Key variants via Alzheimer's Disease Sequencing Project whole genome sequence data

Authors

Yanbing Wang,[a] Chloé Sarnowski,*[a,b] Honghuang Lin,[c] Achilleas N Pitsillides,[a] Nancy L Heard-Costa,[a,d] Seung Hoan Choi,[a] Dongyu Wang, [a] Joshua C Bis,[e] Elizabeth E Blue,[f,g] Alzheimer's Disease Neuroimaging Initiative (ADNI), Eric Boerwinkle,[b] Philip L De Jager,[h,i] Myriam Fornage,[b,j] Ellen M Wijsman,[k] Sudha Seshadri,[d,l,m] Josée Dupuis,[a,n] Gina M Peloso,[a] Anita L DeStefano[a,d], for the Alzheimer's Disease Sequencing Project (ADSP)

Affiliations

a. Department of Biostatistics, Boston University, School of Public Health, Boston, MA, USA
b. Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA
c. Department of Medicine, University of Massachusetts Chan Medical School, Worcester, MA, USA
d. The Framingham Heart Study, Framingham, MA, USA
e. Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA
f. Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, WA, USA
g. Brotman Baty Institute, Seattle, WA, USA
h. Center for Translational & Computational Neuroimmunology, Department of Neurology, Columbia University Irving Medical Center, New York, NY, USA
i. Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University Irving Medical Center, New York, NY, USA

j. Brown Foundation Institute of Molecular Medicine, McGovern Medical School, University of Texas Health Science Center at Houston, Houston, TX, USA
k. Div. of Medical Genetics and Dept. Biostatistics Statistical Genetics Lab, University of Washington, Seattle, WA, USA
l. Glenn Biggs Institute for Alzheimer's & Neurodegenerative Diseases, The University of Texas Health Science Center at San Antonio, San Antonio, TX, USA
m. Boston University School of Medicine, Department of Neurology, Boston, MA, USA
n. Department of Epidemiology, Biostatistics and Occupational Health, School of Population and Global Health, McGill University, Montreal, Canada

Corresponding authors
Dr. Chloé Sarnowski, email: Chloe.Sarnowski@uth.tmc.edu (editorial office correspondence)
Dr. Anita L Destefano, email: adestef@bu.edu

Drs. Yanbing Wang and Chloé Sarnowski contributed equally to this work. Drs. Gina Peloso and Anita L DeStefano contributed equally to the supervision of this work.

Abbreviations: Alzheimer's disease (AD), the Alzheimer's Disease Sequencing Project (ADSP), Black/African-American (AA), Combined Multivariate and Collapsing (CMC), White/European-ancestry (EA), Genome-Wide Association Studies (GWAS), Genetic Relationship Matrix (GRM), Hispanic/Latino (HI), Linkage Disequilibrium (LD), Minor Allele Count (MAC), Minor Allele Frequency (MAF), Mild Cognitive Impairment (MCI), Principal Component Analysis (PCA), Quality Control (QC), Sequence Kernel Association Test (SKAT), variant-Set Test for Association using Annotation infoRmation (STAAR), Whole Genome Sequencing (WGS)

Abstract

INTRODUCTION: Genome-wide association studies (GWAS) have identified loci associated with Alzheimer's disease (AD) but did not identify specific causal genes or variants within those loci. Analysis of whole genome sequence (WGS) data, which interrogates the entire genome and captures rare variations, may identify causal variants within GWAS loci.

METHODS: We performed single common variant association analysis and rare variant aggregate analyses in the pooled population (N cases=2,184, N controls=2,383) and targeted analyses in sub-populations using WGS data from the Alzheimer's Disease Sequencing Project (ADSP). The analyses were restricted to variants within 100 kb of 83 previously identified GWAS lead variants.

RESULTS: Seventeen variants were significantly associated with AD within five genomic regions implicating the genes OARD1/NFYA/TREML1, JAZF1, FERMT2, and SLC24A4. KAT8 was implicated by both single variant and rare variant aggregate analyses.

DISCUSSION: This study demonstrates the utility of leveraging WGS to gain insights into AD loci identified via GWAS.

Keywords: Alzheimer's disease; Whole Genome Sequencing; Association Analyses; Single Nucleotide Variations; Diverse Populations; Genome Wide Association Study

1. Introduction

Alzheimer's disease (AD), the most common cause of dementia, has been ranked as the 6th leading cause of death in the United States and the 5th leading cause of death in older people (≥65 years old). Although the role of genetic factors in the development of AD has been widely recognized, genome-wide association studies (GWAS) typically identify regions or loci rather than specific genes and/or variants. Additionally, the loci identified by GWAS only explain a small portion of the total heritability of AD (h^2_{AD} = [0.58–0.79]).[1] Next-generation sequencing technology applied in diverse populations as part of the Alzheimer's Disease Sequencing Project (ADSP) may help to elucidate the genetic architecture of AD, and thus, aid in the development of effective strategies to diagnose, prevent and treat AD.[2]

A recent large GWAS totalling 111,326 clinically diagnosed/'proxy' AD cases and 677,663 controls has identified over 70 loci associated with AD and related dementias.[3] However, the characterization of these loci remains incomplete. Leveraging whole genome sequence (WGS) data that encompasses the full spectrum of genetic variation including common and rare variants might identify important AD genes within these GWAS loci and provide a better understanding of the biological mechanisms involved in the pathophysiology of AD. Previous studies used WGS to identify genetic loci associated with AD.[3-6] A family-based study conducted in 2,247 subjects from NIMH/NIA with replication in 1,669 independent participants from the ADNI/ADSP identified 13 novel AD candidate loci with rare-variant signals (FNBP1L, SEL1L, LINC00298, PRKCH, C15ORF41, C2CD3, KIF2A, APC, LHX9, NALCN, CTNNA2, SYTL3, and CLSTN2).[3] More recently, the same team investigated association of groups of rare variants in the same datasets using a sliding-window approach and identified two novel genes (DTNB and DLG2) associated with AD.[4] Additional studies conducted in Asian populations highlighted the importance of increasing representation of understudied population groups and value of WGS to uncover population-specific genetic loci.[5,6]

In this work, we focused on deep interrogation of known AD GWAS loci[7] using the ADSP WGS data. The ADSP aims to identify protective or risk genetic contributors for AD in populations with diverse ancestry. The ADSP has generated single nucleotide variant and

insertion/deletion (indel) calls based on WGS data from 4,789 participants, which are publicly available (R1 data release https://dss.niagads.org/datasets/ng00067-v1/). The goal of the current study is not replication of prior GWAS findings as we are underpowered to do so. In addition, the ADSP sample in the current analyses is not independent of the sample used in Bellenguez et al.[7] Instead we aim to provide a more comprehensive look at GWAS loci.

We conducted single variant association analyses and rare variant aggregation association tests using the R1 WGS data of ADSP to identify specific genetic variants, genes and non-coding regions associated with AD within previously identified AD loci. We also examined multi-ancestry evidence for AD associations through population-specific analyses in White/European-ancestry (EA), Black/African-American (AA) and Hispanic/Latino (HI) subgroups, and a multi-population meta-analysis. The insights gained from our analysis will contribute to a better understanding of the AD pathogenesis and to potentially identify new targets for AD drug and treatment.

2. Methods

2.1 Study Participants

Data from the ADSP is available to qualified investigators via the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS) (https://dss.niagads.org/). This study was done under an approved NIAGADS research use statement and local Institutional Review Board approval. The current analyses focused on participants with WGS data in the NIAGADS file set named "R1 5K WGS Project Level VCF". WGS data have been generated in multiple cohorts as part of the ADSP. The ADSP data included in this study are comprised of distinct phases including the Discovery, Discovery Extension, and Augmentation phases. The Discovery phase WGS was generated from individuals of multiplex AD families as previously described.[8-10] The Discovery Extension phase consisted of a family component and a case control component. The Discovery Extension family component WGS was generated on additional members of selected families from the Discovery phase as well as

members of 77 additional families. A set of 114 Hispanic control individuals was also sequenced with the family component.

A focus of the Discovery Extension case control component was to increase the diversity of the ADSP samples. The ADSP Discovery Extension WGS was generated on 3,082 individuals, with approximately one third from EA, AA, and HI populations. In the ADSP Discovery and Discovery Extension phases sequencing was performed at three sequencing centers via the National Human Genome Research Institute (NHGRI). Sequence data for ADSP Augmentation Studies were supported by NIA and private funding and are shared with the research community via NIAGADS. The ADSP data coordinating center, the Genomic Center for AD (GCAD), produced a jointly called and quality controlled (QC'ed) data set for WGS10 that included the ADSP WGS Discovery, Discovery Extension, and from the Augmentation phase, the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. Details of studies included in the ADSP can be found at NIAGADS under dataset: NG00067 ADSP Umbrella Study (https://dss.niagads.org/datasets/ng00067/).

2.2 WGS Quality Control

Low-quality variants were filtered out based on the GCAD provided flags, which were generated separately for the Family, Case-Control, and ADNI sub-studies.[10] In addition, GCAD provided the ABHet ratio computed as (the total reference reads over all heterozygous genotypes)/(total alternative and reference reads over all heterozygous genotypes). A variant was excluded if it failed the GATK Variant Quality Score Recalibration (VQSR) filter, all genotypes were missing, was monomorphic, or if it had low call rate across all studies. Additional filtering was implemented within sub-study. If a variant had high read depth (>500 reads) within a study or had ABHet < 0.25 or ABHet > 0.75 within a sub-study, all the genotypes within that sub-study were set to missing. After these filters were applied, a final call rate filter of 95% across all sub-studies was implemented.

2.3 AD Phenotype Definition

The ADSP provides different AD status variable definitions for participants included via case-control versus family-based studies. In the current analysis, for individuals in the ADSP case-control study, we defined AD cases as individuals with either prevalent or incident AD. Case-control individuals with no prevalent or incident AD were defined as controls and those with missing status were defined as unknown. In the ADSP family phenotype file, possible values for the AD status variable include no dementia, definite AD, probable AD, possible AD, family-reported AD, other dementia, family reported no dementia, and unknown. For family-based individuals, we defined an AD case as either possible, probable or definite AD. AD controls were defined as individuals coded as no dementia. We redefined individuals with family-reported AD, other dementia, or unknown status as missing AD status. The ADNI phenotype data, which is part of the ADSP Augmentation study, provides information on mild cognitive impairment (MCI) in addition to AD status. Individuals with a current diagnosis of MCI (N=320) were included as AD controls in the current study. After selecting genetically unique individuals with AD status available, a total of 4,567 participants (2,383 controls and 2,184 cases) with WGS were included in the analyses.

2.4 Pooled sample single-variant association analysis

Single-variant association analysis of AD was performed on variants within GWAS loci for participants with both phenotype and genotype data available using GENESIS.[11] Principal component analysis (PCA) was performed as described in the supplemental methods to assess and adjust for genetic ancestry of the study participants (**Figure A1**). The WGS samples included in the ADSP R1 WGS data set were sequenced across four sequencing centers (Baylor College of Medicine Human Genome Sequencing Center, The Broad Institute, McDonnell Genome Institute at Washington University School of Medicine, and Illumina) and 2 sequencing platforms (Illumina HiSeq 2000/2500, and Illumina HiSeq X Ten). In order to control for the effects from study design and technical differences, we generated indicator variables (study×sequencing center×sequencing platform) with 10 categories based on **Table 1**. We considered these indicator variables as technical covariates and defined case-control x

Broad x HiSeq X Ten, which had the largest number of observations, as the reference group. We used a generalized logistic mixed-effects model to account for relatedness through a genetic relationship matrix (GRM). The GRM was estimated based on the same variants used in the PCA. We included sex, the technical covariates, and PC2 (based on a Bonferroni corrected significant $p < 0.0016$ for testing 32 PCs) as covariates in the null model. We performed the analysis across autosomes, and kept variants satisfying the criteria: call rate higher than 95% and minor allele count (MAC) higher than 20.

To determine if significant variants identified provided distinct signals from the lead GWAS variants,[7] conditional analyses were performed in loci displaying significant associations. Genotype data, coded as 0/1/2, for the lead GWAS variants in these loci was included in the null model in addition to the covariates. Association analyses conditioned on the lead GWAS variants were then rerun for the loci of interest.

2.5 Population specific association analysis

We conducted population specific analyses (null model and association analyses) for AD using GENESIS, accounting for genetic relatedness using a GRM. We defined three population groups (EA, AA, and HI). We selected a total of 2,144 EA participants based on PCA analysis performed using both the ADSP and the Human Genome Diversity Project (HGDP). Only participants who were not outliers based on 6 standard deviations (SD) from the mean for PCs 1-4 calculated in the European HGDP groups (Adygei, Basque, French, BergamoItalian, Orcadian, Russian, Sardinian, and Tuscan) were retained. We selected a total of 1,028 AA and 1,548 HI participants based on reported race and ethnicity. A total of 38 participants who identified as both African-American and Hispanic were placed in the Hispanic population. We included in the null model, in each population group, covariates associated with AD status at $P \leq 0.05$. The EA null model included sex, ADSP family study status, Illumina sequencing center, HiSeq X Ten platform, PC 2, PC 9, and PC 15. The HI null model included sex, all sequencing centers, HiSeq X Ten platform, PC 13, PC 16, and PC 17. The AA null model included sex, Illumina sequencing center, and PC 1. We performed association analyses, in

each population group, and retained the results with call rate higher than 95% and minor allele count (MAC) higher than 20. In addition, we performed a multi-population meta-analysis using three different models (fixed-effect, random-effect, and Han & Eskin's modified random-effect) implemented in Metasoft[12] by combining the population specific results satisfying the criteria of a within population MAC higher than 10. We then kept the meta-analysis results passing a total MAC across population groups higher than 20.

2.6 Gene-based tests

We tested the association of aggregate groups of low frequency (minor allele frequency (MAF) < 5%) or rare (MAF < 1%) genetic variants with AD status. Annotation for all called variants was generated using Ensembl VEP91 by the ADSP annotation working group. We selected missense or loss of function (lof) genetic variants based on the most severe variant consequence according to the ADSP Annotation WG Ranking Process and listed in the annotation file (frameshift variant, inframe deletion, inframe insertion, missense variant, protein-altering variant, splice acceptor variant, splice donor variant, start lost, stop gained, and stop lost). We conducted Sequence Kernel Association Test (SKAT, mmskat) and burden tests (combined multivariate and collapsing (CMC), emmaxCMC) with EPACTS (Efficient and Parallelizable Association Container Toolbox) using mixed-effect models adjusted for sex, technical covariates, and PCs significantly associated with AD status (PC2) accounting for genetic relatedness (GRM).

2.7 Non-coding rare variant analysis

For non-coding rare variant analysis, we used annotations from WGSA v0.8[13] including annotations from ANNOVAR, VEP, SnpEff, COSMIC and SPIDEX. We conducted rare variant analysis using the variant-Set Test for Association using Annotation infoRmation (STAAR) method,[14] which was developed to boost power of rare variant analyses by effectively incorporating both variant functional categories and multiple complementary functional annotations while accounting for relatedness and population structures. We used the same

covariates (sex, technical covariates, and significant PCs) in the model as in single-variant analysis. The GRM was incorporated to account for relatedness among samples.

We aggregated sites that overlap enhancers and promoters around gene transcription start sites (TSS). The promoters within 5KB of a TSS that overlap DNase hypersensitivity sites (DHS) are defined as at least one WGSA H3K4me3 annotation for brain tissues (E067, E068, E069, E070, E071, E072, E073, E074, E081, E082), and the enhancers within 20KB of a TSS are defined by EnhancerFinder in Brain. We incorporated annotations from WGSA in the analysis, which include MAF, functional scores (GERP, GenoCanyon, RegulomeDB, FUNSEQ, CADD, Fathmm, EIGEN-PC), and the ENCODE score (DNASE). We then transformed the annotation scores to phred-scaled scores using -10×log10(rank(-score)/M), where M is the total number of variants tested in the analysis.

2.8 Focus on GWAS loci

Given the limited power to detect novel loci with the current sample size, we focused on exploiting WGS to provide insights on previously reported AD GWAS loci. We used the variants listed in Supplemental Table 5 in Bellenguez et. al as the previously reported AD GWAS top variants. For single variant association analyses, we looked up these lead variants in the ADSP WGS data. We then assessed ADSP WGS associations within 100KB of each lead GWAS variant. For gene-based and non-coding rare variant analysis, we obtained the results for genes or regions in the 100KB window around each lead variant. We included genes or non-coding regions for which any portion overlapped with the specified window. Using this paradigm, we identified 303 genes within 100kb of the index SNPs.

In general, we defined a threshold for statistical significance equal to 0.05/number of statistical tests and a suggestive threshold as 1/number of statistical tests. Within a 100kb window, many single variant tests were highly correlated. Therefore, we computed the effective number of independent tests using the simpleM approach[15] and used the effective number of tests in the denominator when computing a window-specific threshold for single variant association

testing. Effective number of tests were computed across the pooled sample and within each population sub-group (**Table A1**). **Figure 1** provides an overview of our analysis workflow.

We leveraged publicly-available multi-omic resource generated by applying quantitative trait locus (xQTL) analyses to RNA sequence and DNA methylation from the dorsolateral prefrontal cortex of 411 older adults from the Religious Orders Study (ROS) and Memory and Aging Project (MAP) studies[16] to look-up the main genetic variants from the pooled association analysis.

3. Results

3.1 Description of ADSP data

After the QC of the ADSP data release NG00067.v2, there were over 95 million variants across 4,733 participants. A total of 4,567 individuals (2,383 controls, 2,184 cases) have available AD status and contributed to the analyses, among which 807 are from the ADSP family study, 2,963 are from the ADSP case-control study, and 797 are from ADNI. The participants included in the analyses were more likely to be women (61.8%) than men. The distribution of study design membership, sequencing centers, and sequencing platforms is summarized in **Table 1**.

3.2 Pooled sample single-variant association analysis

Genome-wide, there were about 20 million variants with call rate higher than 95% and MAC higher than 20 in the pooled sample analysis. Our model that included GRM and PC adjustments showed acceptable type-I error control ($\lambda = 1.05$, **Figure A2**). As expected, the strongest association was observed at the *APOE* locus, where the major *APOE* variant rs429358 ($p = 7.2 \times 10^{-77}$) was the top hit.

Among the specific lead GWAS variants from Bellenguez et al,[7] none reached the strict significance threshold ($p < 6 \times 10^{-4}$, Bonferroni correction for the total number of variants tested) in the pooled sample association analysis. Using the suggestive significance threshold ($p < 0.012$; 1/83), we found associations for rs7401792 ($p = 7.3 \times 10^{-4}$, MAF= 49.2%) in the

306 *SLC24A4* locus, rs75932628 ($p = 3 \times 10^{-3}$, MAF= 0.35%) in the *TREM2* locus, rs616338 (p=
307 4.4 × 10-3, MAF= 0.71%) in the *ABI3* locus, rs1358782 ($p = 4.9 \times 10^{-3}$, MAF= 22.9%) in the
308 *RBCK1* locus, rs1160871 ($p = 6.9 \times 10^{-3}$, MAF= 40.2%) in the *JAZF1* locus, and rs602602 (p
309 = 0.12, MAF= 27.7%) in the *MINDY2* locus. Full results for the 83 lead GWAS variants are
310 provided in the supplement (**Table A2**).
311 Applying the significance thresholds based on the effective number of tests within 100 kb
312 windows around the lead GWAS variants (**Table A1**), we identified 17 significant variants in
313 the single variant association analysis in the pooled sample (**Table 2**). These 17 variants occur
314 in five genomic regions on chromosomes 6, 7, 14, and 16. Forest plots for the top variant in
315 each of these five regions are presented in **Figure 2**. Examination of linkage disequilibrium
316 (LD) patterns show near perfect LD among the variants identified on chromosome 6, on
317 chromosome 7 or for the one region on chromosome 14 with multiple variants. Only a single
318 variant was identified on chromosome 16 and in one region on chromosome 14. Detailed LD
319 information is provided in the supplement (**Figure A3**).
320 Conditional analyses were performed to determine if these associations represented the same
321 signal as the lead GWAS variant from Bellenguez et al or a distinct signal. As shown in **Table**
322 **2**, the inclusion of the lead GWAS variant in the association model did not mitigate the
323 association indicating the variants identified represent a distinct signal from the lead GWAS
324 variants.
325
326 3.3 Population specific single-variant association analysis and multi-population meta-analysis
327 We conducted population specific association analyses in the three main subgroups (N=2,043
328 EA; N=995 AA, and N=1,516 HI participants). There was acceptable type-I error in the
329 population specific analyses and the multi-population meta-analysis (**Figures A4-A7**). We
330 confirmed the significant association of the *APOE* locus (rs429358) in both the population
331 specific analyses and the multi-population meta-analysis. However, as found in previous
332 studies[17,18] the association was weaker in the Hispanic population (beta = 1.17 in EA, 1.02 in
333 AA, and 0.59 in HI).

Population specific single variant analyses identified 23 significant variants in 11 loci within 100 kb of the lead GWAS variants (**Table 3**). Of these, 15 variants (8 loci) were identified in the EA population and 8 variants (3 loci) in the AA population. No significant variants were identified in the HI population. The only overlap in significant variants between the pooled sample and the population specific single-variant analyses was a missense variant in *KAT8*, which was very rare in the AA and HI subsamples (**Table A3**).

3.4 Gene-based tests

QQ plots for the SKAT and burden tests are provided in **Figure A8**. Using the multiple-testing correction threshold for 303 genes ($p < 1.7 \times 10^{-4}$), *KAT8* ($p = 2.2 \times 10^{-5}$, MAF< 5%) lying within 100 KB of a GWAS variant was detected to be significantly associated with AD status by SKAT, and it was also shown as a suggestive association ($p < 3.4 \times 10^{-3}$) using CMC ($p = 9.2 \times 10^{-4}$). SKAT detected suggestive associations in *LAIR1* ($p = 0.0023$, MAF < 5%) and *ATF5* ($p = 5.7 \times 10^{-4}$, MAF < 5%) within 100KB. CMC identified *TREM2* ($p = 3.3 \times 10^{-3}$ for MAF < 1% and 8.9×10^{-4} for MAF< 5%) within 100KB of GWAS variants.

3.5 Non-coding rare variant analysis

QQ plots showed deflated type-I error ($\lambda = 0.75$), most likely due to small sample size, in the STAAR rare-variant analysis. (**Figure A9**). No regions were identified as significant using the STAAR approach. The top STAAR results overlapping GWAS loci are shown in **Table A3**.

3.6 xQTL analysis lookup

We did not identify significant mQTL or eQTL associations for the main genetic variants identified in the pooled analysis. We could not look up some of the less frequent variants on chr7, 14 and 16 in the QTL results as these analyses were restricted to common variants. Suggestive associations have been reported between rs10947950 (chr6) and cg25473438 (beta = 0.24, $P = 1.4 \times 10^{-8}$) and between rs7155002 (chr14) and cg12072028 (beta = 0.19, $P = 5.8 \times 10^{-6}$), **Table A4**. The CpG cg12072028 is located in the intron 1 of *RIN3* and modest

associations have been described between rs7155002 and *RIN3* expression in the brain (beta = 0.04, P = 0.01), **Table A4**.

Discussion

GWAS have been essential in identifying genetic loci associated with AD. However, GWAS loci typically contain scores of genes and thousands of variants. Additional studies are needed to pinpoint specific genes or variants as the ones influencing risk for AD. WGS provides complete genomic sequence and hence enumerates both common and rare variants. WGS therefore has the potential to provide information beyond common variants that are the cornerstone of GWAS. In the current study, we have examined WGS from the ADSP R1 data set focusing on previously implicated regions to better understand important variants within AD GWAS loci in a diverse study sample. We identified 17 significant variants in five genomic regions using single variant association analysis in the pooled sample. The majority of these variants were intronic, although two intergenic and one missense variant were also identified. Bellenguez et al[7] identified multiple lead GWAS variants on chromosome 6, which yielded overlapping 100kb windows defined by our approach. We identified seven significant variants on chromosome 6 that are in nearly complete LD. Six of these variants are located within intronic regions of *OARD1* and *NFYA* genes, and one variant is very close to *APOBEC2*. One variant was intergenic with the closest gene being *TREML1*. This region contains *TREM2*, for which several rare coding variants have been implicated as conferring risk for AD.[19-22] *TREM2* showed suggestive evidence of association in gene-based analyses indicating multiple variants in this region are likely to play an important role in AD. The missense variant rs75932628 was one of the lead GWAS variants from Bellenguez et al. and has been identified as a functional variant for AD.[23-25] The variant rs75932628 has an MAF=0.0035 in the ADSP pooled sample and p=0.003 for single variant association with AD. This suggestive association is driven by the EA population (MAF=0.007) as this variant is less frequently observed in the AA (MAF=0.001) or HI (MAF=0.0003) populations. Our conditional analysis indicates that the variants we identified implicating *OARD1/NFYA/TREML1* have a distinct effect from

rs75932628. *OARD1* encodes a deacylase with a function to catalyse O-acetyl-ADP-ribose during multiple cellular processes. A homozygous mutation could lead to cell death and cause a form of childhood neurodegenerative disorder.[26] *NFYA* encodes a subunit of nuclear transcription factor Y, which is a ubiquitous transcription factor. The gene is involved in post-transcriptional regulation with tissue-specific preference, and it is suppressed in the brain of model mice with Huntington's disease[27] and spinal and bulbar muscular atrophy.[28] *TREML1* encodes a protein belonging to the family of triggering receptors expressed on myeloid cells-like (TREM). A deficiency of *TREML1* might result in haemorrhage due to localized inflammatory lesions.[29]

The five significant variants on chromosome 7 are in a strong LD block and all variants are intronic for *JAZF1*, which encodes a transcriptional repressor. The gene has been linked with diabetes mellitus and cancer, but also has a role in lipid metabolism supporting the connection between lipid levels and AD.[30] The *JAZF1* GWAS variant (rs1160871) is a strong eQTL in microglia and considered a Tier 1 (highly plausible) AD gene.[7] The significance of the variants at this locus was slightly attenuated by conditioning on the nearby GWAS variant (rs1160871), indicating that this may be a shared effect with the lead GWAS variant.

We identified two regions on chromosome 14, with significant variants intronic to *FERMT2* and *SLC24A4*. The association of the intronic variant rs7155002 for *SLC24A4* was slightly attenuated by the conditional analysis indicating that this may be a shared effect with the lead GWAS variant. Lookup in brain xQTL data shed light on potential biological regulatory mechanisms in *RIN3* that has also been implicated in AD.[31,32] *FERMT2* encodes plekstrin homology domain-containing family C member 1 and is known to be involved in APP metabolism.[33] The under-expression of *FERMT2* was associated with mature APP level increment in the cell surface.[33] Previous studies reported that *FERMT2* is also involved in cardiac and skeletal muscle development[34] and cancer progression.[35,36] *SLC24A4* encodes a member of the potassium-dependent sodium/calcium exchanger protein family and is associated with neural development.[37] A homozygous mutation in *SLC24A4* may cause amelogenesis imperfecta[38,39] but the function of *SLC24A4* in AD is not clear yet.

A rare missense variant (rs201871085, MAF = 0.0108 in the pooled sample) within *KAT8* (lysine acetyltransferase 8) on chromosome 16 was significantly associated with AD. *KAT8* was also significant in the low-frequency variant gene-based analyses. *KAT8* encodes a member of the MYST histone acetylase protein family that has a characteristic MYST domain containing an acetyl-CoA-binding site, a chromodomain typical of proteins which bind histones, and a C2HC-type zinc finger. This gene has been recently identified by two large-scale GWAS of clinically diagnosed AD and family history of AD[40,41] and by a novel knockoff method when applied to the ADSP Data.[42] Aberrant expression patterns of *KAT8* might be associated with AD progression.[43] *KAT8* appears like a promising candidate gene that is involved in cerebral development[44] and may play a role in neurodegeneration in both AD and Parkinson's disease.[45,46] We were not able to look-up the variant rs201871085 in the brain xQTL data. This variant might not have been analyzed due to a low frequency or a low quality of imputation, thus highlighting the importance of leveraging whole genome sequence data.

The ADSP represents a diverse population sample, although in this early release of ADSP WGS data the sample size within a specific population is limited (N_{EA} = 2,043, N_{AA} = 995, N_{HI} = 1,516). Population specific analyses provide information about patterns of allele frequency for AD associated variants among populations. Among the five loci identified as significant in the pooled single variant association analysis, two regions (chr6 & 16) displayed EA-specific associations and corresponded to low frequency variants in EA that were rare in other population groups. One signal (chr7) was driven by a variant common in AA with a low frequency in HI, and rare in EA. Finally, two regions (chr14) were driven by HI signals with one variant common in all population groups, and one variant common in HI but rare in EA and AA. All these results are summarized in **Table A3**.

The signals identified only in AA in the population specific analyses (chr4 and 14) corresponded to SNPs common in AA that were less common in HI and rare in EA. A few signals identified only in EA corresponded to variants that were rare in all population groups (chr8, 17 and 20). Two low frequency signals identified only in EA (chr16 and 17) corresponded to SNPs that were rare in AA and HI. Finally, three signals identified only in EA

(chr5, 14, and 16) corresponded to SNPs that were common in different population groups. All these results are summarized in **Table 3**.

A strength of this study is the analysis of WGS data that was jointly called and QC'ed by a single data coordinating center. The diversity in genetic ancestry of the participants included is another strength. Despite this diversity, a limitation of the study is the moderate sample size within each population analysed. To overcome this limitation, main analyses were focused on the pooled sample. The ADSP is ongoing with larger WGS data sets being publicly released and planned. Future analyses with larger sample size may yield additional insights, specifically for population specific effects. The current study demonstrates the importance of leveraging whole genome sequence data to gain insights into loci identified via GWAS and highlights the contribution of low frequency variants to AD risk.

References

1. Karlsson IK, Escott-Price V, Gatz M, et al. Measuring heritable contributions to alzheimer's disease: Polygenic risk score analysis with twins. *Brain Commun*. 2022;4(1):fcab308. doi: 10.1093/braincomms/fcab308.

2. Beecham GW, Bis JC, Martin ER, et al. The alzheimer's disease sequencing project: Study design and sample selection. *Neurol Genet*. 2017;3(5):e194. doi: 10.1212/NXG.0000000000000194.

3. Prokopenko D, Morgan SL, Mullin K, et al. Whole-genome sequencing reveals new alzheimer's disease-associated rare variants in loci related to synaptic function and neuronal development. *Alzheimers Dement*. 2021;17(9):1509-1527. doi: 10.1002/alz.12319.

4. Prokopenko D, Lee S, Hecker J, et al. Region-based analysis of rare genomic variants in whole-genome sequencing datasets reveal two novel alzheimer's disease-associated genes: DTNB and DLG2. *Mol Psychiatry*. 2022;27(4):1963-1969. doi: 10.1038/s41380-022-01475-0.

5. Shigemizu D, Asanomi Y, Akiyama S, Mitsumori R, Niida S, Ozaki K. Whole-genome sequencing reveals novel ethnicity-specific rare variants associated with alzheimer's disease. *Mol Psychiatry*. 2022;27(5):2554-2562. doi: 10.1038/s41380-022-01483-0.

6. Park J, Park I, Youm EM, et al. Novel alzheimer's disease risk variants identified based on whole-genome sequencing of APOE epsilon4 carriers. *Transl Psychiatry*. 2021;11(1):296-9. doi: 10.1038/s41398-021-01412-9.

7. Bellenguez C, Kucukali F, Jansen IE, et al. New insights into the genetic etiology of alzheimer's disease and related dementias. *Nat Genet*. 2022;54(4):412-436. doi: 10.1038/s41588-022-01024-z.

8. Tsapanou A, Scarmeas N, Gu Y, et al. Data from a cross-sectional study on apolipoprotein E (APOE-epsilon4) and snoring/sleep apnea in non-demented older adults. *Data Brief*. 2015;5:351-353. doi: 10.1016/j.dib.2015.09.014.

9. Chung J, Zhang X, Allen M, et al. Genome-wide pleiotropy analysis of neuropathological traits related to alzheimer's disease. *Alzheimers Res Ther*. 2018;10(1):22-z. doi: 10.1186/s13195-018-0349-z.

10. Naj AC, Lin H, Vardarajan BN, et al. Quality control and integration of genotypes from two calling pipelines for whole genome sequence data in the alzheimer's disease sequencing project. *Genomics*. 2019;111(4):808-818. doi: 10.1016/j.ygeno.2018.05.004.

11. Gogarten SM, Sofer T, Chen H, et al. Genetic association testing using the GENESIS R/bioconductor package. *Bioinformatics*. 2019;35(24):5346-5348. doi: 10.1093/bioinformatics/btz567.

12. Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet*. 2011;88(5):586-598. doi: 10.1016/j.ajhg.2011.04.014.

13. Liu X, White S, Peng B, et al. WGSA: An annotation pipeline for human genome sequencing studies. *J Med Genet*. 2016;53(2):111-112. doi: 10.1136/jmedgenet-2015-103423.

14. Li X, Li Z, Zhou H, et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet*. 2020;52(9):969-983. doi: 10.1038/s41588-020-0676-4.

15. Gao X, Starmer J, Martin ER. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol*. 2008;32(4):361-369. doi: 10.1002/gepi.20310.

16. Ng B, White CC, Klein H, et al. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. Nat Neurosci. 2017;20(10):1418-1426. doi: 10.1038/nn.4632.

17. Blue EE, Horimoto ARVR, Mukherjee S, Wijsman EM, Thornton TA. Local ancestry at APOE modifies alzheimer's disease risk in caribbean hispanics. Alzheimers Dement. 2019;15(12):1524-1532. doi: 10.1016/j.jalz.2019.07.016.

18. Campos M, Edland SD, Peavy GM. Exploratory study of apolipoprotein E epsilon4 genotype and risk of alzheimer's disease in mexican hispanics. J Am Geriatr Soc. 2013;61(6):1038-1040. doi: 10.1111/jgs.12292.

529 19. Jonsson T, Stefansson H, Steinberg S, et al. Variant of TREM2 associated with the risk
530 of alzheimer's disease. *N Engl J Med*. 2013;368(2):107-116. doi: 10.1056/NEJMoa1211103.
531 20. Li R, Wang X, He P. The most prevalent rare coding variants of TREM2 conferring risk of
532 alzheimer's disease: A systematic review and meta-analysis. *Exp Ther Med*. 2021;21(4):347.
533 doi: 10.3892/etm.2021.9778.
534 21. Jin SC, Benitez BA, Karch CM, et al. Coding variants in TREM2 increase risk for
535 alzheimer's disease. *Hum Mol Genet*. 2014;23(21):5838-5846. doi: 10.1093/hmg/ddu277.
536 22. Guerreiro R, Wojtas A, Bras J, et al. TREM2 variants in alzheimer's disease. *N Engl J*
537 *Med*. 2013;368(2):117-127. doi: 10.1056/NEJMoa1211851.
538 23. Song WM, Joshita S, Zhou Y, Ulland TK, Gilfillan S, Colonna M. Humanized TREM2
539 mice reveal microglia-intrinsic and -extrinsic effects of R47H polymorphism. *J Exp Med*.
540 2018;215(3):745-760. doi: 10.1084/jem.20171529.
541 24. Guerreiro R, Wojtas A, Bras J, et al. TREM2 variants in alzheimer's disease. *N Engl J*
542 *Med*. 2013;368(2):117-127. doi: 10.1056/NEJMoa1211851.
543 25. Xiang X, Piers TM, Wefers B, et al. The Trem2 R47H alzheimer's risk variant impairs
544 splicing and reduces Trem2 mRNA and protein in mice but not in humans. *Mol*
545 *Neurodegener*. 2018;13(1):49-6. doi: 10.1186/s13024-018-0280-6.
546 26. Sharifi R, Morra R, Appel CD, et al. Deficiency of terminal ADP-ribose protein
547 glycohydrolase TARG1/C6orf130 in neurodegenerative disease. *EMBO J*. 2013;32(9):1225-
548 1237. doi: 10.1038/emboj.2013.51.
549 27. Yamanaka T, Miyazaki H, Oyama F, et al. Mutant huntingtin reduces HSP70 expression
550 through the sequestration of NF-Y transcription factor. *EMBO J*. 2008;27(6):827-839. doi:
551 10.1038/emboj.2008.23.
552 28. Katsuno M, Adachi H, Minamiyama M, et al. Disrupted transforming growth factor-beta
553 signaling in spinal and bulbar muscular atrophy. *J Neurosci*. 2010;30(16):5702-5712. doi:
554 10.1523/JNEUROSCI.0388-10.2010.

555 29. Washington AV, Gibot S, Acevedo I, et al. TREM-like transcript-1 protects against
556 inflammation-associated hemorrhage by facilitating platelet aggregation in mice and
557 humans. *J Clin Invest*. 2009;119(6):1489-1501. doi: 10.1172/JCI36175.
558 30. Li L, Yang Y, Yang G, et al. The role of JAZF1 on lipid metabolism and related genes in
559 vitro. *Metabolism*. 2011;60(4):523-530. doi: 10.1016/j.metabol.2010.04.021.
560 31. Shen R, Zhao X, He L, et al. Upregulation of RIN3 induces endosomal dysfunction in
561 alzheimer's disease. Transl Neurodegener. 2020;9(1):26-1. doi: 10.1186/s40035-020-00206-
562 1.
563 32. Bhattacharyya R, Teves CAF, Long A, Hofert M, Tanzi RE. The neuronal-specific
564 isoform of BIN1 regulates beta-secretase cleavage of APP and abeta generation in a RIN3-
565 dependent manner. Sci Rep. 2022;12(1):3486-4. doi: 10.1038/s41598-022-07372-4.
566 33. Chapuis J, Flaig A, Grenier-Boley B, et al. Genome-wide, high-content siRNA screening
567 identifies the alzheimer's genetic risk factor FERMT2 as a major modulator of APP
568 metabolism. *Acta Neuropathol*. 2017;133(6):955-966. doi: 10.1007/s00401-016-1652-z.
569 34. Dowling JJ, Gibbs E, Russell M, et al. Kindlin-2 is an essential component of intercalated
570 discs and is required for vertebrate cardiac structure and function. *Circ Res*.
571 2008;102(4):423-431. doi: 10.1161/CIRCRESAHA.107.161489.
572 35. Shen Z, Ye Y, Dong L, et al. Kindlin-2: A novel adhesion protein related to tumor
573 invasion, lymph node metastasis, and patient outcome in gastric cancer. *Am J Surg*.
574 2012;203(2):222-229. doi: 10.1016/j.amjsurg.2011.06.050.
575 36. Sossey-Alaoui K, Pluskota E, Szpak D, Plow EF. The Kindlin2-p53-SerpinB2 signaling
576 axis is required for cellular senescence in breast cancer. *Cell Death Dis*. 2019;10(8):539-z.
577 doi: 10.1038/s41419-019-1774-z.
578 37. Larsson M, Duffy DL, Zhu G, et al. GWAS findings for human iris patterns: Associations
579 with variants in genes that influence normal neuronal pattern development. *Am J Hum*
580 *Genet*. 2011;89(2):334-343. doi: 10.1016/j.ajhg.2011.07.011.

581 38. Parry DA, Poulter JA, Logan CV, et al. Identification of mutations in SLC24A4, encoding
582 a potassium-dependent sodium/calcium exchanger, as a cause of amelogenesis imperfecta.
583 *Am J Hum Genet*. 2013;92(2):307-312. doi: 10.1016/j.ajhg.2013.01.003.
584 39. Khan SA, Khan MA, Muhammad N, et al. A novel nonsense variant in SLC24A4 causing
585 a rare form of amelogenesis imperfecta in a pakistani family. *BMC Med Genet*.
586 2020;21(1):97-6. doi: 10.1186/s12881-020-01038-6.
587 40. Jansen IE, Savage JE, Watanabe K, et al. Genome-wide meta-analysis identifies new
588 loci and functional pathways influencing alzheimer's disease risk. *Nat Genet*.
589 2019;51(3):404-413. doi: 10.1038/s41588-018-0311-9.
590 41. Marioni RE, Harris SE, Zhang Q, et al. GWAS on family history of alzheimer's disease.
591 *Transl Psychiatry*. 2018;8(1):99-6. doi: 10.1038/s41398-018-0150-6.
592 42. He Z, Liu L, Wang C, et al. Identification of putative causal loci in whole-genome
593 sequencing data via knockoff statistics. *Nat Commun*. 2021;12(1):3152-4. doi:
594 10.1038/s41467-021-22889-4.
595 43. Chen F, Chen H, Jia Y, Lu H, Tan Q, Zhou X. miR-149-5p inhibition reduces alzheimer's
596 disease beta-amyloid generation in 293/APPsw cells by upregulating H4K16ac via KAT8.
597 *Exp Ther Med*. 2020;20(5):88. doi: 10.3892/etm.2020.9216.
598 44. Li L, Ghorbani M, Weisz-Hubshman M, et al. Lysine acetyltransferase 8 is involved in
599 cerebral development and syndromic intellectual disability. *J Clin Invest*. 2020;130(3):1431-
600 1445. doi: 10.1172/JCI131145.
601 45. Dumitriu A, Golji J, Labadorf AT, et al. Integrative analyses of proteomics and RNA
602 transcriptomics implicate mitochondrial processes, protein folding pathways and GWAS loci
603 in parkinson disease. *BMC Med Genomics*. 2016;9:5-y. doi: 10.1186/s12920-016-0164-y.
604 46. Wainberg M, Andrews SJ, Tripathy SJ. Shared genetic risk loci between alzheimer's
605 disease and related dementias, parkinson's disease, and amyotrophic lateral sclerosis.
606 Alzheimers Res Ther. 2023;15(1):113-3. doi: 10.1186/s13195-023-01244-3.
607
608

Acknowledgments

We thank the contributors who collected samples used in this study, as well as patients and their families, whose help and participation made this work possible.

ADSP: Data for this study were prepared, archived, and distributed by the National Institute on Aging Alzheimer's Disease Data Storage Site (NIAGADS) at the University of Pennsylvania (U24-AG041689), funded by the National Institute on Aging (accession NG00067). The full acknowledgement statement for the ADSP can be found at: https://dss.niagads.org/datasets/ng00067/

ADNI: Data used in preparation of this article were obtained through NIAGADS. 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.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Sources of Funding

This work was funded through U01 AG058589 and U01 AG068221 from the National Institute on Aging.

ADSP: All relevant funding is listed in the full acknowledgement statement for the ADSP that can be found at: https://dss.niagads.org/datasets/ng00067/

ADNI: Data collection and sharing for ADNI was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.;Janssen Alzheimer Immunotherapy

Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.;Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Disclosures

The authors do not have declarations of interest to report. The funding sources of this study had no role in the study design, the collection, the analysis or the interpretation of data, in the writing of the report, or in the decision to submit the article for publication.

Figure captions

Figure 1. Schematic of the ADSP 5K analysis

Figure 2. Top variants identified from single variant association analysis in the pooled sample within 100kb of the 83 lead GWAS variants

Variant ID is in the form of chromosome:position (effect allele). Positions provided are on build 38. EAF is the effect allele frequency, Meta-RE is the multi-population meta-analysis using a random effect model, Meta-FE is the multi-population meta-analysis using a fixed effect model. The p-value for Meta-RE is calculated using Han and Eskin's random effects model. The effect size and its 95% CI is not shown for variants with a minor allele count (MAC) < 10 in population specific analysis.

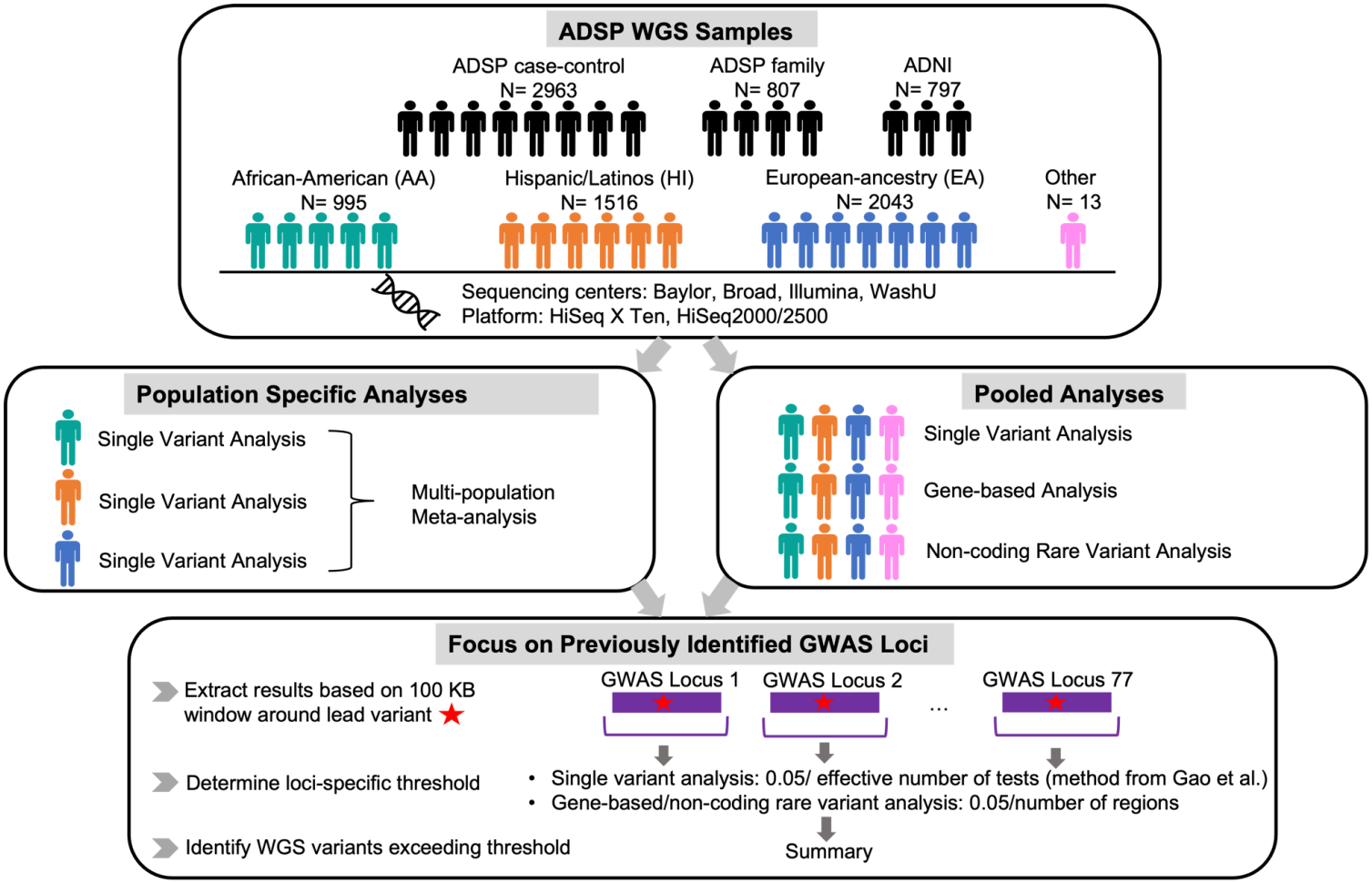

Figure 1. Schematic of ADSP 5K analysis

Color should be used.

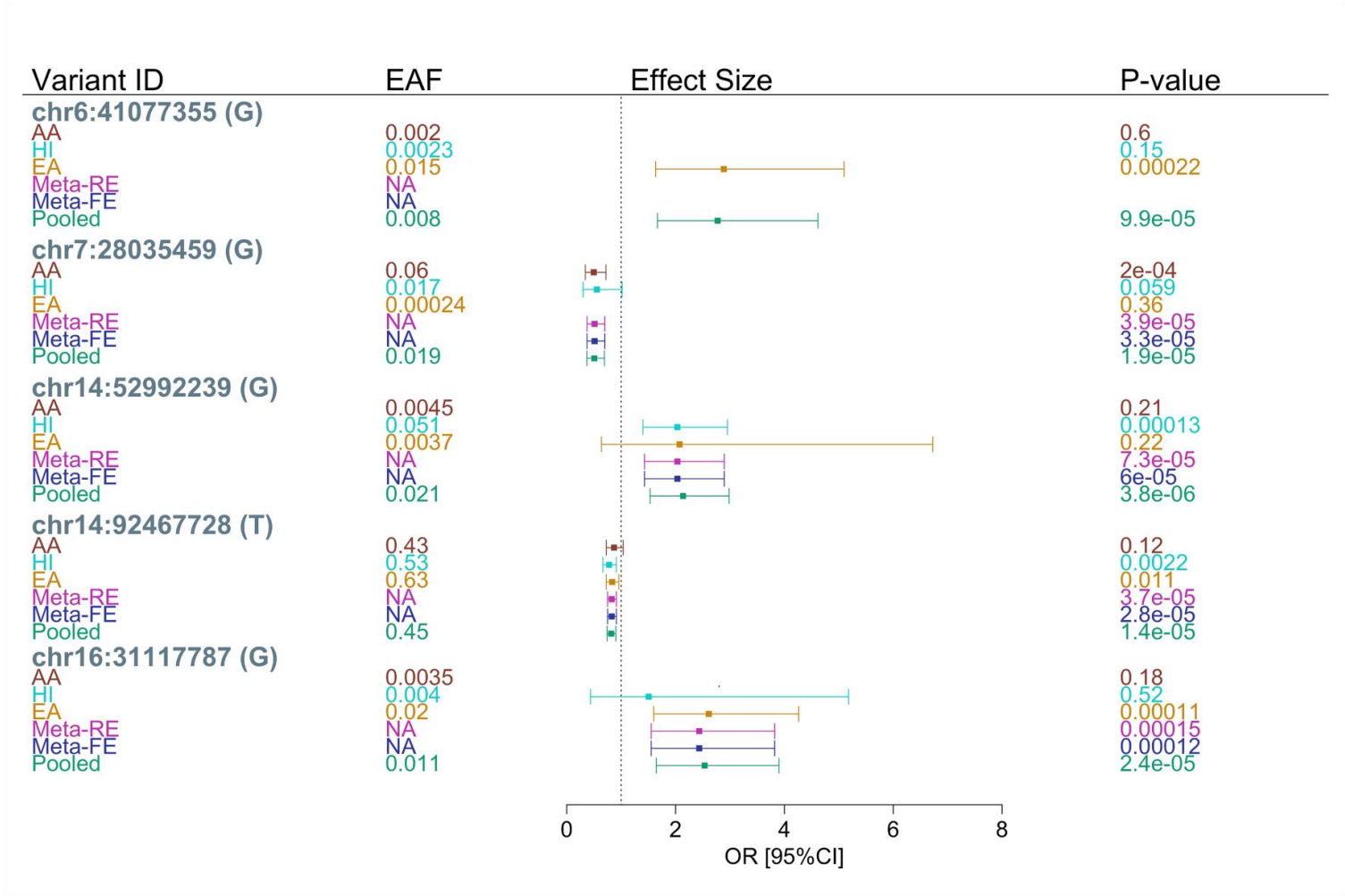

Figure 2. Top variants identified from single variant association analysis in the pooled sample within 100kb of the 83 lead GWAS variants

Color should be used.

Table 1. Characteristics of the participants included in the ADSP R1 data set

		All (N=4,567)	AA* (N=995)	HI* (N=1,516)	EA* (N=2,043)	Other (N=13)
Age (sd)		76.9 (8.3)	79.2 (7.6)	75.1 (8.5)	77.1 (8.4)	77.3 (6.6)
Alzheimer's Disease (%)						
	Case	2184 (47.8%)	463 (46.5%)	795 (52.4%)	921 (45.1%)	5 (46.2%)
	Control	2383 (52.2%)	532 (53.5%)	721(47.6%)	1122 (54.9%)	8 (53.8%)
Sex (%)						
	Female	2822 (61.8%)	710 (71.4%)	1020 (67.3%)	1086 (53.2%)	6 (38.5%)
	Male	1745 (38.2%)	285 (28.6%)	496 (32.7%)	957 (46.8%)	7 (61.5%)
Study (%)						
	ADNI	797 (17.5%)	26 (2.6%)	10 (0.7%)	750 (36.7%)	11 (84.6%)
	ADSP- case control	2963 (64.9%)	944 (94.9%)	1107 (73.0%)	911 (44.6%)	1 (7.7%)
	ADSP- family	807 (17.7%)	25 (2.5%)	399 (26.3%)	382 (18.7%)	1 (7.7%)
Sequencing center (%)						
	Baylor	1241 (27.2%)	0 (0)	1103 (72.8%)	138 (6.8%)	0 (0)
	Broad	1263 (27.7%)	2 (0.2%)	286 (18.9%)	974 (47.7%)	1 (7.7%)
	Illumina	797 (17.5%)	26 (2.6%)	10 (0.7%)	750 (36.7%)	11 (84.6%)

		All (N=4,567)	AA* (N=995)	HI* (N=1,516)	EA* (N=2,043)	Other (N=13)
	WashU	1266 (27.7%)	967 (97.2%)	117 (7.7%)	181 (8.9%)	1 (7.7%)
Platform (%)						
	HiSeq X Ten	3227 (70.7%)	965 (97.0%)	1186 (78.2%)	1074 (52.6%)	2 (15.4%)
	HiSeq2000/2500	1340 (29.3%)	30 (3.0%)	330 (21.8%)	969 (47.4%)	11 (84.6%)

*AA: Black/African-American, HI: Hispanic/Latino; EA: White/European ancestry. Populations defined are described in the Methods.

Table 2. Significant variants from single variant association analysis in the pooled ADSP sample within 100kb of the 83 lead GWAS variants

						Pooled Single Variant Association Analysis			Conditional Analysis	
Chr:Pos:A2:A1*	**rsid**	**Gene**	**Location**	**GWAS variants†**	**GWAS loci†**	**MAF**	**Pvalue**	**Beta**	**Pvalue**	**Beta**
6:41067923:C:T	rs115774857	OARD1 (close gene APOBEC2)	intronic	rs143332484, rs75932628, rs10947943	TREM2, UNC5CL	0.0080	1.0E-04	1.016	6.6E-05	1.047
6:41077355:A:G	rs145520578	OARD1, NFYA	intronic			0.0080	9.9E-05	1.018	6.4E-05	1.049
6:41077511:A:C	rs115202236		intronic			0.0081	9.9E-05	1.018	6.4E-05	1.049
6:41082030:G:A	rs12200736		intronic	rs143332484, rs75932628, rs60755019, rs10947943	TREM2, TREML2, UNC5CL	0.0080	1.0E-04	1.018	6.5E-05	1.048
6:41083056:C:T	rs10947945		intronic			0.0080	1.0E-04	1.018	6.4E-05	1.049
6:41088533:T:C	rs12210716		intronic			0.0080	1.0E-04	1.018	6.4E-05	1.049
6:41147490:T:G	rs12199328	intergenic (close gene TREML1)		rs143332484, rs75932628, rs60755019	TREM2, TREML2	0.0080	1.0E-04	1.018	6.5E-05	1.048
6:41173956:G:A‡	rs10947950‡	intergenic	intergenic			0.0128	1.4E-04	0.783	9.1E-05	0.806
7:28034934:T:A	rs73683942	JAZF1	intronic	rs1160871	JAZF1	0.0188	7.5E-05	-0.627	3.3E-03	-0.624

						Pooled Single Variant Association Analysis			Conditional Analysis	
Chr:Pos:A2:A1*	**rsid**	**Gene**	**Location**	**GWAS variants†**	**GWAS loci†**	**MAF**	**Pvalue**	**Beta**	**Pvalue**	**Beta**
7:28034935:C:G	rs78789160		intronic			0.0188	7.5E-05	-0.627	3.3E-03	-0.624
7:28035459:GAGAT:G	no rsids		intronic			0.0188	1.9E-05	-0.677	3.9E-03	-0.674
7:28037452:A:C	rs73683943		intronic			0.0191	3.8E-05	-0.644	3.6E-03	-0.641
7:28042506:C:G	rs60825597		intronic			0.0196	7.8E-05	-0.607	3.3E-03	-0.6
14:52885670:T:TA	rs1310103853	FERMT2	intronic	rs17125924	FERMT2	0.0174	4.5E-06	0.806	3.8E-06	0.817
14:52932032:A:G	rs60609189		intronic			0.0189	7.4E-06	0.768	8.4E-06	0.767
14:52992239:A:G	rs12431954	intergenic				0.0205	3.8E-06	0.763	5.6E-06	0.752
14:92467728:C:T	rs7155002	SLC24A4	intronic	rs12590654, rs7401792	SLC24A4	0.4460	1.4E-05	-0.198	4.6E-03	-0.26
16:31117787:C:G	rs201871085	KAT8 (close gene BCKDK)	missense variant, non coding transcript exon variant	rs889555	BCKDK	0.0108	2.4E-05	0.926	2.9E-05	0.916

*A1 corresponds to the alternate (effect) allele; positions provided are on build 38

† The GWAS variants and GWAS loci are based on the GWAS list from Bellenguez et al (PMID: 35379992)

‡ Variant identified as significant in the multi-population meta-analysis (not in the pooled analysis)

MAF: Minor Allele Frequency

Table 3. Significant variants from single variant association analysis in the population subsamples within 100kb of the 83 lead GWAS variants

	White / European Ancestry					Black / African American					Hispanic / Latino				
Chr:Pos:A2:A1*	**N**	**Freq**	**MAC**	**Pvalue**	**Beta**	**N**	**Freq**	**MAC**	**Pvalue**	**Beta**	**N**	**Freq**	**MAC**	**Pvalue**	**Beta**
4:11052797:T:C	2040	0.0002	1	3.04E-01	2.069	984	0.137	270	8.37E-05	-0.534	1513	0.059	180	4.94E-01	0.113
4:11053332:C:T	2043	0.0002	1	3.04E-01	2.068	995	0.137	273	8.66E-05	-0.530	1516	0.059	179	4.83E-01	0.116
5:180211637:C:T	2042	0.143	584	7.14E-05	0.385	995	0.052	104	8.00E-02	-0.351	1514	0.110	333	9.48E-01	0.009
5:180214978:G:A	2042	0.143	586	5.37E-05	0.392	994	0.051	101	1.45E-01	-0.296	1514	0.109	330	9.39E-01	0.010
5:180216117:C:T	2041	0.143	582	3.69E-05	0.400	995	0.050	100	1.72E-01	-0.278	1512	0.096	289	6.05E-01	0.073
5:180222862:G:A	2042	0.144	590	3.70E-05	0.398	995	0.079	157	3.94E-01	-0.144	1516	0.124	377	9.75E-01	0.004
5:180224704:C:T	2041	0.145	591	2.85E-05	0.403	989	0.081	160	3.81E-01	-0.147	1511	0.124	375	9.48E-01	-0.008
8:11868930:C:G	2043	0.007	27	6.52E-05	1.627	995	0.003	6	5.52E-01	-0.494	1516	0.009	26	3.26E-01	-0.414
11:86068255:T:G	2042	0.810	774	6.19E-01	0.044	992	0.886	226	5.53E-05	0.576	1516	0.811	572	8.58E-01	-0.019
11:86068268:A:G	2042	0.810	774	6.19E-01	0.044	990	0.887	224	2.49E-05	0.607	1515	0.811	572	8.45E-01	-0.021
11:86072833:A:G	2042	0.815	755	6.67E-01	0.038	992	0.888	222	5.37E-05	0.577	1516	0.816	559	7.35E-01	-0.037
11:86186529:G:A	2038	0.142	577	8.10E-02	0.169	993	0.264	525	1.21E-04	-0.403	1516	0.245	742	1.52E-01	0.137
14:105740487:C:T	1863	0.524	1772	2.12E-05	-0.479	981	0.308	605	5.33E-01	0.074	1446	0.430	1244	7.48E-02	0.204
15:63375962:G:A	2042	0.0002	1	2.53E-01	2.318	994	0.101	200	1.18E-04	0.591	1514	0.035	105	9.71E-01	0.008

	White / European Ancestry					Black / African American					Hispanic / Latino				
Chr:Pos:A2:A1*	**N**	**Freq**	**MAC**	**Pvalue**	**Beta**	**N**	**Freq**	**MAC**	**Pvalue**	**Beta**	**N**	**Freq**	**MAC**	**Pvalue**	**Beta**
15:63376200:A:G	2042	0.0002	1	2.53E-01	2.317	994	0.106	210	1.24E-04	0.576	1515	0.040	122	7.96E-01	-0.052
16:31117787:C:G	1917	0.020	76	1.11E-04	0.961	991	0.004	7	1.77E-01	1.033	1489	0.004	12	5.16E-01	0.412
16:86357432:T:C	2043	0.088	359	3.53E-05	-0.497	995	0.132	262	7.55E-01	-0.043	1516	0.097	293	4.62E-01	0.100
17:46724128:T:C	2043	0.007	30	3.58E-05	1.607	995	0.002	3	5.80E-01	-0.643	1515	0.006	17	3.75E-01	0.514
17:46747538:C:T	2043	0.007	28	3.69E-05	1.649	994	0.002	3	5.79E-01	-0.645	1515	0.006	17	3.76E-01	0.513
17:58269710:G:A	2042	0.016	67	7.96E-05	1.069	995	0.004	8	7.23E-01	0.256	1515	0.005	14	7.84E-01	-0.169
20:56407698:G:A	2037	0.004	16	3.43E-05	2.328	991	0.001	2	8.15E-01	0.339	1516	0.002	6	8.86E-01	-0.172
20:56490678:G:T	2041	0.005	19	4.74E-05	2.021	995	0.002	4	4.93E-01	-0.703	1516	0.002	6	8.86E-01	-0.172
20:56505267:G:A	2041	0.005	20	1.57E-05	2.095	995	0.002	4	4.93E-01	-0.703	1513	0.002	6	8.83E-01	-0.177

*A1 corresponds to the alternate (effect) allele; positions provided are on build 38

Variants with a minor allele count (MAC) < 10 in a population subsample were not included in the meta-analysis