genetic Analyses

We used Genome-wide Complex Trait Analysis (GCTA)[26], a tool that implements the methods described in Yang et al. [27], Lee et al. [28], and Yang et al. [29] to estimate the phenotypic variance explained by known AD genes and tagged by SNPs on the SNP arrays. Briefly, GCTA uses a mixed linear model and treats the effects of SNPs as random effects, effectively testing all the SNPs together for effect (in contrast to GWAS, which considers each SNP individually). We used age, sex, and 10 principal components as covariates. For the analyses in which we examined unexplained phenotypic variance, we also controlled for the 11 known AD markers (Table 1). The 11 known AD markers are the AlzGene.org top hits and are the markers with replicable evidence for association with AD. Each of these markers is present in our dataset except rs9349407 in CD2AP. As proxy we used rs9296559, which is in very high LD with rs9349407 ($r^2=1$). We specified a population prevalence of LOAD at 0.13 [30].
2.3 Ethics Statement

All study procedures were approved by the Institutional Review Boards of Brigham Young University and the University of Washington.

3. Results

We estimated the variance in AD case-control status focusing first on the 11 known AD markers (Table 1). Together these markers account for 7.8% (standard error 0.03) of the phenotypic variance (Table 3). Next, we estimated the explained phenotypic variance for each chromosome (Figure 1). Chromosome 19 accounts for the highest proportion of phenotypic variance.

In all, the 2,042,116 SNPs in the HapMap imputed ADGC dataset explain 33.1% of phenotypic variance (genetic variance of 0.0711, standard error 0.0072). The APOE ε2 and ε4 alleles account for 5.9% (standard error 0.03) of the phenotypic variance (Table 3). The other 9 known high frequency SNPs identified in GWAS explain an additional 1.9% (standard error 0.03)(Table 3). After controlling for these 11 markers, an additional 25.3% of the total phenotypic variance (genetic variance of 0.046, standard error 0.006) is explained with as-yet unidentified variants (Table 3). The remaining phenotypic variance explained by each chromosome after controlling for the 11 known markers is shown in Figure 2. SNPs on chromosomes 1, 4, 5 and 17 account for the largest percentage of remaining unexplained phenotypic variance compared to other
chromosomes, each accounting for more than 2% (Figure 2). Chromosomes 9, 14, and 21 account for the least (<0.0001% each); however, there is unexplained phenotypic variance on all the autosomes. There is no relationship between explained variance and chromosome length (p-value = 0.8), or number of genes per chromosome (p-value = 0.7).

4. Discussion

A clear understanding of the genetic architecture of Alzheimer’s disease provides the foundation of information needed to cure this terrible disease. While many large GWAS for AD have been performed and several replicable loci have been identified (as referenced in Table 1), relatively little phenotypic variance is explained by these variants. Our estimates of phenotype variance explained by common genetic variants and by the APOE locus are higher than those of Lee et al. [19]. We estimated total phenotypic variance explained by common SNPs to be ~33%. In contrast Lee et al. [19] estimated ~24%. In our study we used genotyped and HapMap imputed SNPs, whereas Lee et al. [19] used only directly genotyped SNPs. Inclusion of imputed SNPs improved heritability estimates and suggests that imputed SNPs should be included in such studies. In addition to using imputed variants, our dataset was larger and our controls were clinically ascertained. Differences in the estimates for APOE (~6% in this study compared to ~4% in Lee et al. [19]) could be due to these same characteristics as well as the direct genotyping of the APOE ε2 and ε4 alleles in our samples as opposed to the use of proxy markers. Regardless, both studies provide evidence that a considerable amount of variance in AD is explained by yet unidentified genetic variation. Together these results
show that there is clearly much work to be done if we are to solve the genetic architecture of AD. GWAS with sample sizes performed to date are able to identify common variants with moderate to small effect sizes. Results of GWAS in AD and other conditions suggest there may be a large number of such variants and that additional variants of this type can be identified by increasing sample sizes. However, additional loci detected with the GWAS strategy will likely have effects either similar to or smaller than SNPs already identified. The range of SNPs identifiable by current GWAS [20,31-36] is marked on Figure 3 by the large box bordered by dots (the GWAS search space), with recent GWAS hits inside the labeled oval. The GWAS being conducted by the International Genomics of Alzheimer’s Project represents a substantial increase in sample size and will undoubtedly identify additional common loci with small effects on AD risk. Nevertheless, it is unlikely that many common variants of even modest effect size remain to be identified.

There are still many AD variants that remain to be identified, however, and these variants exist on every autosome (Figure 2). Variants with large effects are almost certainly present in very low frequencies or they would have been identified in GWAS. While such variants are unlikely to be detected using traditional GWAS due to limitations of $r^2$ based “tagging” for alleles with different frequencies [37] the current analysis allows for high D’ values between common alleles and rare variants of large effect to contribute to the explained variance. These rare variants of large effect appear in the smaller box bordered with dashes in Figure 3. To date, identified alleles of this type have clear functional effects and large effect sizes compared to associated alleles from GWAS. Detecting rare
variants of large effect requires different experimental designs than GWAS such as sequencing causal loci. Exome chip array studies target known variation in coding regions, even those of very low frequency; this may prove a promising and economical approach. However, accurately genotyping variants of less than 1-2% using these arrays is quite challenging, and for variants that are present below these frequencies other approaches are required.

Two seemingly contradictory hypotheses exist about the architecture of complex disease: the common disease/common variant hypothesis and the multiple rare variant hypothesis. In the first, many common variants of small effect size collectively explain disease risk, while in the second, rare variants, some with large effect and high penetrance, explain disease risk. However, as suggested by Singleton et al. [38] these two hypotheses are not mutually exclusive and the genetics of complex diseases are likely a hybrid of the two. Singleton et al. [38] suggests that both common and rare variants that increase or decrease disease risk are likely to be found in the same loci and coined the phrase “pleomorphic risk loci”. To date, AD genetics research has largely focused on common variants that influence disease risk, likely due to technological and financial constraints. However, the advent of next generation sequencing (NGS) and falling costs of this technology have made it possible to expand AD research to include searching for rare variants. Recently, this technology was used to identify a functional variant that protects against Alzheimer’s disease in the amyloid precursor protein (APP)[17]. Additionally, two groups recently used NGS to identify additional, likely functional, variants associated with AD in the triggering receptor expressed on myeloid 2 (TREM2) gene
The TREM2 variant is present in about 1% of the general population and has a high odds ratio (2.9 to 5.1 depending on the dataset). Likewise, the APP variant is extremely rare (frequency of 0.00038), but confers a large protective effect on carriers. Larger scale applications of this technology and careful study design are likely to identify additional variants and further explain the remaining phenotypic variance in AD.

Family-based studies are also an effective application of NGS. These studies require carefully ascertained families and accurate pedigree data and can be used to identify high effect, low frequency variants (located in the box with longer dashes in Figure 3). Family-based studies are especially powerful because large effect, low frequency disease-causing (or disease-modifying) sequence features, some of which may be unique to a single family, are likely to segregate, at least partially, with disease status. These approaches have not yet been extensively applied in AD research. Nevertheless, family-based studies utilizing large-scale genome or exome sequencing have recently been used to identify disease-causing variants in several Mendelian [39-41] and complex disorders [42,43].

It is also possible that gene-gene interactions account for much of the unexplained variance in AD status [44]. These interactions are widespread and common [45,46] and approaches to understand the effects of epistatic interactions exist and continue to mature [44,47]. Several interesting candidate interactions have been identified and Ebbert et al. 2013 (accepted) recently demonstrated that allowing interactions improves the diagnostic utility of the known AD markers. Unfortunately, the complexity of this problem and the
extremely large samples sizes required to perform agnostic screens for gene-gene interactions make it very difficult to conduct effective screens for these effects.

AD is a highly complex disease with substantial genetic and environmental components. Our results suggest that genetic variance accounts for ~30% of phenotypic variance, but over 75% of this phenotypic variance remains unexplained by currently identified AD genes. Future AD genetics research must leverage larger samples and novel technologies such as NGS to identify rare, high penetrant variation and gene-gene interactions that are likely to explain the remaining genetic and phenotypic variance in AD.

Genetic research in AD has followed roughly the same model as the study of other complex diseases; largely focusing on the identification of common variants of modest effect using association studies. Scientist in many disease fields have successfully identified numerous associated variants (this is a small representative sample [48-57]). The transition from a focus on common variants to a focus on the identification of low frequency variants is now underway. These rare, functionally relevant markers are often more easily characterized than common variants of small effect. This will lead to strong and testable hypotheses for the development of therapeutics, thus accelerating the progress toward effective prevention and treatment.

5. Acknowledgments

Also, we would like to thank Mark T.W. Ebbert for his insightful comments on the manuscript.
5. 1 *Alzheimer’s Disease Genetics Consortium*

Biological samples and associated phenotypic data used in primary data analysis were stored at the Principal Investigator’s institutions, and at the National Cell Repository for Alzheimer’s Disease (NCRAD), at the NIA Genetics of Alzheimer’s Disease Data Storage Site (NIAGADS) at the University of Pennsylvania, and the NIA Alzheimer’s Disease Genetics Consortium Data Storage Site at the University of Pennsylvania. Contributors to the genetic analysis data included Principal Investigators on projects that were individually funded by NIA, other NIH institutes, or private entities.

The National Institutes of Health, National Institute on Aging (NIH-NIA) supported this work through the following grants: ADGC, U01 AG032984, RC2 AG036528; NACC, U01 AG016976; NCRAD, U24 AG021886; NIA LOAD, U24 AG026395, U24 AG026390; Banner Sun Health Research Institute P30 AG019610; Boston University, P30 AG013846, U01 AG10483, R01 CA129769, R01 MH080295, R01 AG017173, R01 AG025259, R01AG33193; Columbia University, P50 AG008702, R37 AG015473; Duke University, P30 AG028377, AG05128; Emory University, AG025688; Group Health Research Institute, U01 AG06781, U01 HG004610; Indiana University, P30 AG013854; Johns Hopkins University, P50 AG005146, R01 AG020688; Massachusetts General Hospital, P50 AG005134; Mayo Clinic, P50 AG016574; Mount Sinai School of Medicine, P50 AG005138, P01 AG002219; New York University, P30 AG08051, MO1RR00096, UL1 RR029893, 5R01AG012101, 5R01AG022374, 5R01AG013616, 1RC2AG036502, 1R01AG035137; Northwestern University, P30 AG013854; Oregon
Health & Science University, P30 AG008017, R01 AG026916; Rush University, P30 AG010161, R01 AG019085, R01 AG15819, R01 AG17917, R01 AG30146; TGen, R01 NS059873; University of Alabama at Birmingham, P50 AG016582, UL1RR02777; University of Arizona, R01 AG031581; University of California, Davis, P30 AG010129; 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 AG005131; University of California, San Francisco, P50 AG023501, P01 AG019724; University of Kentucky, P30 AG028383, AG05144; University of Michigan, P50 AG008671; University of Pennsylvania, P30 AG010124; University of Pittsburgh, P50 AG005133, AG030653; University of Southern California, P50 AG005142; University of Texas Southwestern, P30 AG012300; University of Miami, R01 AG027944, AG010491, AG027944, AG021547, AG019757; University of Washington, P50 AG005136; Vanderbilt University, R01 AG019085; and Washington University, P50 AG005681, P01 AG03991. The Kathleen Price Bryan Brain Bank at Duke University Medical Center is funded by NINDS grant # NS39764, NIMH MH60451 and by Glaxo Smith Kline. Genotyping of the TGEN2 cohort was supported by Kronos Science. The TGen series was also funded by NIA grant AG034504 to AJM, The Banner Alzheimer’s Foundation, The Johnnie B. Byrd Sr. Alzheimer’s Institute, the Medical Research Council, and the state of Arizona and also includes samples from the following sites: Newcastle Brain Tissue Resource (funding via the Medical Research Council, local NHS trusts and Newcastle University), MRC London Brain Bank for Neurodegenerative Diseases (funding via the Medical Research Council), South West Dementia Brain Bank (funding via numerous sources including the Higher Education
Funding Council for England (HEFCE), Alzheimer’s Research Trust (ART), BRACE as well as North Bristol NHS Trust Research and Innovation Department and DeNDRoN), The Netherlands Brain Bank (funding via numerous sources including Stichting MS Research, Brain Net Europe, Hersenstichting Nederland Breinbrekend Werk, International Parkinson Fonds, Internationale Stiching Alzheimer Onderzoek), Institut de Neuropatologia, Servei Anatomia Patologica, Universitat de Barcelona. 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, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc, F. Hoffman-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 AG024904, RC2 AG036535, K01 AG030514. We thank Drs. D. Stephen Snyder and Marilyn Miller from NIA who are ex-officio ADGC members. Support was also from the Alzheimer’s Association (LAF, IIRG-08-89720; MP-V, IIRG-05-14147) and the US Department of Veterans Affairs Administration, Office of Research and Development, Biomedical Laboratory Research Program. P.S.G.-H. is supported by Wellcome Trust, Howard Hughes Medical Institute, and the Canadian Institute of Health Research.

5.2 Disclosure Statement

The authors declare no conflicts of interest.
References


Figure Legends

Figure 1. Unexplained Alzheimer’s disease variance, by chromosome. In this figure we show phenotypic variance, by chromosome, explained by all SNPs. Error bars correspond to standard error.
Figure 2. Unexplained Alzheimer’s disease variance, by chromosome, excluding known Alzheimer’s disease markers. This figure is the same as Figure 1 except we have removed variance explained by known Alzheimer’s disease markers. Error bars correspond to standard error.
**Figure 3.** Variant search space. Real and hypothetical variants are graphed by effect size (y-axis) and population frequency (x-axis). Known Alzheimer’s disease SNPs are blue circles and hypothetical SNPs are red squares. The large box on the right outlined with dots, is the GWAS search space and the smaller box on the left, outlined with dashes, is the next-generation sequencing search space. SNPs identified in recent GWAS are located in the labeled oval inside the GWAS search space. Known Alzheimer’s disease SNPs are those found in Table 1 as well as APP and TREM2, which are both labeled on the graph.
Table 1. Late-onset Alzheimer’s disease associated genes/variants.

<table>
<thead>
<tr>
<th>Variant</th>
<th>Gene</th>
<th>Abbreviation</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs429358</td>
<td>Apolipoprotein E (ε4 allele)[58]</td>
<td>APOE</td>
<td>3.685</td>
</tr>
<tr>
<td>rs7412</td>
<td>Apolipoprotein E (ε2 allele)[59]</td>
<td>APOE</td>
<td>0.621</td>
</tr>
<tr>
<td>rs744373</td>
<td>Bridging Integrator 1[32]</td>
<td>BIN1</td>
<td>1.166</td>
</tr>
<tr>
<td>rs11136000</td>
<td>Clusterin[33,60]</td>
<td>CLU</td>
<td>0.879</td>
</tr>
<tr>
<td>rs3764650</td>
<td>ATP-binding cassette, sub-family A (ABC1), member 7[31]</td>
<td>ABCA7</td>
<td>1.229</td>
</tr>
<tr>
<td>rs3818361</td>
<td>Complement component (3b/4b) receptor 1 (Knops blood group)[60]</td>
<td>CR1</td>
<td>1.174</td>
</tr>
<tr>
<td>rs3851179</td>
<td>Phosphatidylinositol binding clathrin assembly protein[33]</td>
<td>PICALM</td>
<td>0.879</td>
</tr>
<tr>
<td>rs610932</td>
<td>Membrane-spanning 4-domains, subfamily A, member 6A[31]</td>
<td>MS4A6A</td>
<td>0.904</td>
</tr>
<tr>
<td>rs3865444</td>
<td>CD33 molecule[20,31]</td>
<td>CD33</td>
<td>0.893</td>
</tr>
<tr>
<td>rs670139</td>
<td>Membrane-spanning 4-domains, subfamily A, member 4E[31]</td>
<td>MS4A4E</td>
<td>1.079</td>
</tr>
<tr>
<td>rs9296559</td>
<td>CD2-associated protein[20,31]</td>
<td>CD2AP</td>
<td>1.117</td>
</tr>
</tbody>
</table>

The dbSNP identification number, gene name, gene abbreviation, and odds ratio for each of the top variants from the Alzgene.org meta-analyses. The SNP in CD2AP, rs9349407, is not present in this sample, so we used rs9296559 instead as a proxy. These two SNPs are close together (1,108 base pairs apart) and in very high LD ($r^2=1$).
We report total individuals, sex, case-control status, and average age for the 10,922 individuals analyzed in this report.
Table 3. Summary of genetic and phenotypic variance measurements.

<table>
<thead>
<tr>
<th>SNP Set</th>
<th>Genetic Variance (Standard Error)</th>
<th>% Phenotypic Variance (Standard Error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All SNPs</td>
<td>0.071 (0.0072)</td>
<td>33.12% (0.033)</td>
</tr>
<tr>
<td>APOE ε2/ε4</td>
<td>0.020 (0.0066)</td>
<td>5.92% (0.033)</td>
</tr>
<tr>
<td>All known AD markers (Table 1)</td>
<td>0.025 (0.0066)</td>
<td>7.78% (0.033)</td>
</tr>
<tr>
<td>All SNPs excluding known AD markers</td>
<td>0.046 (0.0060)</td>
<td>25.34% (0.033)</td>
</tr>
</tbody>
</table>

In this table we summarize our results, showing genetic and percent phenotypic variance for four different subsets of SNPs: all SNPs in the ADGC dataset, the two APOE alleles (ε2 and ε4), all known AD markers (as listed in Table 1), and all SNPs excluding those in Table 1.
Supplementary Table 1. Missingness rates for covariants and case-control status

<table>
<thead>
<tr>
<th>Covariates/Case-control Status</th>
<th># of individuals missing the given covariate (% missing)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1491 (7.57%)</td>
</tr>
<tr>
<td>Sex</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Case-control Status</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>APOE (rs429358/rs7412)</td>
<td>579 (2.94%)</td>
</tr>
<tr>
<td>10 Principal Components</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>BIN1 (rs744373)</td>
<td>2872 (14.58%)</td>
</tr>
<tr>
<td>CLU (rs11136000)</td>
<td>3009 (15.28%)</td>
</tr>
<tr>
<td>PICALM (rs3851179)</td>
<td>211 (1.07%)</td>
</tr>
<tr>
<td>MS4A6A (rs610932)</td>
<td>30 (0.15%)</td>
</tr>
<tr>
<td>CD33 (rs3865444)</td>
<td>392 (1.99%)</td>
</tr>
<tr>
<td>MS4A4E (rs670139)</td>
<td>192 (0.98%)</td>
</tr>
<tr>
<td>CD2AP (rs9296559)</td>
<td>2374 (12.06%)</td>
</tr>
<tr>
<td>CR1 (rs3818361)</td>
<td>54 (0.27%)</td>
</tr>
<tr>
<td>ABCA7 (rs3764650)</td>
<td>2964 (15.05%)</td>
</tr>
</tbody>
</table>

The Alzheimer’s Disease Genetics Consortium dataset consists of 19,692 total individuals. We removed any individuals missing any of the covariates (listed here) or case-control status (included in this table).