Supplementary Online Content

Chen JA, Wang Q, Davis-Turak J, et al. A multi-ancestral genome-wide exome array study of Alzheimer disease, frontotemporal dementia, and progressive supranuclear palsy. *JAMA Neurol*. Published online February 23, 2015. doi:10.1001/jamaneurol.2014.4040

This supplementary material has been provided by the authors to give readers additional information about their work.

eMethods. Supplemental Methods

Subject recruitment and diagnosis. Patients are referred to these research studies primary through referrals from clinicians in the community. All participants in these project underwent a standard multidisciplinary diagnostic assessment including neurological history and examination, nursing assessment, laboratory evaluation, and a previously described neuropsychological assessment of memory, executive function, visuospatial ability, language, and mood[1]. AD patients were diagnosed by applying NINCDS-ADRDA criteria for probable Alzheimer's disease based on neurological and neuropsychological examination, brain imaging and laboratory assessments to rule out other causes of dementia. All subjects and/or their proxies signed informed consents for genetic studies.

Genotyping. Polymorphisms at APOE (rs429358 and rs7412) were genotyped using the pre-designed TaqMan genotyping assays; a polymorphism tagging MAPT H1/H2 (rs1560310) was genotyped using a custom TaqMan genotyping assay. Exome array genotyping using the Illumina HumanExome arrays was performed by the UCLA Neuroscience Genomics Core using an Illumina iScan confocal laser scanner. For the discovery cohort, samples were randomly assigned to arrays (12 samples per array) and genotyped at the same time. For the replication cohort, samples were randomly assigned to arrays (12 samples per array) and genotyped at the same time, separate from the discovery cohort.

Data pre-processing. First, known and cryptically related individuals were removed. Cryptic relatedness was determined by IBD estimation in PLINK version 1.07[2]. The set of exome array variants was pruned to 18,250 SNPs in approximate linkage equilibrium based on pairwise

genotypic correlation, using a window of 50 SNPs, a step size of 5 SNPs, and an r^2 threshold of 0.5. A total of 32 samples with either self-reported familial relationships or with proportion of alleles IBD greater than 0.2 were removed from further analysis. Second, 23,475 variants with lower than a 98% genotyping rate (21,729 in total), or not in Hardy-Weinberg equilibrium (a *p*-value threshold of 10^{-4}; 2,308 in total) were excluded. Following quality control procedures, the initial discovery cohort was comprised of 226,797 variants genotyped in 216 patients with AD, 163 patients with FTD, 48 patients with PSP, and 200 non-demented controls. Of these 627 patients, analysis of X-chromosome homozygosity using PLINK (--check-sex) identified that gender was consistent in 621 (99.0%) of them (3 cases in which gender could not be called, and 3 cases with discordant gender call). All data was processed using GRCh37/hg19 coordinates.

Heritability estimation. The variance explained by the subset of variants was determined using GCTA version 1.13[3]. A genetic relationship matrix (GRM) was computed for all of the subjects, using 1) all of the typed variants, 2) the exonic content of the chip, or 3) the exonic content of the chip with lower than 5% minor allele frequency within our total cohort. A principal components analysis was performed on the GRM using GCTA. Then, GCTA's REML analysis was used to determine the variance explained by each set of variants, using the first four principal components as a covariate to correct for genetic ancestry.

Variant-level association testing. Variant-level association testing was performed using logistic regression implemented in PLINK[2]. In order to correct for population stratification, multidimensional scaling was performed within PLINK to extract the first four multidimensional scaling axes. These four dimensions were then taken as covariates in the logistic model. In order

to confirm the validity of this approach, association was also performed using FaST-LMM, a linear mixed model method[4] that has been shown to correct for test statistic inflation in the presence of population structure confounders and cryptic relatedness. To avoid problems of "proximal contamination", a subset of 18,875 polymorphisms with minor allele frequency (MAF) greater than 5%, genotyping rate greater than 99%, and in approximate linkage equilibrium based on pairwise genotypic correlation, using a window of 50 SNPs, a step size of 5 SNPs, and an r^2 threshold of 0.5 was used to estimate the genetic similarity matrix. A suggestive p-value threshold of $1x10^{-5}$ was used for the initial screening stage, as described for previous association studies[5]. Power calculations mentioned in the text were performed using the QUANTO version 1.2.4 software (University of Southern California, Los Angeles, CA), assuming a log-additive genetic model. For power calculations and GCTA modeling, population prevalence of neurodegenerative disease in elderly individuals was estimated at 11% in AD[6], 0.022% in FTD[7], and 0.008% in PSP[8].

Gene-level association testing. Gene-level association testing was performed using the Sequence Kernel Association Test (SKAT)[9] and implemented in the R package 'SKAT', R version 2.15.1 (R Foundation for Statistical Computing, Vienna, Austria). To control for population structure, the first four principal components of the genotyping data were input as covariates into the SKAT program. The designed exonic content of the exome array, comprising a subset of 195,728 non-synonymous variants (missense and nonsense) and splice variants, were selected from the quality-controlled exome array content for gene-level testing. A false discovery rate (FDR) was estimated by generating *B*=100 permutations of the genotyping data by randomly shuffling the case and control status of each subject and performing the SKAT analysis, as

previously described.[10] Briefly, each SKAT p-value was tested as a potential threshold *d* for the test statistic. An estimate of the FDR controlled at each p-value threshold *d* was calculated as follows:

$$\mathrm{FDR}(d) = \frac{FP(d)}{TP(d)}(1 - S)$$

Here, *FP*(*d*) is an estimate of the number of false positive genes at the threshold *d*, given by

$$FP(d) = \sum_{b=1}^{B} \sum_{i} I(p_{i,b} \leq d)/B$$

where *B* is the total number of permutations (100), *I* is the indicator function, and $p_{i,b}$ is the p-value of the *i*th gene in the *b*th permutation. Similarly, *TP*(*d*) is an estimate of the total number of positive genes (sum of true and false positives), calculated by

$$TP(d) = \sum_{i} I(p_i \leq d)$$

where p_i is the p-value of the *i*th gene in the experimental dataset. The number of true positive genes (*S*) was conservatively assumed to be vanishingly small compared to the total number of genes. We selected an FDR threshold of 50% to prioritize genes for analysis in a follow-up replication cohort for Alzheimer's disease, and an FDR threshold of 15% for suggestive results for FTD and PSP (for which additional patients were not available). For the second stage of testing, the p-value threshold for a single gene was determined using Bonferroni correction,

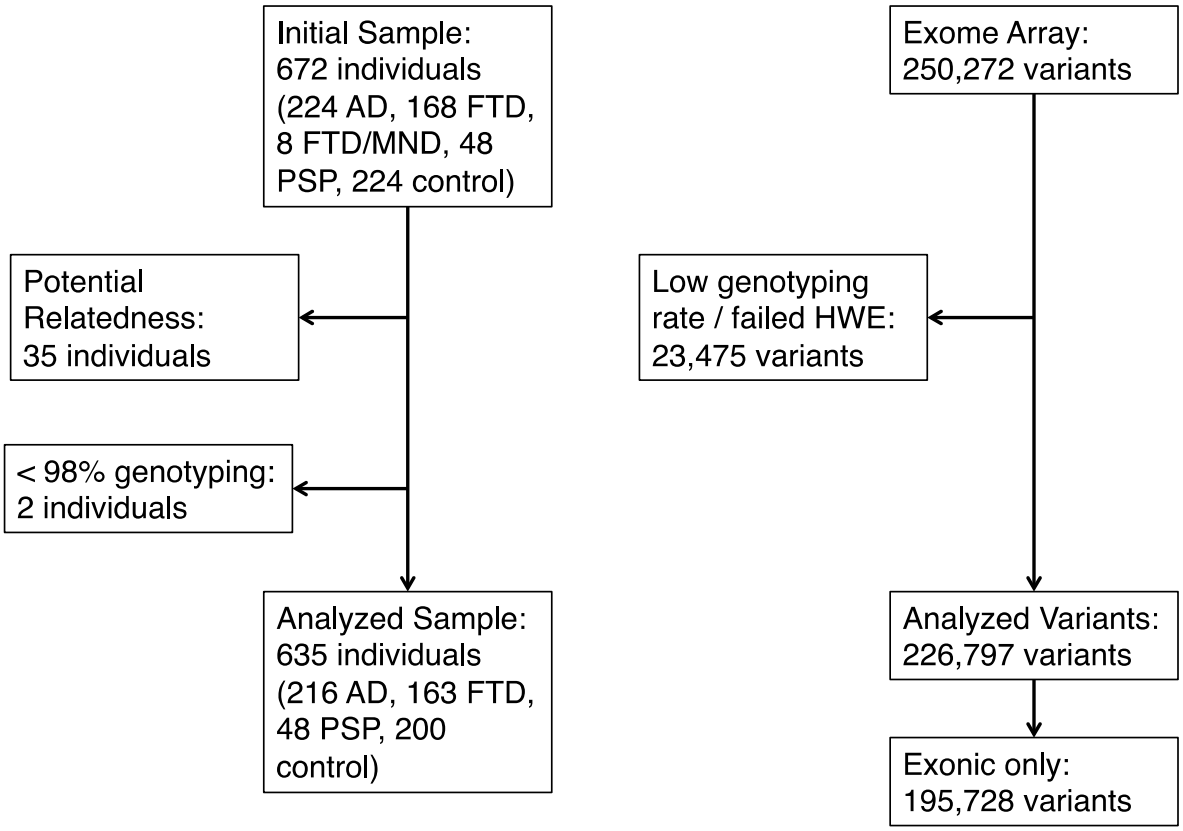
where the number of hypothesis tests was calculated as the product of the total number of genes that passed the first stage of testing and the number of ancestral groups tested in the sample (4), at a family-wise error rate of 0.05.

Expression analysis of publically available data. The expression of mRNA in the brains of patients with Alzheimer's disease and non-demented controls was described by Zhang et al.[11], who measured gene expression (in each of prefrontal cortex, visual cortex, and cerebellum) in 415 cases and 171 controls using microarray. Data presented in the Zhang et al. manuscript was accessed from the Sage Bionetworks Synapse service (ID syn4505). Expression data corresponding to *DYSF* (NM_003494) or *PAXIP1* (NM_007349) was extracted. The expression data had been corrected for technical covariates such as age, gender, post-mortem interval, RNA integrity, and others. Differential expression between patients with Alzheimer's disease and non-demented controls was assessed using the Welch two-sample t-test implemented in R. An additional dataset of microarray expression data described by Webster et al.[12] from various brain regions (frontal, temporal, parietal, and cerebellar cortices) of 176 Alzheimer's disease patients and 188 controls was similarly analyzed. Data presented in the Webster et al. manuscript was accessed from the NCBI Gene Expression Omnibus (GEO Accession GSE15222). This dataset contained a probe for *DYSF* expression, but did not measure the expression of *PAXIP1*. Expression data had been adjusted for technical covariates such as gender, age, APOE status, post-mortem interval, brain region, and others, and was rank invariant normalized.

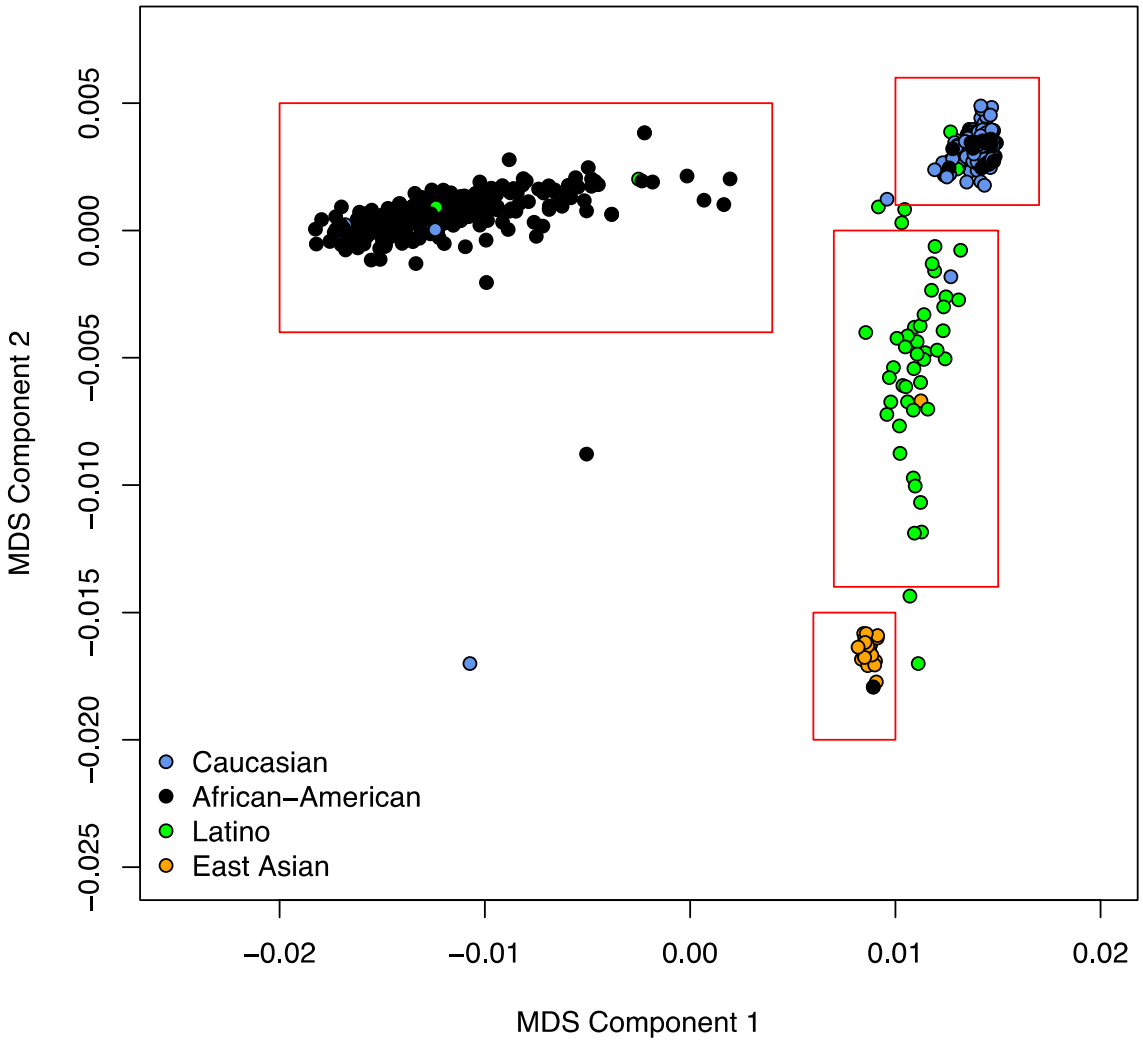
References:

1. Kramer JH, Jurik J, Sha SJ, et al. Distinctive Neuropsychological Patterns in Frontotemporal Dementia, Semantic Dementia, And Alzheimer Disease. *Cogn. Behav. Neurol.* 2003;16(4):211-218.
2. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 2007;81(3):559-575.
3. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.* 2011;88(1):76-82.
4. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. *Nat. Methods.* 2011;8(10):833-835.
5. Duggal P, Gillanders E, Holmes T, Bailey-Wilson J. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genomics.* 2008;9(1):516.
6. Hebert LE, Weuve J, Scherr PA, Evans DA. Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. *Neurology.* 2013;80(19):1778-1783.
7. Borroni B, Alberici A, Grassi M, et al. Is Frontotemporal Lobar Degeneration a Rare Disorder? Evidence from a Preliminary Study in Brescia County, Italy. *J. Alzheimers Dis.* 2010;19(1):111-116.
8. Bower JH, Maraganore DM, McDonnell SK, Rocca WA. Incidence of progressive supranuclear palsy and multiple system atrophy in Olmsted County, Minnesota, 1976 to 1990. *Neurology.* 1997;49(5):1284-1288.
9. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am. J. Hum. Genet.* 2011;89(1):82-93.
10. Xie Y, Pan W, Khodursky AB. A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics.* 2005;21(23):4280-4288.
11. Zhang B, Gaiteri C, Bodea L-G, et al. Integrated Systems Approach Identifies Genetic Nodes and Networks in Late-Onset Alzheimer's Disease. *Cell.* 2013;153(3):707-720.
12. Webster JA, Gibbs JR, Clarke J, et al. Genetic Control of Human Brain Transcript Expression in Alzheimer Disease. *Am. J. Hum. Genet.* 2009;84(4):445-458.

eFigure 1. Flowchart for quality control procedures for the GIFT cohort and exome array variants.

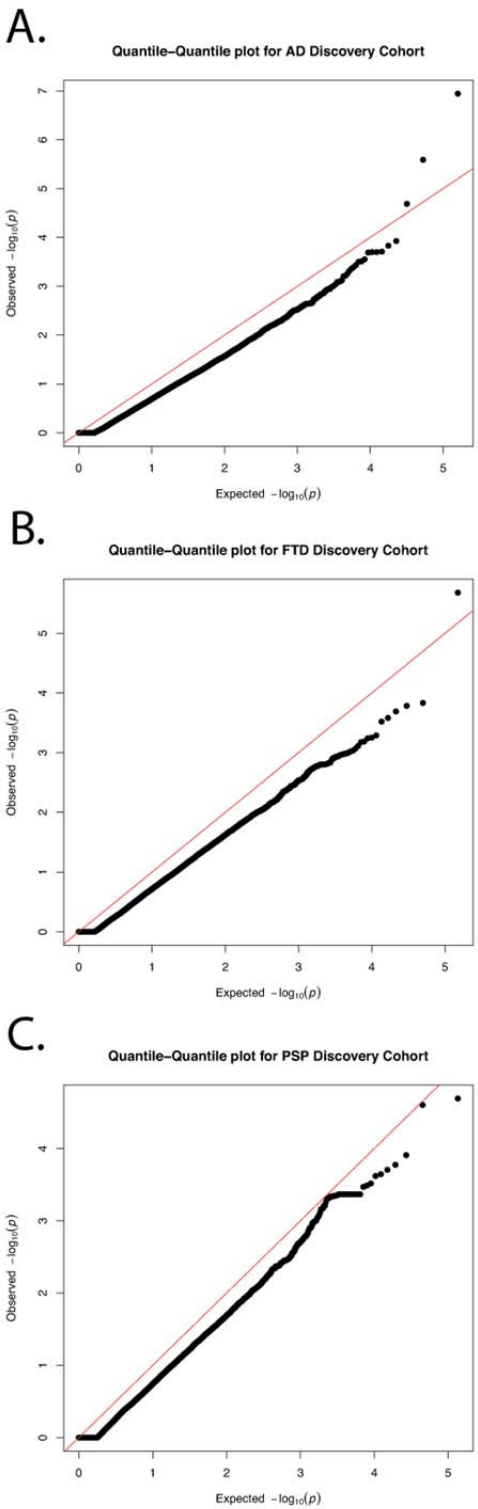


eFigure 2. Scatterplot demonstrating the population structure within the replication cohort evidenced from multidimensional scaling (MDS).



The first two MDS components are shown. Clusters were defined manually using the displayed boxes; subjects with reported ancestry that differed from the majority of the cluster were removed from further analysis.

eFigure 3. Quantile-quantile plots for the variant-level association statistics calculated from the discovery cohort for a) Alzheimer's disease, b) frontotemporal dementia, and c) progressive supranuclear palsy.



eTable 1. Replication of associations for Alzheimer's disease genome-wide association studies.

SNP	Chr.	hg19 Coord.	Mapped Gene	Risk Allele	Lit. Risk Allele	Reference	Disc. Cohort Minor Allele Odds Ratio	Disc. Cohort p-value	Rep. Cohort: European p-value	Rep. Cohort: African-American p-value	Rep. Cohort: Latino p-value	Rep. Cohort: Asian p-value
rs7539409	1	84254735	*TTLL7*	A	A	Heinzen et al., 2009	0.88	0.53	0.6704	0.80	0.65	1
rs3818361	1	207784968	*CR1*	A	A	Hollingworth et al., 2011 (Nat Genet)	1.4	0.095	0.5354	0.97	0.20	0.020
rs6701713	1	207786289	*CR1*	A	A	Naj et al., 2011	1.4	0.095	0.5354	0.97	0.20	0.020
rs12044355	1	231844347	*DISC1*	C	Unknown	Beecham et al., 2009	0.98	0.87	0.9954	0.31	0.94	0.45
rs4676049	2	109635257	Intergenic	T	T	Naj et al., 2010	1.6	0.085	0.9931	0.83	0.11	0.053
rs12989701	2	127887985	*BIN1*	A	A	Hu et al., 2011	1.1	0.56	0.6585	0.26	0.64	NA
rs7561528	2	127889637	*BIN1*	A	A	Naj et al., 2011	1.2	0.35	0.2664	0.45	0.46	0.67
rs744373	2	127894615	*BIN1*	G	G	Hu et al., 2011; Hollingworth et al., 2011 (Nat Genet); Antunez et al., 2011	1.3	0.11	0.2419	0.84	0.019	0.81
rs2121433	2	149557860	*EPC2*	C	Unknown	Kim et al., 2010	0.97	0.85	0.6582	0.64	0.95	0.55
rs4499362	2	149568396	*EPC2*	C*	T	Kim et al., 2010	0.81	0.22	0.4551	0.51	0.81	0.36
rs727153	4	155654421	*LRAT*	T*	C	Abraham et al., 2008	1.2	0.24	0.8745	0.40	0.096	0.47
rs11754661	6	151207078	*MTHFD1L*	G*	A	Naj et al., 2010	0.89	0.67	0.5114	0.73	0.33	1
rs4298437	7	103625877	*RELN*	T	T	Kramer et al., 2010	1.05	0.77	0.117	0.93	0.088	0.21
rs11767557	7	143109139	*EPHA1*	C*	T	Naj et al., 2011	1.2	0.35	0.6068	0.96	0.74	1
rs11782819	8	10334781	*MSRA*	T	Unknown	Kramer et al., 2010	1.1	0.70	0.8328	0.15	0.037	0.20
rs11136000	8	27464519	*CLU*	C	C	Harold et al., 2009	0.91	0.53	0.08604	0.18	0.83	0.94
rs1532278	8	27466315	*CLU*	C	C	Naj et al., 2011	0.92	0.54	0.2425	0.11	0.87	0.94
rs569214	8	27487790	*CLU*	T*	G	Antunez et al., 2011	1.008	0.95	0.4479	0.033	0.47	0.055
rs62209	10	11000339	*CUGBP2*	G	G	Wijsman et al, 2011	0.83	0.32	0.5445	0.67	0.57	0.57
rs4509693	10	102501571	*PAX2*	C	C	Heinzen et al., 2009	1.2	0.35	0.02471	0.22	0.87	NA
rs610932	11	59939307	*MS4A6A*	T	T	Naj et al., 2011; Hollingworth et al., 2011 (Nat Genet)	1.2	0.26	0.6622	0.88	0.55	0.61
rs1562990	11	60023087	*MS4A*	C*	A	Antunez et al., 2011	1.1	0.71	0.8374	0.89	0.60	0.59
rs2373115	11	78091150	*GAB2*	A*	C	Reiman et al., 2007	1.1	0.59	0.7161	0.13	0.13	0.038
rs536841	11	85787824	*PICALM*	T	T	Antunez et al., 2011	0.94	0.69	0.1247	0.73	0.71	0.43
rs561655	11	85800279	*PICALM*	A	A	Naj et al., 2011	0.94	0.69	0.3638	0.16	0.49	0.49
rs3851179	11	85868640	*PICALM*	C	C	Harold et al., 2009	0.95	0.75	0.2186	0.49	0.40	0.83
rs11610206	12	47639526	*FAM113B*	C	Unknown	Beecham et al., 2009	1.006	0.98	0.3417	0.028	0.35	0.13
rs690705	13	34654918	*RFC3*	G	G	Heinzen et al., 2009	1.04	0.79	0.4689	0.28	0.67	0.08
rs11159647	14	84775209	Intergenic	A	A	Bertram et al., 2008	1.1	0.62	0.1393	0.37	0.75	0.59
rs3764650	19	1046520	*ABCA7*	G	G	Hollingworth et al., 2011 (Nat Genet)	1.1	0.68	0.4365	0.90	0.90	0.78
rs2061333	19	44614208	*ZNF224*	G	Unknown	Beecham et al., 2009	1.1	0.63	0.04709	0.33	0.92	0.22
rs157580	19	45395266	*TOMM40*, *APOE*	A	A	Feulner et al, 2009; Kim et al., 2010; Antunez et al., 2011	0.71	0.019	0.1557	0.011	0.52	0.54

rs2075650	19	45395619	*TOMM40*, *APOE*	G	G	Lambert et al., 2009; Harold et al., 2009; Heinzen et al., 2009; Seshadri et al., 2010; Naj et al., 2010; Kim et al., 2010	2.2	$2.1x10^{-5}$	0.02292	0.46	0.14	0.48
rs439401	19	45414451	*APOE*	C	C	Kim et al., 2010; Antunez et al, 2011	0.69	0.012	0.125	0.0065	0.044	0.53
rs4420638	19	45422946	*APOE*	G	G	Coon et al., 2007; Webster et al., 2007; Li et al., 2007	2.3	$2.6x10^{-6}$	0.01208	0.019	0.62	0.33
rs3826656	19	51726613	*CD33*	A*	G	Bertram et al., 2008	0.90	0.52	0.8777	0.049	0.11	0.49
rs3865444	19	51727962	*CD33*	A*	C	Naj et al., 2011	1.02	0.89	0.32	0.042	0.80	0.98
rs7364180	22	42218856	*CCDC134*	G	Unknown	Kim et al., 2010	1.03	0.85	0.7377	0.38	0.49	0.11
rs2573905	X	91402220	*PCDH11X*	T	Unknown	Carrasquillo et al., 2009	0.83	0.28	0.29	0.37	0.29	0.80

eTable 2. Replication of associations for progressive supranuclear palsy genome-wide association studies.

SNP	Chr.	hg19 Coord.	Mapped Gene	Risk Allele	Lit. Risk Allele	Reference	Minor Allele Odds Ratio	p-value
rs1411478	1	180962282	*STX6*	A	A	Hoglinger et al., 2011	1.416	0.1473
rs6687758	1	222164948	Intergenic	G	G	Hoglinger et al., 2011	1.26	0.407
rs6547705	2	87044316	*CD8B*	A	A	Hoglinger et al., 2011	0.8465	0.556
rs7571971	2	88895351	*EIF2AK3*	T	T	Hoglinger et al., 2011	1.407	0.1633
rs1768208	3	39523003	*MOBP*	T	T	Hoglinger et al., 2011	1.616	0.04625
rs6852535	4	123478716	*IL2/IL21*	A*	G	Hoglinger et al., 2011	1.012	0.9604
rs12203592	6	396321	*IRF4*	C	C	Hoglinger et al., 2011	0.849	0.5826
rs242557	17	44019712	*MAPT*	A	A	Hoglinger et al., 2011	1.38	0.1585
rs8070723	17	44081064	*MAPT*	A	A	Hoglinger et al., 2011	0.1796	0.000431

* does not match literature risk allele

eTable 3. Gene-level association statistics for Alzheimer's disease using the Sequence Kernel Association Test (SKAT) for genes with FDR < 50% in the discovery cohort.

Gene	No. Typed Variants	Disc. AD p-value	Disc. FTD p-value	Disc. PSP p-value	Rep. European p-value	Rep. African-American p-value	Rep. Latino p-value	Rep. Asian p-value	HGMD Alzheimer's Disease Gene
DYSF	84	$5.5x10^{-5}$	0.15	0.22	0.076	0.45	0.10	1	FALSE
PAXIP1	7	0.00023	0.21	0.43	0.30	0.81	0.016	0.037	FALSE
TOP1MT	33	0.00029	0.0062	1	0.31	0.0060	0.40	0.40	FALSE
C3orf1	8	0.00039	0.36	0.82	0.70	0.10	0.56	0.81	FALSE
SETDB1	14	0.00041	0.78	0.18	0.48	0.30	1	0.23	FALSE
CRISPLD1	12	0.00045	0.23	0.081	0.71	1	1	0.66	FALSE