RESEARCH ARTICLE OPEN ACCESS

A Fast and Robust Strategy to Remove Variant-Level Artifacts in Alzheimer Disease Sequencing Project Data

Michael E. Belloy, PhD, Yann Le Guen, PhD, Sarah J. Eger, BA, Valerio Napolioni, PhD,
Michael D. Greicius, MD, MPH, and Zihuai He, PhD

Correspondence
Dr. Belloy
mbelloy@stanford.edu

Neurol Genet 2022;8:e200012. doi:10.1212/NXG.0000000000200012

Abstract

Background and Objectives

Exome sequencing (ES) and genome sequencing (GS) are expected to be critical to further elucidate the missing genetic heritability of Alzheimer disease (AD) risk by identifying rare coding and/or noncoding variants that contribute to AD pathogenesis. In the United States, the Alzheimer Disease Sequencing Project (ADSP) has taken a leading role in sequencing AD-related samples at scale, with the resultant data being made publicly available to researchers to generate new insights into the genetic etiology of AD. To achieve sufficient power, the ADSP has adapted a study design where subsets of larger AD cohorts are collected and sequenced across multiple centers, using a variety of sequencing platforms. This approach may lead to variable variant quality across sequencing centers and/or platforms. In this study, we sought to implement and evaluate filters that can be applied fast to robustly remove variant-level artifacts in the ADSP data.

Methods

We implemented a robust quality control procedure to handle ADSP data. We evaluated this procedure while performing exome-wide and genome-wide association analyses on AD risk using the latest ADSP whole ES (WES) and whole GS (WGS) data releases (NG00067.v5).

Results

We observed that many variants displayed large variation in allele frequencies across sequencing centers/platforms and contributed to spurious association signals with AD risk. We also observed that sequencing platform/center adjustment in association models could not fully account for these spurious signals. To address this issue, we designed and implemented variant filters that could capture and remove these center-specific/platform-specific artifactual variants.

Discussion

We derived a fast and robust approach to filter variants that represent sequencing center-related or platform-related artifacts underlying spurious associations with AD risk in ADSP WES and WGS data. This approach will be important to support future robust genetic association studies on ADSP data, as well as other studies with similar designs.

From the Department of Neurology and Neurological Sciences (M.E.B., Y.L.G., S.J.E., M.D.G., Z.H.), Stanford University, CA; Institut du Cerveau—Paris Brain Institute—ICM (Y.L.G.), France; School of Biosciences and Veterinary Medicine (V.N.), University of Camerino, Italy; and Quantitative Sciences Unit (Z.H.), Department of Medicine, Stanford University, CA.

Go to Neurology.org/NG for full disclosures. Funding information is provided at the end of the article.

The Article Processing Charge was funded by 5R01AG06620603.

Glossary

AD = Alzheimer disease; **ADSP** = Alzheimer Disease Sequencing Project; **APP** = amyloid precursor protein; **CHIP** = clonal hematopoiesis of indeterminate potential; **ES** = exome sequencing; **gnomAD** = Genome Aggregation Database; **GS** = genome sequencing; **MC** = Monte Carlo; **PC** = principal component; **QC** = quality control; **SNP** = single-nucleotide polymorphism; **WES** = whole-exome sequencing; **WGS** = whole-genome sequencing.

Late-onset Alzheimer disease (AD) is marked by a strong genetic component, with heritability estimates ranging from 59% to 79%.[1,2] Largely supported by single-nucleotide polymorphism (SNP) genotyping arrays and variant imputation, large-scale meta-analyses of genome-wide association studies have so far implicated more than 50 loci relevant to AD in individuals of European ancestry.[2-6] Despite these important advances, most risk variants identified so far have common allele frequencies, and it is estimated that only approximately half of the genetic heritability of AD has been captured, such that much of the genetic component of AD remains to be identified.[2] In response to this observation, there has been a shift to start using exome sequencing (ES) or genome sequencing (GS) to help capture rare and/or coding variants that contribute to AD risk, which has led to several recent initial successes.[7-15]

In the United States, the Alzheimer Disease Sequencing Project (ADSP) has taken a leading role in the sequencing of AD-related samples at scale, with resultant data being made publicly available to researchers to generate new insights into the genetic etiology of AD. The ADSP has pursued both "whole" ES (WES) and "whole" GS (WGS) approaches (although it should be noted that these for now do not actually provide whole coverage due to technical limitations), where most recently, the focus is increasingly on GS. To achieve sufficient power to support analyses of sequencing data and rare variants, the ADSP has adapted a study design where subsets of larger AD cohorts are collected and sequenced across multiple centers, using a variety of sequencing platforms.[16-18] This in turn can lead to "center" or "platform" effects that traditionally are accounted for by using center/platform covariate adjustment. However, a prior study using a prior version of the ADSP WES discovery phase observed that center/platform covariate adjustment could not account for variable variant qualities across centers and platforms, which in turn may lead to spurious associations or affect the identification of AD-associated risk variants.[19]

Since then, the ADSP has further expanded its efforts and as of 2021, provides their WES and WGS data sets on 20.5 and 16.9 k individuals, respectively, across diverse ancestries.[18] In our exploratory analyses of these data, we observed many variants that displayed large variation in allele frequencies across centers/platforms and contributed to spurious association signals with AD risk, that is, associations that passed at least the common suggestive significance for genome-wide association studies ($p < 1 \times 10^{-5}$) but were of a (likely) artifactual nature. Similar to the prior study,[19] we also observed that platform/center adjustment could not fully account for these signals.

Beyond center/platform adjustment, several strategies have been proposed to handle such artifacts in ES and GS data.[20-22] Notably, preprocessing of UK Biobank SNP array data has previously shown that filters that capture variants displaying large genotype variations across batches/arrays (assessed by the Fisher exact tests) can importantly help remove variants that represent batch or array effects.[23] Because the latter approach is reasonably fast to implement and robust, in this study, we designed and implemented similar filters that aimed to capture and remove center-specific/platform-specific artifactual variants in ASDP WES and WGS data. We additionally tested filters containing putatively artifactual variants identified in the Genome Aggregation Database (gnomAD) reference database.[24] All filters were designed such that they can be implemented post hoc to association analyses, leaving flexibility to researchers to either run full-sample analyses with robust variant quality control (QC) or to identify variants that require targeted analyses. This study summarized the effect of these filters on genome-wide and exome-wide AD association findings in ADSP and proposed they can be used as a fast approach to robustly remove artifactual variants, thereby supporting initial explorations of the ADSP data.

Methods

Ascertainment of Genotype and Phenotype Data

Genotype data for individuals with AD-related clinical outcome measures were available from the ADSP WES and WGS data. Notably, the ADSP performed targeted sequencing of samples in case-control (majority), family-based, population-based, and longitudinal cohorts, performing sequencing across multiple sequencing centers and using various sequencing platforms (eTable 1 and 2, links.lww.com/NXG/A536). Ascertainment of genotype/phenotype data for these samples is described in detail elsewhere.[18,25] In addition to the ADSP samples, we also had access to several publicly available SNP microarray and WGS data sets (eTable 1), largely comprising data from the Alzheimer Disease Genetics Consortium. The latter have a large degree of sample overlap with ADSP. To ensure the most up-to-date and parsimonious phenotypes, we performed a cross-sample genotype/phenotype harmonization, which is summarized in eMethods.

Standard Protocol Approvals, Registrations, and Patient Consents

Participants or their caregivers provided written informed consents in the original studies. This study protocol was granted an exemption by the Stanford Institutional Review

Board because the analyses were conducted on "de-identified, off-the-shelf" data.

Genetic Data QC and Processing

The ADSP WES and WGS data (NG00067.v5) were joint called by the ADSP following the SNP/Indel Variant Calling Pipeline and data management tool used for analysis of GS and ES for the ADSP.[25] The WES data were currently released only for biallelic variants, which the ADSP has quality controlled. The WGS data were released for biallelic and multiallelic variants separately, which the ADSP had not yet quality controlled. The current analyses of ADSP WGS were restricted to biallelic variants, to which we applied the Variant Quality Score Recalibration QC filter (PASS variants; GATK v4.1).[26] The WES/WGS data were available in genome build hg38, which we annotated using dbSNP153 variant identifiers.

Genetic data underwent standard QC. Detailed descriptions of all processing procedures and sequential sample filtering steps are listed in eMethods and eTables 3 and 4 (links.lww.com/NXG/A536). For the purpose of the presented genetic association analyses, only non-Hispanic individuals of European ancestry were considered to focus on the largest ancestry population (SNPweights v2.1; eFigure 1).[27] Principal component (PC) analysis of genotyped SNPs provided PCs capturing population substructure (PC-AiR, eFigure 2).[28] In both the WES and WGS data, variants with a genotyping rate less than 95%, deviating from the Hardy-Weinberg equilibrium in the full sample or in controls ($p < 10^{-6}$), and a minor allele count less than 10 were excluded. After this standard QC, the total number of remaining variants was 224,270 for ADSP WES and 14,772,936 for ADSP WGS.

Primary Filters to Remove Sequencing Center-Related/Platform-Related Variant-Level Artifacts

We designed filters to assess whether there were significant deviations in genotype distributions for any given variant across sequencing centers and platforms. To avoid bias from frequency differences across cases and controls, we assessed only genotypes in control individuals.

The primary filters made use of the fast Fisher exact test as implemented by Plink (v.1.9; command: fisher).[29] However, this test can currently be implemented by comparing only 2 groups at a time (e.g., 2 genotyping centers), while we observed variant issues across multiple groups. We therefore compared every individual sequencing center/platform with all others and combined the p values from the multiple tests through the Cauchy combination test[29] (code available at: github.com/yaowuliu/ACAT). Variants with a combined p value lower than the heuristic threshold of 1×10^{-5} were flagged to be filtered. We note that in this design, there is no need to adjust the p value threshold regarding the number of centers/platforms because the Cauchy combination test inherently accounts for this.

We additionally tested 2 other types of sequencing center-based/platform-based variant filters. On one hand, we performed the χ^2 tests (R v.3.6.0; command: chisq.test) that considered all sequencing centers or platforms at once. Variants with a p value lower than the heuristic threshold of 1×10^{-5} were flagged to be filtered. On the other hand, we performed the Fisher tests with Monte Carlo (MC) simulation of p values (R v.3.6.0; command: fisher.test(simulate.p.value = T)) that considered all sequencing centers or platforms. The MC approach was chosen to allow feasible run times. Variants with a p value lower than the heuristic threshold of 1×10^{-3} were flagged to be filtered (this threshold reflects that the p values from MC simulation are less small than those obtained for the other tests).

The 3 filters were compared for speed by calculating the time needed to derive the respective variant filters on a 1 MB genetic region of chromosome 1 in ADSP WGS. Computing time was evaluated on a single central processing unit from an 80-core Xeon Gold 6138T processor @ 2.00 GHz.

Filters From the gnomAD

In addition to the filters proposed earlier, we used the gnomAD data base (v3.1.1) reference to identify potential variant artifacts.[24] Specifically, we created filters for variants that have the following: (1) a "non-PASS" flag in gnomAD, corresponding to those that did not pass gnomAD sample QC filters and may thus be more prone to sequencing issues; (2) an "LCR" flag in gnomAD, corresponding to those located in a low complexity region and may thus be more prone to low coverage, read misalignment, and subsequent genotype issues; (3) a differential frequency of more than 10% between our current samples and non-Finish European participants in gnomAD, which may indicate an issue with those variants in our samples. The 3 gnomAD filters were evaluated with the goal of supporting the primary ADSP WES/WGS center-based/platform-based variant filters.

Filters for Discordant Variants Across Duplicate Samples

A final set of filters was designed to flag variants that had a discordant genotype across any duplicate sample. Notably, the ADSP WES and WGS data contain a few hundred duplicate samples, generally covering multiple sequencing centers and/or platforms. Discordant variants across such duplicates therefore provide a reference of artifactual variants that should be removed and are largely reflecting center-related/platform-related genotyping issues. We evaluated these filters with the primary goal of comparing them with the primary ADSP WES/WGS center-based/platform-based variant filters, as well as the gnomAD-based variant filters. In a secondary goal, we also assessed to what extent these duplicate discordant variant filters themselves could handle center-related/platform-related variant issues that drove observations of spurious association signals.

Table 1 Sample Demographics

Samples		Diagnosis		Sex	Age	_APOE_ status	
Name	Participants after QC (N)	Type	N	Female, N (%)	Age, mean (SD)	_APOE_*4-pos	_APOE_*2-pos
ADSP WES	11,573	CN	5,418	3,152 (58.2)	85.4 (6.5)	926 (17.1)	1,057 (19.5)
		AD	6,155	3,619 (58.8)	75.4 (8.6)	2,938 (47.7)	493 (8.0)
ADSP WGS	6,533	CN	2,949	1,791 (60.7)	81.6 (6.6)	1,075 (36.4)	204 (6.9)
		AD	3,584	2,051 (57.2)	76.7 (8.3)	2,078 (58.0)	177 (4.9)

Abbreviations: AD, Alzheimer disease; ADSP, Alzheimer Disease Sequencing Project; CN, cognitively normal; QC, quality control; WES, whole-exome sequencing; WGS, whole-genome sequencing.
Samples were restricted to those passing genetic/phenotypic QC, being non-Hispanic, and being of European ancestry.

Statistical Analyses, Variant Annotation, and Visualization

Exome-wide and genome-wide association studies on AD case-control status were conducted on ADSP WES and WGS, respectively, using LMM-BOLT (v.2.3.5). LMM-BOLT uses a Bayesian mixture model that allows the inclusion of related individuals by adjusting for the genetic relationship matrix,[30] thereby maximizing sample size and power. Given the current minor allele count thresholds, the approximate 50-50 ratio of cases to controls and sample sizes exceeding 5,000 participants for both ADSP WES and WGS, the resultant test statistics are expected to be well-calibrated.[30] After analyses, association statistics were transformed back to a logistic scale taking into account the case fraction.[30] Per convention, variants were considered at suggestive ($p \le 1 \times 10^{-5}$) or genome-wide ($p \le 5 \times 10^{-8}$) significance.

Case-control association analyses considered 2 models. Model 1 included covariates for sex, _APOE_*4 dosage, _APOE_*2 dosage, and the first 5 genetic PCs. We did not adjust for age because we previously showed that this can lead to significant power loss when the age of cases is younger than that for controls,[15] which is true for ADSP, given their initial design to prioritize old controls and young cases (Table 1 and eTables 5 and 6, links.lww.com/NXG/A536). Model 2 was the same as model 1 but additionally included covariates for sequencing center and platform. Variant filters were then applied to summary statistics using data.table functions in R v.3.6.0.

The _APOE_ locus (1 Mb region centered on _APOE_) was removed from all summary statistics. Independent loci were determined by sliding window when no variants with $p \le 1 \times 10^{-5}$ were observed within 200Kb from one another. The Manhattan plots provide RefSeq curated gene annotations for the gene closest (<500Kb) to the top significant variant per locus. Only variants with $p \le 1 \times 10^{-6}$ were annotated to improve visualization. Suggestive significance levels were indicated by gray dotted lines and green dots for variants. Genome-wide significance levels were indicated by black solid lines and red dots for variants. Variant densities were indicated at the bottom of the Manhattan plots (dark green = low density, yellow = medium density, and red = high density). Plots were generated using the R package CMplot.[31]

Data Availability

The specific data repository and identifier for each cohort is indicated in eTable 1 (links.lww.com/NXG/A536) of the supplement. Code for the Cauchy combination test is available at: github.com/yaowuliu/ACAT. Summary statistics and variant filters are available on application at: niagads.org/. All data used in the discovery analyses are available on application to the following:

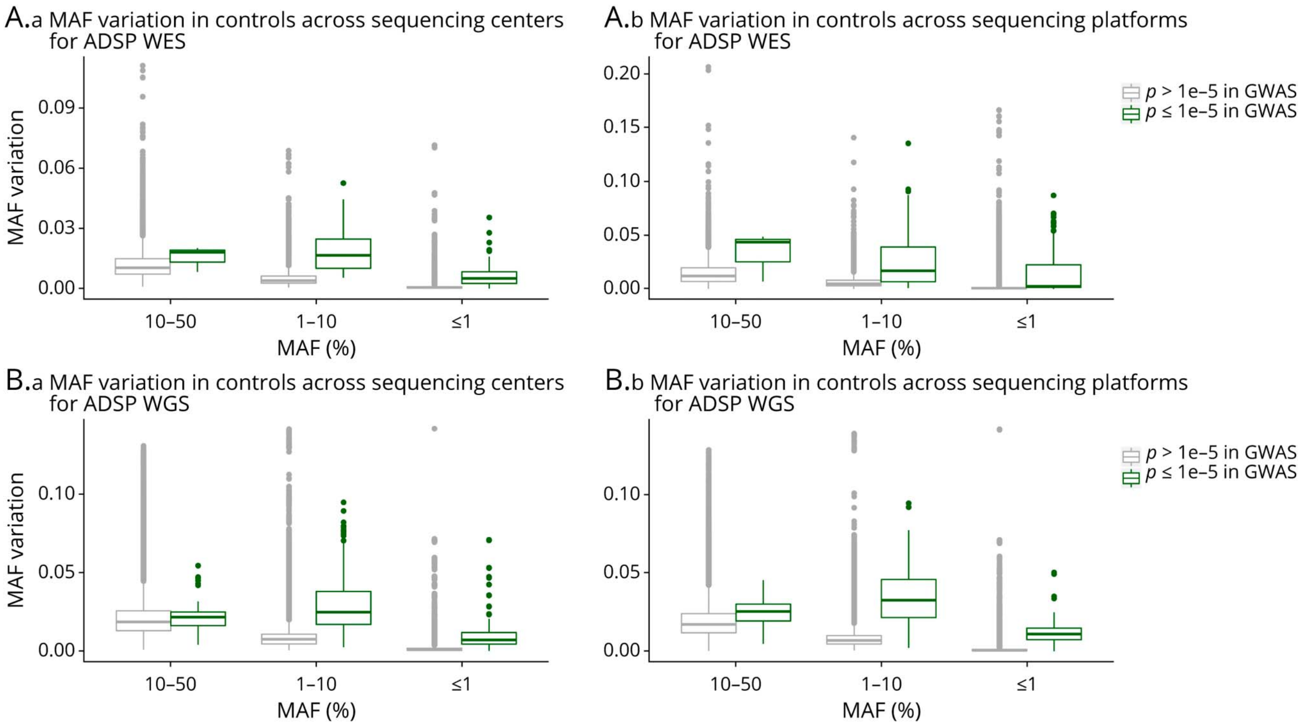
- dbGaP (ncbi.nlm.nih.gov/gap/)
- NIAGADS (niagads.org/)
- LONI (ida.loni.usc.edu/)
- Synapse (synapse.org/)
- Rush (radc.rush.edu/)
- NACC (naccdata.org/).

Results

Sample demographics are summarized in Table 1, with per center/platform demographics in eTables 5 and 6 (links.lww.com/NXG/A536). In initial exome-wide and genome-wide analyses using model 1, we observed many spurious associations ($p \le 1\text{e}{-5}$). We identified that variants underlying these spurious signals displayed increased variation in allele frequency across sequencing centers/platforms for the full frequency range (Figure 1, A and B). We also observed that such variants could not consistently be accounted for by adjustment for sequencing center/platform in model 2; a specific example of such a variant is provided in Figure 1C.

Based on these observations, 3 versions of filters were designed and evaluated for their capacity to capture putative center-related/platform-related variant artifacts. In assessing computing time, the filter using the Fisher exact test implemented in Plink followed by the Cauchy combination of p values implemented in R proved to be the fastest, taking 32 seconds to be constructed using a single central processing unit for a 1 Mb region in ADSP WGS (5,402 variants). Comparatively,

Figure 1 Variant Artifacts Across Different Sequencing Centers/Platforms Drive Spurious Associations in ADSP WES and WGS data



A.a MAF variation in controls across sequencing centers for ADSP WES

A.b MAF variation in controls across sequencing platforms for ADSP WES

B.a MAF variation in controls across sequencing centers for ADSP WGS

B.b MAF variation in controls across sequencing platforms for ADSP WGS

C.a Sequencing center

		ADSP_WES_Baylor	ADSP_WES_Broad	ADSP_WES_CHOP	ADSP_WES_CU_IGM	ADSP_WES_MGI	ADSP_WES_Otogenetics	ADSP_WES_UM_HIHG	ADSP_WES_UW_GenomeSciences	ADSP_WES_WashU
All	WT	2,193	3,590	1	719	746	552	48	19	2,989
	HET	3	411	0	0	0	0	12	0	1
	HOM	0	0	0	0	0	0	0	0	0
CN	WT	1,096	1,473	0	667	356	143	14	0	1,518
	HET	2	80	0	0	0	0	1	0	1
	HOM	0	0	0	0	0	0	0	0	0

C.b Sequencing platform

		Illumina_HiSeq_2000	Illumina_HiSeq_2000/2500	Illumina_HiSeq_2500	Illumina_HiSeq_4000
All	WT	10,342	48	1	466
	HET	415	12	0	0
	HOM	0	0	0	0
CN	WT	4,977	14	0	276
	HET	83	1	0	0
	HOM	0	0	0	0

In initial exome-wide and genome-wide association studies of ADSP WES and WGS, we observed many spurious associations ($p \le 1e{-}5$) using model 1 (i.e., not adjusting for sequencing center/platform; cf. Figures 2A and 3A). On inspection of these signals, it was notable that these variants displayed large variation in genotype counts across sequencing centers/platforms. The MAF variation in controls for all analyzed variants is visualized in (A.a-b) for ADSP WES and in (B.a-b) for ADSP WGS. (C.a-b) A specific example of a variant showing spurious association is provided. This variant, rs199707443, has an MAF of 0.003% in non-Finnish Europeans in Genome Aggregation Database v3.1.1, contrasting the 411 heterozygote counts in the Broad sequencing center. Notably, this particular variant still showed genome-wide significant association with Alzheimer disease risk even after sequencing center/platform adjustment (cf. Figure 2B). ADSP, Alzheimer Disease Sequencing Project; CN, cognitively normal; HET, heterozygote; HOM, homozygote; MAF, minor allele frequency; WT, wild type; WES, whole-exome sequencing; WGS, whole-genome sequencing.

constructing the χ^2 test filter implemented in R took 93 seconds, while the Fisher test with MC filter implemented in R took 128 seconds. Given the faster speed, as well as the expected higher robustness provided by an exact test, we present the filter using the Fisher exact test implemented in Plink as the primary filter, while the other 2 represent supporting analyses. Throughout the remainder of the article, we will use the term "filtered" to describe variants that were removed by filters and the term "non-filtered" to describe variants that were not removed by filters.

The Fisher exact center-based/platform-based variant filters showed they strongly reduced the number of spurious associations observed with model 1 in ADSP WES (Figure 2, A and C) and WGS (Figure 3, A and C). When further adjusting for sequencing center/platform in model

Figure 2 The Proposed Center-Based/Platform-Based Variant Filters Remove Spurious Associations in Alzheimer Disease Sequencing Project Whole-Exome Sequencing

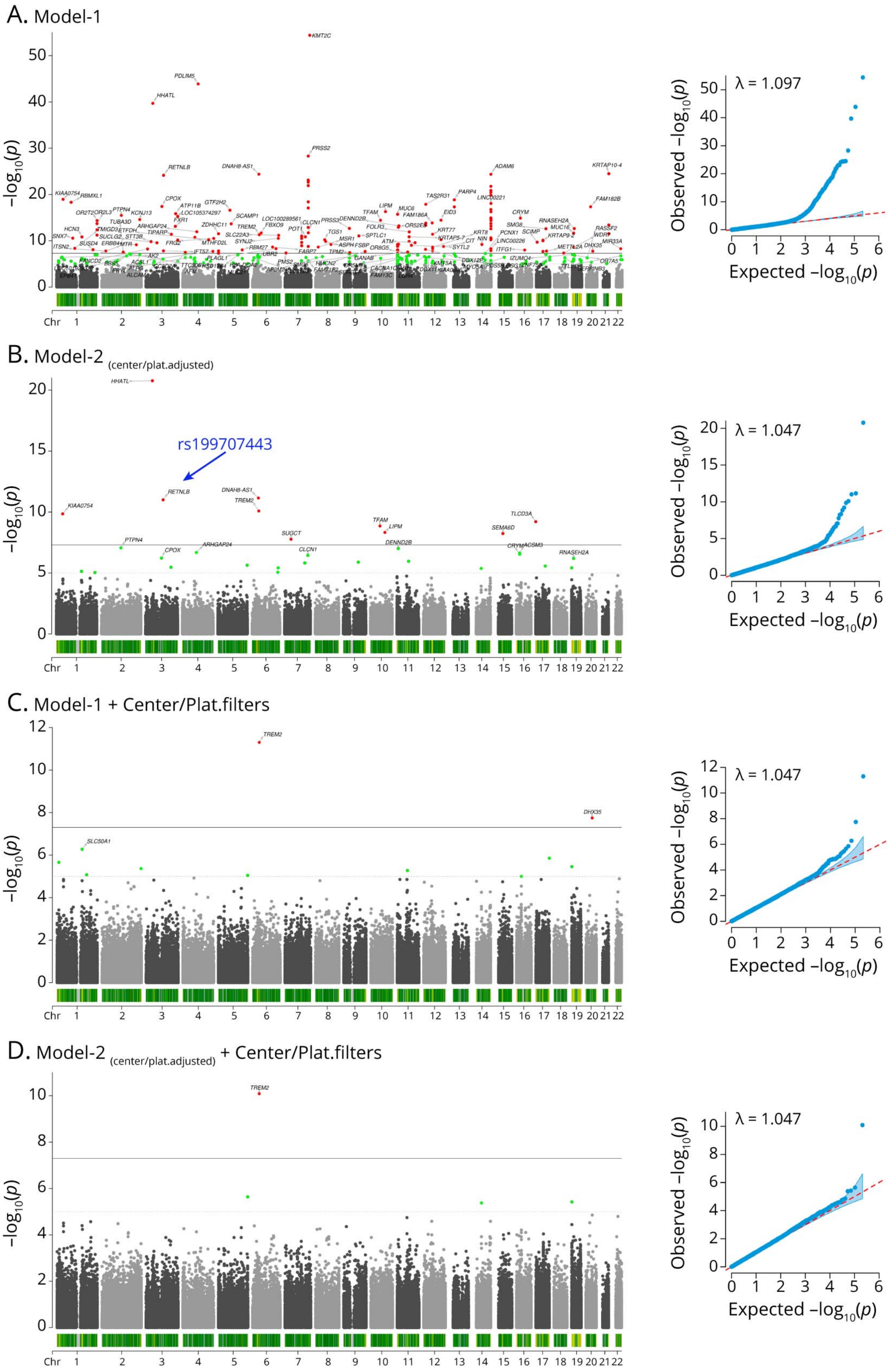


Figure shows the Manhattan (left) and quantile-quantile (right) plots. (A) Model 1 indicates many spurious hits. (B) Model 2 shows that adjustment for center/platform can reduce many, but not all, spurious hits. The variant described in Figure 1C is highlighted by the blue arrow. (C) Filters remove most spurious hits. (D) Further adjustment for center/platform removes few additional spurious hits.

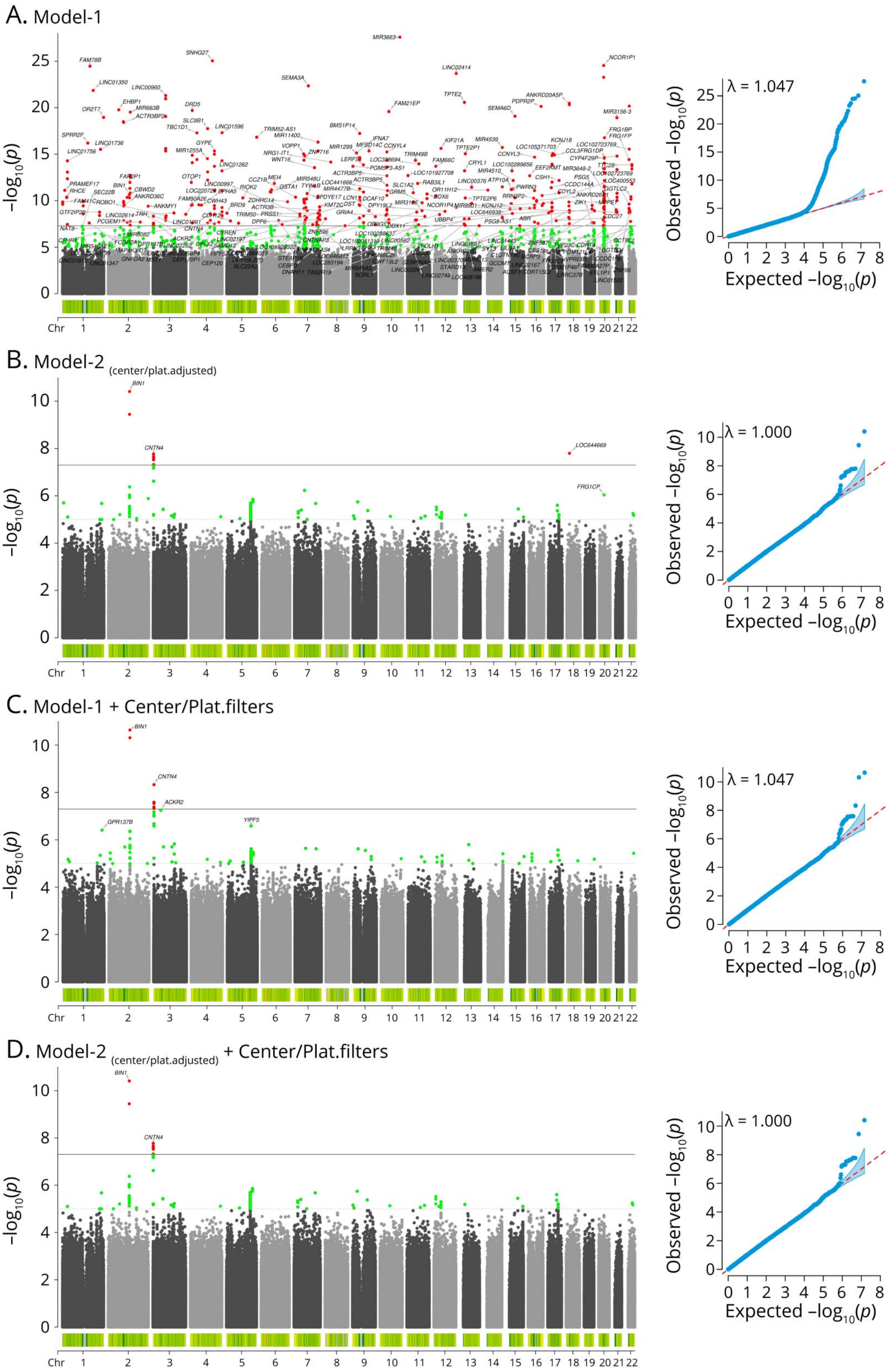
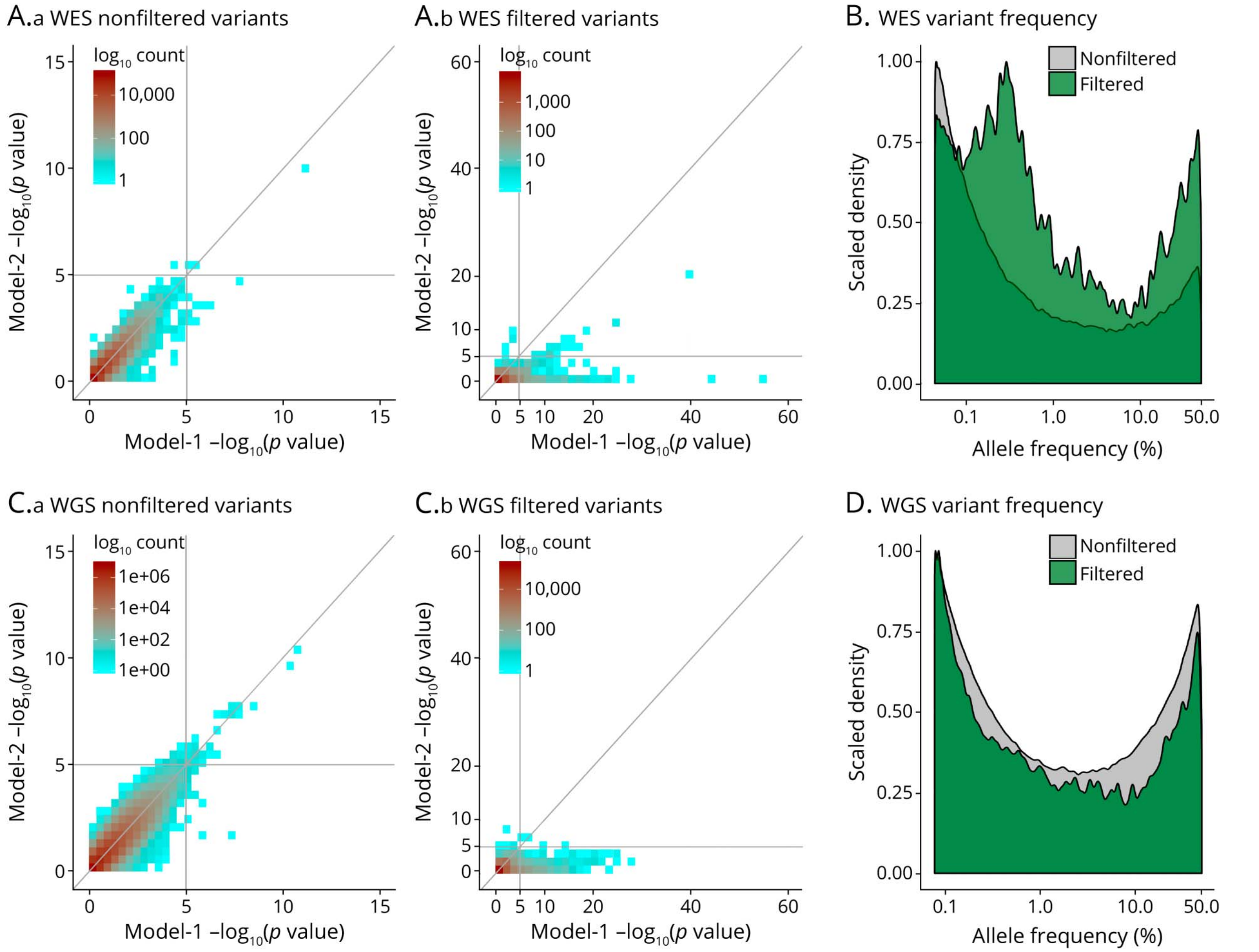


Figure shows the Manhattan (left) and quantile-quantile (right) plots. (A) Model 1 indicates many spurious hits. (B) Model 2 shows that adjustment for center/platform can reduce many, but not all, spurious hits. (C) Filters remove most spurious hits. (D) Further adjustment for center/platform removes few additional spurious hits.

Figure 4 Metrics of Variants Removed by the Proposed Center-Based/Platform-Based Variant Filters



(A.a, A.b, and B) ADSP WES. (C.a, C.b, and D) ADSP WGS. (A.a and C.a) Variants that passed filters showed largely consistent *p* values across model 1 and model 2 case-control association analyses, with only few variants remaining that reach suggestive significance in model 1 but lose suggestive significance on center/platform adjustment in model 2 (lower right quadrant). (A.b and C.b) Variants that were removed by filters showed many inconsistent *p* values across models 1 and 2, consistent with center-related/platform-related variant artifacts that could not fully be accounted for by model 2. (B and D) Frequency density plots, comparing variants that were filtered/removed with those that were not filtered. Note that variants were consistently filtered across the full frequency range, with increased density at frequencies <1% or >10% in ADSP WES. ADSP, Alzheimer Disease Sequencing Project; WES, whole-exome sequencing; WGS, whole-genome sequencing.

2, spurious associations appeared essentially absent in ADSP WES (Figure 2D) and WGS (Figure 3D). Notably, the spurious associations were not detected by the genomic inflation, as for instance, the genomic control factor (λ) was consistent prior to and after applying variant filters in ADSP WGS for the respective models (Figure 3). The slightly larger λ for ADSP WES in model 1 prior to applying the variant filters (Figure 2A) indicated that the large number of spurious variants regarding the relatively small total set of variants was likely driving some modest inflation. Consistent observations were made for the other 2 center-based/platform-based variant filters (eFigures 3-6, links.lww.com/NXG/A536). When intersecting variants identified across these 3 sets of filters, the filter derived from the Fisher exact test implemented in Plink overlapped strongly (>96%) with the other 2 filters that in turn showed less overlap (eFigure 7). This was consistent with the Fisher exact test being the most conservative and robust.

A closer inspection of the center-based/platform-based variant filters showed that nonfiltered variants displayed fairly concordant *p* values across models 1 and 2, whereas filtered variants showed many discrepancies (Figure 4, A and C). This was consistent with the filtered variants driving spurious associations. In addition, it was apparent that filters removed variants across the full frequency range (Figure 4, B and D) consistent with the increased minor allele frequency variation across all frequency ranges for variants underlying spurious association signals (Figure 1, A and B).

We then assessed to what extent the gnomAD-based filters could remove the observed spurious associations. A visual

Table 2 Alzheimer Disease Sequencing Project Whole-Exome Sequencing Variants Passing Suggestive Significance After Applying Center-Based/Platform-Based Filters

Variant info							**Model 1**				**Model 2**				**Filters**			
GENE	**CHR**	**BP**	**dbSNP153 ID**	**Effect allele**	**Other allele**	**Effect allele frequency (%)**	**OR**	**95% CI (lb)**	**95% CI (ub)**	***p* Value**	**OR**	**95% CI (lb)**	**95% CI (ub)**	***p* Value**	**Center Fisher P**	**Platf. Fisher P**	**gnomAD filter**	**Duplicate check**
DRAXIN	1	11,709,400	rs769650621	C	A	0.45	2.44	1.69	3.53	**2.2E−06**	1.83	1.28	2.63	1.0E−03	2.2E−04	0.97	Non-PASS	Discordant
SLC50A1	1	155,136,277	rs765315443	C	T	0.37	2.82	1.88	4.22	**5.3E−07**	2.06	1.38	3.05	3.6E−04	2.3E−05	0.99	PASS	Ok
LAMC1-AS1	1	183,135,182	rs1385675950	A	C	0.18	3.69	2.08	6.56	**8.4E−06**	2.70	1.54	4.72	5.2E−04	0.10	0.99	PASS	Discordant
LOC150935	2	239,780,397	rs1355381797	C	A	0.15	4.39	2.34	8.25	**4.3E−06**	3.17	1.71	5.86	2.4E−04	0.99	0.99	PASS	Ok
RASGEF1C	5	180,127,602	rs57288534	T	C	18.67	0.87	0.81	0.92	**9.0E−06**	0.86	0.81	0.92	**2.3E−06**	0.76	0.34	PASS	Ok
TREM2	6	41,161,514	rs75932628	T	C	0.69	2.82	2.10	3.79	**5.0E−12**	2.58	1.94	3.44	**8.2E−11**	0.02	0.09	PASS	Ok
HNRNPUL2-BSCL2	11	62,724,366	rs772898628	C	A	0.24	3.14	1.92	5.13	**5.3E−06**	2.21	1.37	3.57	1.2E−03	6.1E−03	0.98	Non-PASS	Discordant
CDKL1	14	50,390,164	rs61981931	T	C	4.82	0.79	0.71	0.89	5.6E−05	0.77	0.69	0.86	**4.2E−06**	1.1E−05	0.02	PASS	Ok
C16orf92	16	30,025,807	rs11544328	C	A	46.34	0.89	0.85	0.94	**1.0E−05**	0.91	0.87	0.95	8.0E−05	0.83	1.7E−03	PASS	Discordant
ZNF750	17	82,831,739	rs751362098	G	A	0.39	2.61	1.77	3.85	**1.4E−06**	2.02	1.38	2.96	2.8E−04	1.5E−04	0.99	Non-PASS	Discordant
ABCA7	19	1,042,810	rs3764645	G	A	46.71	0.89	0.85	0.94	**3.5E−06**	0.89	0.85	0.94	**3.8E−06**	0.22	3.9E−04	PASS	Ok
DHX35	20	39,018,815	rs779184241	A	G	0.32	3.49	2.26	5.38	**1.8E−08**	2.56	1.67	3.91	1.4E−05	2.5E−04	0.99	PASS	Discordant

Abbreviations: BP, base pair; CHR, chromosome; CI, confidence interval; gnomAD, Genome Aggregation Database; OR, odds ratio; PASS, flag indicating variant passed gnomAD quality control.
Variants shown passed suggestive significance in either model 1 or model 2. Note that many variants that lose suggestive significance after center/platform adjustment in model 2 have fairly small *p* values (but above threshold) in the center/platform Fisher tests and/or have a non-PASS flag in gnomAD or are flagged by the duplicate discordant variant filter. This suggests there is added value in using model 2 and/or applying the gnomAD and duplicate discordant variant filters to reduce spurious signals or that model 1 without gnomAD filters can be used contingent on post hoc assessment of the association signal's robustness. Bolded entries indicate $p \leq 1 \times 10^{-5}$.

assessment of the Manhattan plots showed that the gnomAD-based filters could only account for a part of the spurious associations (eFigures 8 and 9, links.lww.com/NXG/A536). Similarly, a closer inspection of the gnomAD-based filters showed that they mainly removed variants with frequencies <1% (eFigure 10). The *p* values across models 1 and 2 further showed many discrepancies for both nonfiltered and filtered variants (although fewer for nonfiltered variants). In sum, the gnomAD-based filters could remove some spurious signals but were less effective than the center-based/platform-based variant filters.

We further assessed to what extent the duplicate discordant variants filters could remove the observed spurious associations. The Manhattan plots showed that the duplicate discordant variant filters could account for many of the spurious associations, but several remained when using model 1, while when using model 2, the Manhattan plots looked similar to those using the center-based/platform-based variant filters (eFigures 11 and 12, links.lww.com/NXG/A536). A closer inspection of the duplicate discordant variant filters similarly showed they mainly removed variants with frequencies >10% and did not remove a set of variants that lose suggestive significance when going from model 1 to model 2 (eFigure 13). An illustrative example of such a variant is listed in eTable 7, confirming these variants represent genotyping issues that more ideally should be removed from the data. In sum, the duplicate discordant filters could remove many spurious signals but were less effective than the center-based/platform-based variant filters, yet more effective than the gnomAD-based variant filters.

We also sought to understand the overlap between the different proposed filters. The 3 gnomAD-based variant filters appeared to show little overlap with one another (eFigure 14, links.lww.com/NXG/A536) and overlapped with less than 20% of the variants in the center-based/platform-based variant filters (eFigure 15). Furthermore, in ADSP WES and WGS, 32% and 14% of duplicate discordant variants overlapped center-based/platform-based variant filters, respectively, while vice versa 31% and 15% of center-based/platform-based filtered variants overlapped duplicate discordant variants (eFigure 16). In the same comparison, 28% and 49% of duplicate discordant variants overlapped gnomAD-based variant filters, respectively, while vice versa 53% and 17% of gnomAD-based filtered variants overlapped duplicate discordant variants (eFigure 17). In sum, this confirmed that all 3 types of filters captured overlapping as well as unique variants. Notably, the center-based/platform-based and gnomAD-based variant filters could capture a subset of reference artifactual variants present in the duplicate discordant variant filters but identified many additional signals that represented likely artifactual variants and contributed to spurious association signals. An overview of the number of variants and spurious signals removed for all respective filters and models is summarized in eTable 8.

Then, we sought to assess whether the use of these different types of variant filters could omit the need for adjusting for sequencing center/platform as implemented in model 2, which may be desirable for certain studies or research questions. We thus inspected all variants that passed suggestive significance in either model 1 or model 2 in ADSP WES (Table 2) and WGS (eTable 9, links.lww.com/NXG/A536) after applying the center-based/platform-based filters (which we showed removed the most spurious signals). We observed that many variants that lose suggestive significance after center/platform adjustment in model 2 have fairly small (above threshold) *p* values in the center-based/platform-based Fisher exact tests and/or are covered in the gnomAD-based and duplicate discordant variant filters. Similarly, assessing the Manhattan plots and variant metrics suggested that the gnomAD-based and/or duplicate discordant variant filters removed few additional variants underlying spurious signals (eFigures 18-23). Notably, we also observed in ADSP WGS that center/platform adjustment for some variants led to somewhat more significant *p* values, thereby increasing the number of suggestive hits (eTables 8 and 9). This could reflect improved model fits after center/platform adjustment by accounting for case/control imbalances or other factors. Overall, these observations suggest there may be added value in using model 2 and/or applying the gnomAD-based filters to reduce spurious signals. Obviously, adding the duplicate discordant variant filters will inherently remove artifactual signals and help reduce spurious signals.

Last, as a robustness check, we compared association statistics from the current ADSP WES analyses with variants that we identified in a prior study using a prior version of the ADSP WES data and observed highly concordant findings (eTable 10, links.lww.com/NXG/A536).[15]

Discussion

We present a fast and robust approach to filter variants that represent sequencing center-related or platform-related artifacts underlying spurious associations with AD risk in ADSP WES and WGS data, which cannot fully be accounted for by center/platform covariate adjustment. In addition, we showed that filters comprising variants that may be prone to artifacts, as identified by gnomAD, were less efficient in removing spurious signals but may still have added value on top of the center-based/platform-based filters. Similarly, filters containing variants that were discordant across duplicate samples could remove many, but not all, spurious signals and added onto the center-based/platform-based filters. In sum, the presented filters are important to support future robust studies on ADSP data. In addition, these filters allow flexibility, given that they can be applied in post hoc QC. Researchers may thus inspect filtered variants in targeted analyses in subsets of the ADSP data where no artifactual genotype enrichment is observed (e.g., excluding a single sequencing center/platform that showed an artifactual increase in genotype counts compared with the others, cf. Wickland et al.[19]).

Certain study designs or research questions may benefit from not adjusting by sequencing center/platform (i.e., cohort adjustment). For example, a study that considers specific strata and/or low-frequency variants may observe some colinearity between variant genotype observations and sequencing centers/platforms. However, this does not necessarily indicate artifactual variants and may be driven by chance or variable cohort study designs across samples sequenced by different centers. We observed that the presented center-based/platform-based variant filters could handle nearly all spurious associations when not adjusting for sequencing center/platform in model 1. Inspecting the remaining signals passing suggestive significance, it was apparent that the gnomAD-based and duplicate discordant variant filters could remove a few additional spurious signals. Similarly, the *p* values from the Fisher exact tests across sequencing centers/platforms was fairly small for several variants that passed suggestive significance in their association with AD risk in model 1 but lost suggestive significance on center/platform adjustment in model 2. In sum, we suggest that model 2 with application of center-based/platform-based, gnomAD-based, and duplicate discordant variant filters is the most conservative approach, but model 1 using only center-based/platform-based and duplicate discordant variant filters may reasonably be implemented, contingent on post hoc assessment of the association signals' robustness.

The center-based/platform-based filtering approach will further be valuable beyond the currently presented exome-wide and genome-wide univariate AD risk association analyses in European ancestry samples. Notably, the removal of artifactual variants may lead to improved association statistics in gene-based testing, which is particularly relevant for ES/GS data.[7] The filter approach can also be applied to non-European samples available in ADSP WES/WGS. Last, the approach to check for variant artifacts by comparing genotype distributions across sequencing centers/platforms may also be used in other studies with a similar design as the current ADSP data. Notably, our approach is similar to the one previously applied to the preprocessing of UK Biobank SNP array to remove variants that may represent batch or array effects.[23] In turn, the approach described here and applied to ES/GS data could also be applied to the large amount of SNP array data sets used in large-scale genetic studies of AD.[3]

This study reports exome-wide and genome-wide AD risk association findings for the newly released ADSP 20.5k WES and 16.9k WGS data. After QC and filter implementation, we observed few signals passing the genome-wide significance threshold. In the ADSP WES data, *TREM2* and *ABCA7*—well-established AD risk genes[2,6]—were observed with variants, respectively, at genome-wide and suggestive significance, consistent with observations for similar models in prior studies on the prior ADSP WES discovery phase data.[7,15] Despite observing only 4 variants in ADSP WES that passed suggestive significance in model 2, our findings were overall highly consistent with prior work.[15] We also observed that certain variants identified previously were not present in our current summary statistics (eTable 10, links.lww.com/NXG/A536), reflecting differences in joint calling, QC, and the fact that currently only biallelic data were available for the new ADSP WES data. Notably, the common protective variant on *ABCA7* identified here has not been previously reported (and we confirm it appears to not have been successfully joint called in the prior ADSP WES data; dbGaP accession ID: phs000572). In the ADSP WGS data, in addition to several suggestive hits, *BIN1*—a well-established AD risk gene[2,6]—and *CNTN4* were identified with variants at genome-wide significance. The common protective variant on *CNTN4* appears novel and may be of relevance to AD pathogenesis given that Contactin 4 (CNTN4) is a binding partner of amyloid precursor protein (APP) and CNTN4/APP interaction may play a role in promoting target-specific axon arborization.[32,33] Overall, these initial findings appear promising but suggest that the current ADSP WES/WGS data may still face power limitations limiting the discovery of novel risk variants. As such, gene-based testing, analyses on available non-European ancestry samples, and novel methodological approaches to gain additional power[12,15] will all be crucial to support future advances into disentangling the missing heritability of AD using ADSP samples and other complimentary large-scale sequencing data.

One limitation to the proposed center-based/platform-based and gnomAD-based filters is that, while they robustly remove many artifactual variants, they may potentially remove non-artifactual variants (i.e., false negatives) and thus reduce power or still miss other artifactual variants (i.e., true positives). Theoretically, filtered and nonfiltered variants could be verified for their association with AD in the summary statistics from other large-scale genome-wide association studies using imputed SNP data, but this inherently comes with concerns regarding imputation/genotype quality in those cohorts, as well as challenges to resolve signals below the suggestive significance threshold in ADSP (given its relatively limited sample sizes). As such, a clear assessment of sensitivity and specificity is not directly feasible at the current time. Some false positives may be expected in ADSP WES owning to the fact that the ADSP used a variety of exome capture kits, which were not considered here, because those metadata were not readily available at the current time. Additional false positives may also still be expected for any remaining variants with allele imbalance, which was not assessed in this study.[34] Furthermore, other factors such as imbalance of ancestry, case/control ratios, or age across centers may affect the variant filters and lead to false negatives. However, in data not shown in this study, consistent spurious associations were observed and removed by filters when considering a more homogenous population of North-western European or African ancestry individuals, suggesting ancestry imbalance did not specifically bias the center/platform effects. Similarly, by designing the center/platform filters on controls, there was little concern regarding case/control ratio and age imbalance. However, cohort study design differences may cause control individuals on certain centers/platforms to be enriched in protective variants (e.g., if a given study specifically recruited protected old age *APOE**4 carriers), which could potentially contribute to false negatives. In addition, age in general may represent a confounding factor because clonal hematopoiesis of indeterminate potential (CHIP) contributes to an increased rate of somatic mutations with aging that can confound analyses (particularly in CHIP-associated genes).[35] This may be specifically relevant when the genetic association model does not account for age, as was the case in this study. Last, the gnomAD filters flag variants that were artifactual in gnomAD and are thus prone to technical issues, but not all these variants are necessarily artifactual in the current ADSP data. Future studies may further also consider adapting the gnomAD 10% differential frequency filter to instead make use of a Fisher test, similar as in the primary center-based/platform-based filters. In sum, while the current filters are clearly useful to increase the robustness of association finding in ADSP data, future studies may further implement and evaluate other approaches to handle artifactual variants while validating sensitivity and specificity. Future studies may also consider inspecting target variants or genes without applying the filters proposed here but instead using them as a reference or adapting them, as appropriate.

We present a fast and robust approach to filter variants that represent sequencing center-related or platform-related artifacts underlying spurious associations with AD risk in ADSP WES and WGS data. This approach will be important to support future robust studies on ADSP data, as well as other studies with similar designs.

Acknowledgment

Biological samples used in this study were stored at study investigators' institutions and at the National Cell Repository for Alzheimer Disease (NCRAD) at Indiana University, which receives government support under a cooperative agreement grant (U24 AG21886) awarded by the National Institute on Aging (NIA). The authors thank contributors who collected samples used in this study, as well as patients and their families, whose help and participation made this work possible. Phenotypic data were provided by principal investigators, the NIA-funded Alzheimer Disease Centers (ADCs), the National Alzheimer Coordinating Center (NACC, U01AG016976), and the National Institute on Aging Genetics of Alzheimer Disease Data Storage Site (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by the NIA. Contributors to the Genetic Analysis Data included study investigators on projects that were individually funded by the NIA, and other NIH institutes, and by private US organizations or foreign governmental or nongovernmental organizations. Data for this study were prepared, archived, and distributed by the National Institute on Aging Alzheimer Disease Data Storage Site (NIAGADS) at the University of Pennsylvania (U24-AG041689-01); Alzheimer Disease Genetics Consortium (ADGC), U01 AG032984, RC2 AG036528; NACC, U01 AG016976; NIA-LOAD (Columbia University), U24 AG026395, U24 AG026390, R01AG041797; Banner Sun Health Research Institute P30 AG019610; Boston University, P30 AG013846, U01 AG10483, R01 CA129769, R01 MH080295, R01 AG017173, R01 AG025259, R01 AG048927, R01AG33193, R01 AG009029; Columbia University, P50 AG008702, R37 AG015473, R01 AG037212, R01 AG028786; Duke University, P30 AG028377, AG05128; Emory University, AG025688; Group Health Research Institute, UO1 AG006781, UO1 HG004610, UO1 HG006375, U01 HG008657; Indiana University, P30 AG10133, R01 AG009956, RC2 AG036650; Johns Hopkins University, P50 AG005146, R01 AG020688; Massachusetts General Hospital, P50 AG005134; Mayo Clinic, P50 AG016574, R01 AG032990, KL2 RR024151; Mount Sinai School of Medicine, P50 AG005138, P01 AG002219; New York University, P30 AG08051, UL1 RR029893, 5R01AG012101, 5R01AG022374, 5R01AG013616, 1RC2AG036502, 1R01AG035137; North Carolina A&T University, P20 MD000546, R01 AG28786-01A1; Northwestern University, P30 AG013854; Oregon Health & Science University, P30 AG008017, R01 AG026916; Rush University, P30 AG010161, R01 AG019085, R01 AG15819, R01 AG17917, R01 AG030146, R01 AG01101, RC2 AG036650, R01 AG22018; 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 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, P30 AG053760 and AG063760; University of Pennsylvania, P30 AG010124; University of Pittsburgh, P50 AG005133, AG030653, AG041718, AG07562, AG02365; 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, R01 AG042437; University of Wisconsin, P50 AG033514; Vanderbilt University, R01 AG019085; and Washington University, P50 AG005681, P01 AG03991, P01 AG026276. The Kathleen Price Bryan Brain Bank at Duke University Medical Center is funded by NINDS grant no. 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 AG041232, The Banner Alzheimer Foundation, The Johnnie B. Byrd Sr. Alzheimer Institute, the Medical Research Council, and the state of Arizona and includes samples from the following sites: Newcastle Brain Tissue Resource (funding through the Medical Research Council, local NHS trusts, and Newcastle University), MRC London Brain Bank for Neurodegenerative Diseases (funding through the Medical Research Council), South West Dementia Brain Bank (funding through numerous sources including the Higher Education Funding Council for England [HEFCE], Alzheimer Research Trust [ART], BRACE, as well as

North Bristol NHS Trust Research and Innovation 58 Department and DeNDRoN), The Netherlands Brain Bank (funding through 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. The NACC database is funded by NIA/NIH Grant U01 AG016976. NACC data are contributed by the NIA-funded ADCs: P30 AG019610 (PI Eric Reiman, MD), P30 AG013846 (PI Neil Kowall, MD), P30 AG062428-01 (PI James Leverenz, MD) P50 AG008702 (PI Scott Small, MD), P50 AG025688 (PI Allan Levey, MD, PhD), P50 AG047266 (PI Todd Golde, MD, PhD), P30 AG010133 (PI Andrew Saykin, PsyD), P50 AG005146 (PI Marilyn Albert, PhD), P30 AG062421-01 (PI Bradley Hyman, MD, PhD), P30 AG062422-01 (PI Ronald Petersen, MD, PhD), P50 AG005138 (PI Mary Sano, PhD), P30 AG008051 (PI Thomas Wisniewski, MD), P30 AG013854 (PI Robert Vassar, PhD), P30 AG008017 (PI Jeffrey Kaye, MD), P30 AG010161 (PI David Bennett, MD), P50 AG047366 (PI Victor Henderson, MD, MS), P30 AG010129 (PI Charles DeCarli, MD), P50 AG016573 (PI Frank LaFerla, PhD), P30 AG062429-01(PI James Brewer, MD, PhD), P50 AG023501 (PI Bruce Miller, MD), P30 AG035982 (PI Russell Swerdlow, MD), P30 AG028383 (PI Linda Van Eldik, PhD), P30 AG053760 (PI Henry Paulson, MD, PhD), P30 AG010124 (PI John Trojanowski, MD, PhD), P50 AG005133 (PI Oscar Lopez, MD), P50 AG005142 (PI Helena Chui, MD), P30 AG012300 (PI Roger Rosenberg, MD), P30 AG049638 (PI Suzanne Craft, PhD), P50 AG005136 (PI Thomas Grabowski, MD), P30 AG062715-01 (PI Sanjay Asthana, MD, FRCP), P50 AG005681 (PI John Morris, MD), and P50 AG047270 (PI Stephen Strittmatter, MD, PhD). The genotypic and associated phenotypic data used in the study "Multi-Site Collaborative Study for Genotype-Phenotype Associations in Alzheimer's Disease (GenADA)" were provided by the GlaxoSmithKline, R&D Limited. ROSMAP study data were provided by the Rush Alzheimer Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161, R01AG15819, R01AG17917, R01AG30146, R01AG36836, U01AG32984, U01AG46152, the Illinois Department of Public Health, and the Translational Genomics Research Institute. The AddNeuroMed data are from a public-private partnership supported by EFPIA companies and SMEs as part of InnoMed (Innovative Medicines in Europe), an Integrated Project funded by the European Union of the Sixth Framework program priority FP6-2004-LIFESCIHEALTH-5. Clinical leads responsible for data collection are Iwona Kłoszewska (Lodz), Simon Lovestone (London), Patrizia Mecocci (Perugia), Hilkka Soininen (Kuopio), Magda Tsolaki (Thessaloniki), and Bruno Vellas (Toulouse), imaging leads are Andy Simmons (London), Lars-Olad Wahlund (Stockholm), and Christian Spenger (Zurich), and bioinformatics leads are Richard Dobson (London) and Stephen Newhouse (London). Data collection and sharing for this project was funded by the Alzheimer 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 Association; Alzheimer 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 HealtControlsare; 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. The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of Southern California. The Alzheimer Disease Sequencing Project (ADSP) comprises 2 Alzheimer Disease (AD) genetics consortia and 3 National Human Genome Research Institute (NHGRI)–funded Large-Scale Sequencing and Analysis Centers (LSAC). The 2 AD genetics consortia are the Alzheimer Disease Genetics Consortium (ADGC) funded by NIA (U01 AG032984) and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) funded by NIA (R01 AG033193), the National Heart, Lung, and Blood Institute (NHLBI), other National Institute of Health (NIH) institutes, and other foreign governmental and nongovernmental organizations. The Discovery Phase analysis of sequence data is supported through UF1AG047133 (to Drs. Schellenberg, Farrer, Pericak-Vance, Mayeux, and Haines); U01AG049505 to Dr. Seshadri; U01AG049506 to Dr. Boerwinkle; U01AG049507 to Dr. Wijsman; and U01AG049508 to Dr. Goate, and the Discovery Extension Phase analysis is supported through U01AG052411 to Dr. Goate, U01AG052410 to Dr. Pericak-Vance, and U01 AG052409 to Drs. Seshadri and Fornage. The ADGC cohorts included in ADSP include the following: Adult Changes in Thought (ACT; UO1 AG006781, UO1 HG004610, UO1 HG006375, U01 HG008657), the Alzheimer Disease Centers (ADC; P30 AG019610, P30

AG013846, P50 AG008702, P50 AG025688, P50 AG047266, P30 AG010133, P50 AG005146, P50 AG005134, P50 AG016574, P50 AG005138, P30 AG008051, P30 AG013854, P30 AG008017, P30 AG010161, P50 AG047366, P30 AG010129, P50 AG016573, P50 AG016570, P50 AG005131, P50 AG023501, P30 AG035982, P30 AG028383, P30 AG010124, P50 AG005133, P50 AG005142, P30 AG012300, P50 AG005136, P50 AG033514, P50 AG005681, and P50 AG047270), the Chicago Health and Aging Project (CHAP; R01 AG11101, RC4 AG039085, and K23 AG030944), Indianapolis Ibadan (R01 AG009956 and P30 AG010133), the Memory and Aging Project (MAP; R01 AG17917), Mayo Clinic (MAYO) (R01 AG032990, U01 AG046139, R01 NS080820, RF1 AG051504, and P50 AG016574), Mayo Parkinson Disease controls (NS039764, NS071674, and 5RC2HG005605), University of Miami (R01 AG027944, R01 AG028786, R01 AG019085, IIRG09133827, and A2011048), the Multi-Institutional Research in Alzheimer Genetic Epidemiology Study (MIRAGE; R01 AG09029 and R01 AG025259), the National Cell Repository for Alzheimer's Disease (NCRAD; U24 AG21886), the National Institute on Aging Late-Onset Alzheimer Disease Family Study (NIA-LOAD; R01 AG041797), the Religious Orders Study (ROS; P30 AG10161 and R01 AG15819), the Texas Alzheimer Research and Care Consortium (TARCC; funded by the Darrell K Royal Texas Alzheimer Initiative), Vanderbilt University/Case Western Reserve University (VAN/CWRU; R01 AG019757, R01 AG021547, R01 AG027944, R01 AG028786, P01 NS026630, and Alzheimer Association), the Washington Heights-Inwood Columbia Aging Project (WHICAP; RF1 AG054023), the University of Washington Families (VA Research Merit Grant, NIA: P50AG005136, R01AG041797, NINDS: R01NS069719), the Columbia University Hispanic Estudio Familiar de Influencia Genetica de Alzheimer (EFIGA; RF1 AG015473), the University of Toronto (UT; funded by Wellcome Trust, Medical Research Council, Canadian Institutes of Health Research), and Genetic Differences (GD; R01 AG007584). The CHARGE cohorts are supported in part by the National Heart, Lung, and Blood Institute (NHLBI) infrastructure grant HL105756 (Psaty), RC2HL102419 (Boerwinkle), and the neurology working group is supported by the National Institute on Aging (NIA) R01 grant AG033193.The CHARGE cohorts participating in the ADSP include the following: Austrian Stroke Prevention Study (ASPS), ASPS-Family study, and the Prospective Dementia Registry-Austria (ASPS/PRODEM-Aus), the Atherosclerosis Risk in Communities (ARIC) Study, the Cardiovascular Health Study (CHS), the Erasmus Rucphen Family Study (ERF), the Framingham Heart Study (FHS), and the Rotterdam Study (RS). ASPS is funded by the Austrian Science Fond (FWF) grant number P20545-P05 and P13180 and the Medical University of Graz. The ASPS-Fam is funded by the Austrian Science Fund (FWF) project I904, the EU Joint Programme—Neurodegenerative Disease Research (JPND) in frame of the BRIDGET project (Austria, Ministry of Science), and the Medical University of Graz and the Steiermärkische Krankenanstalten Gesellschaft. PRODEM-Austria is supported by the Austrian Research Promotion agency (FFG) (Project No. 827462) and by the Austrian National Bank (Anniversary Fund, project 15435. ARIC research is conducted as a collaborative study supported by NHLBI contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). Neurocognitive data in ARIC are collected by U01 2U01HL096812, 2U01HL096814, 2U01HL096899, 2U01HL096902, and 2U01HL096917 from the NIH (the NHLBI, NINDS, NIA, and NIDCD) and with previous brain MRI examinations funded by R01-HL70825 from the NHLBI. CHS research was supported by contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the NHLBI with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629, R01AG15928, and R01AG20098 from the NIA. FHS research is supported by NHLBI contracts N01-HC-25195 and HHSN268201500001I. This study was also supported by additional grants from the NIA (R01s AG054076, AG049607, and AG033040 and NINDS (R01 NS017950). The ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and received funding from the European Community's Seventh Framework Program (FP7/2007–2013)/grant agreement HEALTH-F4-2007–201413 by the European Commission under the program "Quality of Life and Management of the Living Resources" of 5th Framework Program (no. QLG2-CT-2002-01254). A high-throughput analysis of the ERF data was supported by a joint grant from the Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043). The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, the Netherlands Organization for Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the municipality of Rotterdam. Genetic data sets are also supported by the Netherlands Organization of Scientific Research NWO Investments (175.010.2005.011, 911-03-012), the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), and the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA), project 050-060-810. All studies are grateful to their participants, faculty, and staff. The content of these studies is solely the responsibility of the authors and does not necessarily represent the official views of the

National Institutes of Health or the US Department of Health and Human Services. The FUS cohorts include the following: the Alzheimer Disease Centers (ADC) (P30 AG019610, P30 AG013846, P50 AG008702, P50 AG025688, P50 AG047266, P30 AG010133, P50 AG005146, P50 AG005134, P50 AG016574, P50 AG005138, P30 AG008051, P30 AG013854, P30 AG008017, P30 AG010161, P50 AG047366, P30 AG010129, P50 AG016573, P50 AG016570, P50 AG005131, P50 AG023501, P30 AG035982, P30 AG028383, P30 AG010124, P50 AG005133, P50 AG005142, P30 AG012300, P50 AG005136, P50 AG033514, P50 AG005681, and P50 AG047270), Alzheimer Disease Neuroimaging Initiative (ADNI) (U19AG024904), Amish Protective Variant Study (RF1AG058066), Cache County Study (R01AG11380, R01AG031272, R01AG21136, RF1AG054052), Case Western Reserve University Brain Bank (CWRUBB) (P50AG008012), Case Western Reserve University Rapid Decline (CWRURD) (RF1AG058267, NU38CK000480), CubanAmerican Alzheimer Disease Initiative (CuAADI) (3U01AG052410), Estudio Familiar de Influencia Genetica en Alzheimer (EFIGA) (5R37AG015473, RF1AG015473, R56AG051876), Genetic and Environmental Risk Factors for Alzheimer Disease Among African Americans Study (GenerAAtions) (2R01AG09029, R01AG025259, and 2R01AG048927), Gwangju Alzheimer and Related Dementias Study (GARD) (U01AG062602), Hussman Institute for Human Genomics Brain Bank (HIHGBB) (R01AG027944, Alzheimer Association “Identification of Rare Variants in Alzheimer Disease”), Ibadan Study of Aging (IBADAN) (5R01AG009956), Mexican Health and Aging Study (MHAS) (R01AG018016), Multi-Institutional Research in Alzheimer’s Genetic Epidemiology (MIRAGE) (2R01AG09029, R01AG025259, and 2R01AG048927), Northern Manhattan Study (NOMAS) (R01NS29993), Peru Alzheimer’s Disease Initiative (PeADI) (RF1AG054074), Puerto Rican 1066 (PR1066) (Wellcome Trust (GR066133/GR080002), European Research Council [340755]), Puerto Rican Alzheimer Disease Initiative (PRADI) (RF1AG054074), Reasons for Geographic and Racial Differences in Stroke (REGARDS) (U01NS041588), Research in African American Alzheimer Disease Initiative (REAAADI) (U01AG052 410), Rush Alzheimer Disease Center (ROSMAP) (P30AG10 161, R01AG15819, R01AG17919), University of Miami Brain Endowment Bank (MBB), and University of Miami/Case Western/North Carolina A&T African American (UM/CASE/NCAT) (U01AG052410, R01AG028786). The four LSACs are as follows: the Human Genome Sequencing Center at the Baylor College of Medicine (U54 HG003273), the Broad Institute Genome Center (U54HG003067), The American Genome Center at the Uniformed Services University of the Health Sciences (U01AG057659), and the Washington University Genome Institute (U54HG003079).

Study Funding

Supported by the Iqbal Farrukh & Asad Jamal Fund, the NIH (AG060747 and AG047366, granted to M.D. Greicius; AG066206 and AG066515, granted to Z. He), the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie (grant agreement No. 890650, granted to Y. Le Guen), and the Alzheimer’s Association (AARF-20-683984, granted to M.E. Belloy).

Disclosure

The authors report no disclosures relevant to the manuscript. Go to Neurology.org/NG for full disclosures.

Publication History

Previously published at medRxiv doi.org/10.1101/2021.10.28.21265577. Received by *Neurology: Genetics* February 2, 2022. Accepted in final form May 31, 2022. Submitted and externally peer reviewed. The handling editor was Suman Jayadev, MD.

Appendix Authors

Name	Location	Contribution
Michael E. Belloy, PhD	Department of Neurology and Neurological Sciences, Stanford University, CA	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data; study concept or design; analysis or interpretation of data
Yann Le Guen, PhD	Department of Neurology and Neurological Sciences, Stanford University, CA Institut du Cerveau—Paris Brain Institute—ICM, France	Major role in the acquisition of data
Sarah J. Eger, BA	Department of Neurology and Neurological Sciences, Stanford University, CA	Major role in the acquisition of data
Valerio Napolioni, PhD	School of Biosciences and Veterinary Medicine, University of Camerino, Italy	Drafting/revision of the article for content, including medical writing for content
Michael D. Greicius, MD, MPH	Department of Neurology and Neurological Sciences, Stanford University, CA	Drafting/revision of the article for content, including medical writing for content; study concept or design; analysis or interpretation of data
Zihuai He, PhD	Department of Neurology and Neurological Sciences, Stanford University, CA Quantitative Sciences Unit, Department of Medicine, Stanford University, CA	Drafting/revision of the article for content, including medical writing for content; study concept or design; analysis or interpretation of data

References

1. Sierksma A, Escott-Price V, De Strooper B. Translating genetic risk of Alzheimer’s disease into mechanistic insight and drug targets. *Science*. 2020;370(6512):61-66.
2. Sims R, Hill M, Williams J. The multiplex model of the genetics of Alzheimer’s disease. *Nat Neurosci.* 2020;23(3):311-322.
3. Kunkle BW, Grenier-Boley B, Sims R, et al. Genetic meta-analysis of diagnosed Alzheimer’s disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat Genet* 2019;51(3):414-430.
4. Jansen IE, Savage JE, Watanabe K, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer’s disease risk. *Nat Genet* 2019;51(3):404-413.
5. de Rojas I, Moreno-Grau S, Tesi N, et al. Common variants in Alzheimer’s disease and risk stratification by polygenic risk scores. *Nat Commun* 2021;12(1):3417. DOI: 10.1038/s41467-021-22491-8.
6. Andrews SJ, Fulton-Howard B, Goate A. Interpretation of risk loci from genome-wide association studies of Alzheimer’s disease. *Lancet Neurol* 2020;19(4):326-335.

7. Bis JC, Jian X, Chen BWK, et al. Whole exome sequencing study identifies novel rare and common Alzheimer's-associated variants involved in immune response and transcriptional regulation. *Mol Psychiatry* 2020;25(8):1859-1875.
8. Patel D, Mez J, Vardarajan BN, et al. Association of rare coding mutations with Alzheimer disease and other dementias among adults of European ancestry. *JAMA Netw Open* 2019;2(3):e191350.
9. Ma Y, Jun GR, Zhang X, et al. Analysis of whole-exome sequencing data for Alzheimer disease stratified by APOE genotype. *JAMA Neurol* 2019;76(9):1099-1108.
10. Blue EE, Thornton TA, Kooperberg C, et al. Non-coding variants in MYH11, FZD3, and SORCS3 are associated with dementia in women. *Alzheimers Dement* 2021;17(2): 215-225.
11. Park JH, Park I, Youm EM, et al. Novel Alzheimer's disease risk variants identified based on whole-genome sequencing of APOE ε4 carriers. *Transl Psychiatry* 2021; 11(1):296. DOI: 10.1038/s41398-021-01412-9.
12. He Z, Liu L, Wang C, et al. Identification of putative causal loci in whole-genome sequencing data via knockoff statistics. *Nat Commun* 2021;12(1):3152.
13. He L, Loika Y, Park Y, Bennett DA, Kellis M, Kulminski AM. Exome-wide age-of-onset analysis reveals exonic variants in ERN1, TACR3 and SPPL2C associated with Alzheimer's disease. *Transl Psychiatry* 2021;11(1):146.
14. 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.
15. Le Guen Y, Belloy ME, Napolioni V, et al. A novel age-informed approach for genetic association analysis in Alzheimer's disease. *Alzheimers Res Ther* 2021;13(1):72.
16. 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. 0000000000200012.
17. Crane PK, Foroud T, Montine TJ, Larson EB. Alzheimer's Disease Sequencing Project Discovery and Replication criteria for cases and controls: data from a community-based prospective cohort study with autopsy follow-up. *Alzheimers Dement* 2017;13(12):1410-1413.
18. NIAGADS. *NG00067—ADSP Umbrella*. 2021. dss.niagads.org/datasets/ng00067/ (accessed 2 November 2021).
19. Wickland DP, Ren Y, Sinnwell JP, et al. Impact of variant-level batch effects on identification of genetic risk factors in large sequencing studies. *PLoS One* 2021;16(4): e0249305.
20. Tom JA, Reeder J, Forrest WF, et al. Identifying and mitigating batch effects in whole genome sequencing data. *BMC Bioinformatics* 2017;18(1):1-12.
21. Browning BL, Yu Z. Simultaneous genotype calling and haplotype phasing improves genotype Accuracy and reduces false-positive associations for genome-wide association studies. *Am J Hum Genet* 2009;85(6):847-861.
22. Carson AR, Smith EN, Matsui H, et al. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics* 2014;15:125.
23. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;562(7726):203-209.
24. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;581(7809):434-443.
25. Leung YY, Valladares O, Chou YF, et al. VCPA: genomic variant calling pipeline and data management tool for Alzheimer's Disease Sequencing Project. *Bioinformatics* 2019;35(10):1768-1770.
26. GATK Team. *GATK Best Practices Workflows*. gatk.broadinstitute.org/hc/en-us/articles/360035894751 (accessed 1 February 2021).
27. Chen CY, Pollack S, Hunter DJ, Hirschhorn JN, Kraft P, Price AL. Improved ancestry inference using weights from external reference panels. *Bioinformatics* 2013;29(11): 1399-1406.
28. Conomos MP, Miller MB, Thornton TA. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet Epidemiol* 2015;39(4):276-293.
29. Liu Y, Xie J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *J Am Stat Assoc* 2020;115(529):393-402.
30. Sun Y, Wu S, Bu G, et al. Glial fibrillary acidic protein—Apolipoprotein E (apoE) transgenic mice: astrocyte-specific expression and differing biological effects of astrocyte-secreted apoE3 and apoE4 lipoproteins. *J Neurosci* 1998; 18(9):3261-3272.
31. Yizhar O, Fenno L, Zhang F, Hegemann P, Diesseroth K. Microbial opsins: A family of single-component tools for optical control of neural activity. *Cold Spring Harb Protoc* 2011;6(3):top102. DOI: 10.1101/pdb.top102.
32. Osterfield M, Egelund R, Young LM, Flanagan JG. Interaction of amyloid precursor protein with contactins and NgCAM in the retinotectal system. *Dev Dis* 2008;135(6):1189-1199.
33. Osterhout JA, Stafford BK, Yoshihara Y, et al. Functional development of the accessory optic article contactin-4 mediates axon-target specificity and functional development. *Neuron* 2015;86(4):985-999.
34. Muyas F, Bosio M, Puig A, et al. Allele balance bias identifies systematic genotyping errors and false disease associations. *Hum Mutat* 2019;40(1):115-126.
35. Holstege H, Hulsman M, van der Lee SJ, van den Akker EB. The role of age-related clonal hematopoiesis in genetic sequencing studies. *Am J Hum Genet* 2020;107(3):575-576.

Neurology® Genetics

A Fast and Robust Strategy to Remove Variant-Level Artifacts in Alzheimer Disease Sequencing Project Data

Michael E. Belloy, Yann Le Guen, Sarah J. Eger, et al.

Neurol Genet 2022;8;
DOI 10.1212/NXG.0000000000200012

This information is current as of August 11, 2022

AMERICAN ACADEMY OF NEUROLOGY®