Whole-Genome Sequencing Analysis Reveals New Susceptibility Loci and Structural Variants Associated with Progressive Supranuclear Palsy

Hui Wang[1,2]*, Timothy S Chang[3]*, Beth A Dombroski[1,2], Po-Liang Cheng[1,2], Vishakha Patil[3], Leopoldo Valiente-Banuet[3], Kurt Farrell[4], Catriona Mclean[5], Laura Molina-Porcel[6,7], Alex Rajput[8], Peter Paul De Deyn[9,10], Nathalie Le Bastard[11], Marla Gearing[12], Laura Donker Kaat[13], John C Van Swieten[13], Elise Dopper[13], Bernardino F Ghetti[14], Kathy L Newell[14], Claire Troakes[15], Justo G de Yébenes[16], Alberto Rábano-Gutierrez[17], Tina Meller[18], Wolfgang H Oertel[18], Gesine Respondek[19], Thomas Arzberger[20,21], Sigrun Roeber[22], Pau Pastor[23,24], Alexis Brice[25], Alexandra Durr[25], Isabelle Le Ber[25], Thomas G Beach[26], Geidy E Serrano[26], Lili-Naz Hazrati[27], Irene Litvan[28], Rosa Rademakers[29,30], Owen A Ross[30], Douglas Galasko[28], Adam L Boxer[31], Bruce L Miller[31], Willian W Seeley[31], Vivanna M Van Deerlin[1], Charles L White III[32], Huw Morris[33], Rohan de Silva[34], John F Crary[4], Alison M Goate[35], Jeffrey S Friedman[36], Yuk Yee Leung[1,2], Giovanni Coppola[3,37], Adam C Naj[1,2,38], Li-San Wang[1,2], Dennis W Dickson[30]#, Günter U Höglinger[19]#, Gerard D Schellenberg[1,2]#, Daniel H Geschwind[3,39,40]#, Wan-Ping Lee[1,2]#

*These authors contributed equally to this work.

#These authors are corresponding authors.

[1]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[2]Penn Neurodegeneration Genomics Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[3]Movement Disorders Programs, Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA
[4]Department of Pathology, Department of Artificial Intelligence & Human Health, Nash Family, Department of Neuroscience, Ronald M. Loeb Center for Alzheimer's Disease, Friedman Brain, Institute, Neuropathology Brain Bank & Research CoRE, Icahn School of Medicine at Mount Sinai, New York, NY, USA.
[5]Victorian Brain Bank, The Florey Institute of Neuroscience and Mental Health, Parkville, Victoria, Australia
[6]Alzheimer's disease and other cognitive disorders unit. Neurology Service, Hospital Clínic, Fundació Recerca Clínic Barcelona (FRCB). Institut d'Investigacions Biomediques August Pi i Sunyer (IDIBAPS), University of Barcelona, Barcelona, Spain
[7]Neurological Tissue Bank of the Biobanc-Hospital Clínic-IDIBAPS, Barcelona, Spain
[8]Movement Disorders Program, Division of Neurology, University of Saskatchewan, Saskatoon, Saskatchewan, Canada
[9]Laboratory of Neurochemistry and Behavior, Experimental Neurobiology Unit, University of Antwerp, Wilrijk (Antwerp), Belgium
[10]Department of Neurology, University Medical Center Groningen, NL-9713 AV Groningen, Netherlands

[11]Fujirebio Europe NV, Technologiepark 6, 9052 Gent, Belgium
[12]Department of Pathology and Laboratory Medicine and Department of Neurology, Emory University School of Medicine, Atlanta, GA, USA
[13]Netherlands Brain Bank and Erasmus University, Netherlands
[14]Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, IN, USA
[15]London Neurodegenerative Diseases Brain Bank, King's College London, London, UK
[16]Autonomous University of Madrid, Madrid, Spain
[17]Fundación CIEN - Centro Alzheimer Fundación Reina Sofía, Madrid, Spain
[18]Department of Neurology, Philipps-Universität, Marburg, Germany
[19]Technische Universität München, German Center for Neurodegenerative Diseases (DZNE), Munich, Germany
[20]Department of Psychiatry and Psychotherapy, University Hospital Munich, Ludwig-Maximilians-University Munich, Germany
[21]Center for Neuropathology and Prion Research, Ludwig-Maximilians-University Munich, Germany
[22]German Brain Bank, Neurobiobank Munich
[23]Unit of Neurodegenerative diseases, Department of Neurology, University Hospital Germans Trias i Pujol, Badalona, Barcelona, Spain.
[24]Neurosciences, The Germans Trias i Pujol Research Institute (IGTP) Badalona, Badalona, Spain.
[25]Sorbonne Université, Paris Brain Institute – Institut du Cerveau – ICM, Inserm U1127, CNRS UMR 7225, APHP - Hôpital Pitié-Salpêtrière, Paris, France
[26]Banner Sun Health Research Institute, Sun City, AZ, USA
[27]University McGill, Montreal, Quebec, Canada
[28]Department of Neuroscience, University of California, San Diego, CA, USA
[29]VIB Center for Molecular Neurology, University of Antwerp, Belgium
[30]Department of Neuroscience, Mayo Clinic Jacksonville, FL, USA
[31]Memory and Aging Center, University of California, San Francisco, CA, USA
[32]University of Texas Southwestern Medical Center, Dallas, TX, USA
[33]Departmento of Clinical and Movement Neuroscience, University College of London, London, UK
[34]Reta Lila Weston Institute, UCL Queen Square Institute of Neurology, London, UK.
[35]Department of Genetics and Genomic Sciences, New York, NY, USA; Icahn School of Medicine at Mount Sinai, New York, NY, USA
[36]Friedman Bioventure, Inc., Del Mar, CA, USA
Department of Genetics and Genomic Sciences, New York, NY, USA
[37]Department of Psychiatry, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA, USA
[38]Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[39]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA
[40]Institute of Precision Health, University of California, Los Angeles, Los Angeles, CA, USA

Key words: Progressive Supranuclear Palsy (PSP), Whole-Genome Sequencing (WGS), Genome-Wide Association Study (GWAS), Structural Variants (SVs), Apolipoprotein E (APOE)

Abstract

Progressive supranuclear palsy (PSP) is a neurodegenerative disease characterized by the accumulation of aggregated tau proteins in astrocytes, neurons, and oligodendrocytes. We performed whole genome sequencing (WGS) and conducted association analysis for single nucleotide variants (SNVs), small insertions/deletions (indels), and structural variants (SVs) in a cohort of 1,718 individuals with PSP and 2,944 control subjects. Analysis of common SNVs and indels confirmed known genetic loci at *MAPT*, *MOBP*, S*TX6*, *SLCO1A2*, *DUSP10*, and *SP1*, and also uncovered novel signals in *APOE*, *FCHO1/MAP1S, KIF13A, TRIM24, TNXB, and ELOVL1*. In contrast to Alzheimer's disease (AD), we observed the *APOE* ε2 allele to be the risk allele and the ε4 allele to be protective, a pattern similar to the association pattern observed in age-related macular degeneration (AMD) but the opposite observed for Alzheimer's disease (AD). Analysis of rare SNVs and indels identified significant association in *ZNF592* and further gene network analysis identified a module of neuronal genes dysregulated in PSP. We also observed seven common SVs associated with PSP on the H1/H2 haplotype region (17q21.31) and in a few other loci: *IGH*, *PCMT1*, *CYP2A13*, and *SMCP*. Particularly, in the H1/H2 haplotype region, there is a burden of rare deletions and duplications ($P = 6.73 \times 10^{-3}$) in PSP. Through WGS, we significantly refine our understanding of the genetic basis of PSP, providing new targets for exploring disease mechanisms and therapeutic interventions.

Introduction

Progressive supranuclear palsy (PSP) is a neurodegenerative disease that is pathologically defined by the accumulation of aggregated tau protein in multiple cortical and subcortical regions, especially involving the basal ganglia, dentate nucleus of the cerebellum midbrain[1]. An isoform of tau harboring 4 repeats of microtubule-binding domain (4R-tau) is particularly prominent in these tau aggregates[2]. Clinical manifestations of PSP include a range of phenotypes, including the initially described and most common, PSP-Richardson syndrome that presents with multiple features, including postural instability, vertical supranuclear palsy, and frontal dementia. However, there are several other phenotypes, such as PSP-Parkinsonism, PSP-Frontotemporal dementia, PSP-freezing of gait, PSP-speech and language disturbances, etc[3]. Presentation of these phenotypes varies widely depending on the distribution and severity of the pathology[4,5].

Currently, the most recognized genetic risk locus for PSP is at the H1/H2 haplotype region covering *MAPT* gene at chromosome 17q21.31[6], where individuals carrying the common H1 haplotype are more likely to develop PSP with an estimated odds ratio (OR) of 5.6[7]. Previous studies usually ascribed the observed association in the H1/H2 haplotype to *MAPT*[6,8,9]. However, recent functional dissection of this region using multiple parallel reporter assays coupled to CRISPRi demonstrated multiple risk genes in the area in addition to *MAPT*, including *KANSL1* and *PLEKMHL1*[10]. Genome-wide association studies (GWASs) in PSP have identified common variants in *STX6*, *EIF2AK3*, *MOBP*, *SLCO1A2*, *DUSP10*, *RUNX2*, and *LRRK2* with moderate effect size[7,11–13]. In addition, variants in *TRIM11* were identified as a genetic modifier of the PSP phenotype when comparing PSP with Richardson syndrome to PSP without Richardson syndrome[14].

To date, no comprehensive analysis of single nucleotide variants (SNVs), small insertions and deletions (indels), and structural variants (SVs) in PSP by whole genome sequencing has been conducted. To gain a more comprehensive understanding of the genetic underpinnings of PSP, we performed whole genome sequencing (WGS) and analyzed SNVs, indels and SVs. As a result, we not only validated previously reported genes but also unveiled new loci that provide novel insights into the genetic basis of PSP.

Results

Common SNVs and *indels associated with PSP*

We conducted whole genome sequencing at 30x coverage (**Methods**) in 4,662 European-ancestry samples (1,718 individuals with PSP of which 1,441 were autopsy confirmed and 277 were clinically diagnosed and 2,944 control subjects, **Table 1**). We successfully replicated the association of known loci at *MAPT, MOBP* and *STX6*[7,11,12] and identified a novel signal in *APOE* with a genome-wide significance of $P < 5 \times 10^{-8}$ (**Fig. 1, Fig. S1, Table 2, Table S1**). Furthermore, eight loci showed suggestive significance ($5 \times 10^{-8} < P < 1 \times 10^{-6}$), including two loci reported genome-wide significant (*SLCO1A2* and *DUSP10*) and one locus (*SP1*) reported suggestive significant in previous studies[11,12], as well as five new loci in *FCHO1*/*MAP1S*, *KIF13A*, *TRIM24*, *ELOVL1* and *TNXB.*

MAPT, MOBP and STX6

In the *MAPT* region, a multitude of SNVs and indels in high linkage disequilibrium (LD) with the H1/H2 haplotype remains the most significant association with PSP (**Fig. S2A**). From our analysis, the prominent signal within the *MAPT* region is rs62057121 ($P = 7.45 \times 10^{-78}$, $\beta = -1.32$, MAF = 0.15). Fine mapping (**Methods**) suggests that rs242561 ($P = 4.49 \times 10^{-74}$, $\beta = -1.23$, MAF = 0.16) is likely to be a causal SNV underling the statistical significance. The SNP rs242561 is located in an enhancer region, containing an antioxidant response element that binds with NRF2/sMAF protein complex. The T allele of rs242561 showed a stronger binding affinity for NRF2/sMAF in ChIP-seq analysis, therefore inducing a significantly higher transactivation of the *MAPT* gene[15]. rs242561 and rs62057151 were both in high LD ($r^2 > 0.9$) with H1/H2 (defined by the 238 bp

deletion in *MAPT* intron 9) and represented the same association signal as the H1/H2. However, in previous studies[16,17], the H1c tagging SNV (rs242557) inside the H1/H2 region was found to be significant when conditioning on H1/H2. We confirmed that rs242557 was genome-wide significant after adjusting for H1/H2 ($P = 3.68 \times 10^{-15}$, $\beta = 0.39$, MAF = 0.42) though in weak LD with H1/H2 ($r^2 = 0.14$). To pinpoint the causal genes underlying the association in H1/H2 requires additional functional study. For example, Cooper *et al.*[10] analyzed transcriptional regulatory activity of SNVs and suggested *PLEKHM1* and *KANSL1* were probable causal genes in H1/H2 besides *MAPT*. In *MOBP* (rs11708828, $P = 7.04 \times 10^{-12}$, $\beta = -0.35$, MAF = 0.46, **Fig. S2B**) and *STX6* (rs10753232, $P = 6.79 \times 10^{-10}$, $\beta = 0.31$, MAF = 0.44, **Fig. S2C**), the associated variants were of high allele frequency and exhibited moderate effect size.

APOE and risk of PSP

One newly identified significant locus from our analysis is the well-known Alzheimer's Disease (AD) risk gene, *APOE*. We observed a significant association between the *APOE* ε2 haplotype and an elevated risk of PSP ($P = 9.57 \times 10^{-16}$, $\beta = 0.87$, MAF = 0.06, **Table 3**, **Fig. S3B**). The *APOE* ε2 haplotype is encoded by rs429358-T and rs4712-T, which is considered a protective allele in AD. The increased risk of *APOE* ε2 in PSP has been previously reported in a Japanese cohort, albeit with a relatively small sample size[18]. Furthermore, Zhao *et al.*[19] confirmed that *APOE* ε2 is linked to increased tau pathology in the brains of individuals with PSP and reported a higher frequency of homozygosity of *APOE* ε2 in PSP with an odds ratio of 4.41. Consistent with these findings, our dataset exhibited a higher frequency of homozygosity of rs7412-T in PSP, yielding an odds ratio of 3.91.

For *APOE* ε4 allele, contrary to its association with AD, we observed that rs429358-C exhibits a protective effect against PSP ($P = 5.71 \times 10^{-18}$, β = -0.60, MAF = 0.16, **Table 3**). The lead SNV demonstrating this protective association from our analysis is rs4420638 ($P = 2.91 \times 10^{-19}$, β = -0.57, MAF = 0.20, **Fig. S3A**), which is in LD ($r^2 = 0.74$) with rs429358. In a previous PSP GWAS conducted by Hoglinger *et al.*[7], another *APOE* ε4 tagging SNV (rs2075650, $r^2 = 0.52$ with rs429358) was also found to be diminished (MAF_case = 0.11 and MAF_control = 0.15) in PSP, although not reaching significance ($P = 1.28 \times 10^{-5}$). Notably, in our analysis, rs2075650 reached genome-wide significance ($P = 3.39 \times 10^{-13}$, β = -0.51, MAF = 0.15). *APOE* ε4 or ε2 displayed an independent effect for PSP risk without a significant epistatic interaction with H1/H2 haplotype ($P > 0.05$) (**Fig. S4**).

Given that our dataset included external controls from ADSP collected for Alzheimer's disease studies, there were a potential selection biases for *APOE* ε4 and ε2 in controls. To address this concern, we broke down the allele frequencies of *APOE* ε4 and ε2 by cohorts (**Table S2**) and indicated cohorts with potential selection bias. The association analysis excluding these cohorts (**Methods**) shows the ε2 SNV (rs7412, $P = 1.23 \times 10^{-12}$, β = 0.70, MAF = 0.06) remained genome-wide significant and ε4 SNV (rs429358, $P = 0.02$, β = -0.16, MAF = 0.14) was nominal significant (**Table S3**, **Table S4**).

Suggestive significant loci

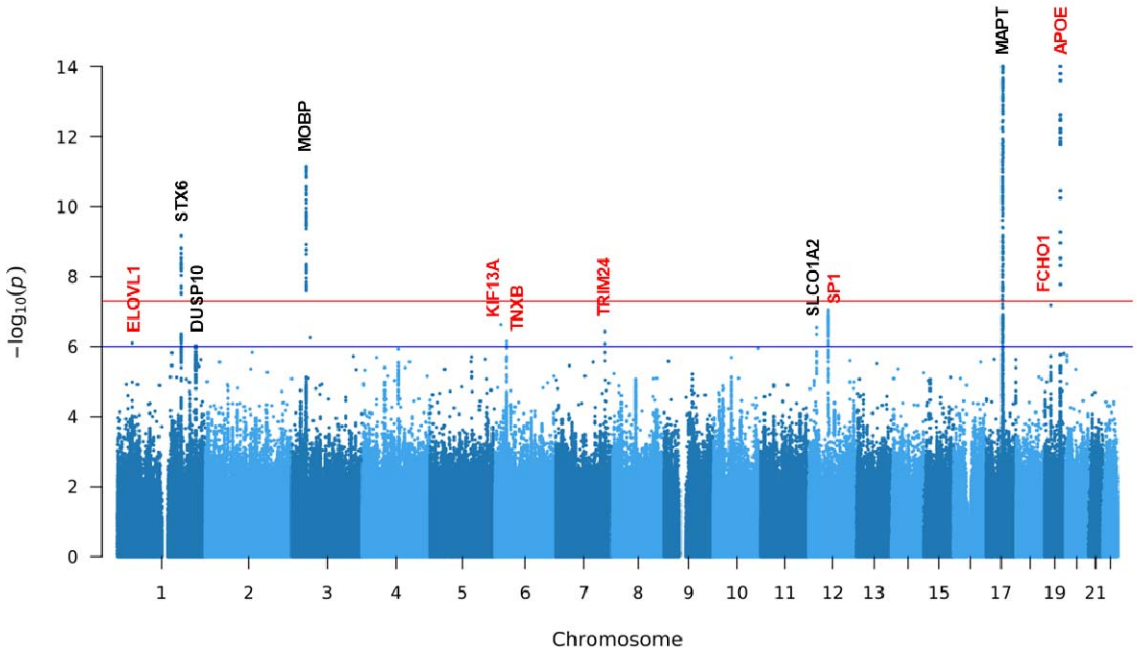
Eight loci were suggestive of significance in our analysis of which three, *SLCO1A2*, *DUSP10*, and *SP1*, were previously reported[11,12]. In *SLCO1A2*, the lead SNV rs74651308 ($P = 2.86 \times 10^{-7}$, β = 0.51, MAF = 0.07, **Fig. S5A**) is intronic and in LD ($r^2 = 0.98$) with missense SNV rs11568563 (*P* =

194 1.45×10^{-6}, $\beta = 0.47$, MAF = 0.07), which was reported in a previous study[11]. About 250 kb
195 upstream of *DUSP10* lies the previously reported SNV rs6687758[11] ($P = 3.36 \times 10^{-6}$, $\beta = 0.29$, MAF
196 = 0.21), which is in LD ($r^2 = 0.98$) with the lead SNV rs12026659 in our analysis ($P = 9.48 \times 10^{-7}$, β
197 = 0.31, MAF = 0.21, **Fig. S5B**). In *SP1*, the reported indel rs147124286[12] ($P = 4.39 \times 10^{-7}$, $\beta = -0.35$,
198 MAF = 0.16) is in LD ($r^2 = 0.995$) with the lead SNV rs12817984 ($P = 8.91 \times 10^{-8}$, $\beta = -0.37$, MAF
199 = 0.16, **Fig. S5C**). Notably, disruption of a transcriptional network centered on *SP1* by causal
200 variants has been implicated previously in PSP[10].

201 Five newly discovered suggestive loci are in *FCHO1/MAP1S*, *KIF13A*, *TRIM24*, *TNXB*, and
202 *ELOVL1*. Within *FCHO1*/*MAP1S*, the most significant signal (rs56251816, $P = 6.57 \times 10^{-8}$, $\beta = 0.35$,
203 MAF = 0.22, **Fig. S6A**) is in the intron of *FCHO1*. rs56251816 is a significant expression
204 quantitative trait locus (eQTL) for both *FCHO1* and *MAP1S* (13 kb upstream of *FCHO1*) in the
205 Genotype-Tissue Expression (GTEx) project[20]. *MAP1S* encodes a microtubule associated protein that
206 is involved in microtubule bundle formation, aggregation of mitochondria and autophagy[21], and
207 therefore, is more relevant than *FCHO1* regarding PSP. *KIF13A*, which encodes a microtubule-based
208 motor protein was also suggestive of significance (rs4712314, $P = 2.37 \times 10^{-7}$, $\beta = 0.27$, AF = 0.51,
209 **Fig. S6B**). The significance in genes involved in microtubule-based processes, such as *MAPT*,
210 *MAP1S* and *KIF13A*, implicates the neuronal cytoskeleton as a convergent aspect of PSP etiology.

211 Other variants with suggestive association evidence include *TRIM24* (rs111593852, $P = 3.75 \times$
212 10^{-7}, $\beta = 0.87$, MAF = 0.02, **Fig. S7A**). *TRIM24* is involved in transcriptional initiation and shows
213 differential expression in individuals with Parkinson disease[22,23]. Another suggestive locus is *TNXB*,
214 located in the major histocompatibility complex (MHC) region on chromosome 6, with the lead SNV

215 rs367364 (P = 7.07 × 10^{-7}, β = -0.37, MAF = 0.13, **Fig. S7B**). Finally, *ELOVL1* yields suggestive
216 evidence of association (rs839764, P = 7.94 × 10^{-7}, β = 0.27, MAF = 0.41, **Fig. S7C**). This gene
217 encodes an enzyme that elongates fatty acids and can cause a neurological disorder with ichthyotic
218 keratoderma, spasticity, hypomyelination and dysmorphic features[24]. Furthermore, we found a few
219 SNV/indels that reached genome-wide or suggestive significance without other supporting variants
220 in LD (**Fig.S1**, **Table S1**). These signals could be due to sequencing errors and need further
221 experimental validation.



222

223 **Fig. 1: Manhattan plot of SNVs/indels for PSP.**
224 Loci with a suggestive or genome-wide significant signal are annotated (novel loci in red and known
225 loci in black). Variants with a P- value below 1 × 10^{-14} are not shown. The red horizontal line
226 represents genome-wide significance level (5 × 10^{-8}). The blue horizontal line represents suggestive
227 significance level (1 × 10^{-6}).

228 **Table 1. Characteristics of study participants.**

	PSP (n = 1,718)		Control (n = 2,944)
	Autopsy Confirmed (n = 1,441)	**Clinical Diagnosed (n = 277)**	
Female	625 (43%)	129 (46%)	1,775 (60%)
Age, y (SD)	68.38 (8.22)	65.72 (7.68)	81.19 (6.01)
***APOE* ε4**[a]			
ε4 carriers	350 (24%)	57 (21%)	905 (32%)
Non-ε4 carriers	1,085 (75%)	216 (78%)	1,913 (65%)
Data missing	6 (0.42%)	4 (1%)	126 (4%)
***APOE* ε2**[b]			
ε2 carriers	234 (16%)	36 (13%)	220 (8%)
Non-ε2 carriers	1,193 (83%)	238 (86%)	2,522 (86%)
Data missing	14 (1%)	3 (1%)	202 (7%)
H2[c]			
H2 carriers	158 (11%)	27 (10%)	1,182 (40%)
Non-H2 carriers	1,283 (89%)	250 (90%)	1,761 (60)
Data missing	0 (0%)	0 (0%)	1 (0.03%)

[a]*APOE* ε4 is represented by the genotypes of rs429358-C.

[b]*APOE* ε2 is represented by the genotypes of rs7412-T.

[c]H2 haplotype is determined by the genotypes of rs8070723-G.

SD, standard deviation.

229 **Table 2. Genome-wide and suggestive significant loci.**

SNV	Chr	Position	Ref	Alt	AF (Alt)	β (Alt)	P	Gene	eQTL/sQTL
Genome-wide Significance ($P < 5 \times 10^{-8}$)									
rs62057121	17	45823394	G	A	0.15	-1.32	7.45×10^{-78}	MAPT	LRRC37A4P[c*]
rs4420638	19	44919689	A	G	0.20	-0.57	2.91×10^{-19}	APOE	TOMM40[b]
rs7412	19	44908822	C	T	0.06	0.87	9.57×10^{-16}	APOE	
rs11708828	3	39458158	C	T	0.46	-0.35	7.04×10^{-12}	MOBP	PRSA[c]
rs10753232	1	180980990	C	T	0.44	0.31	6.79×10^{-10}	STX6	STX6[a*]
Suggestive Significance ($P < 1 \times 10^{-6}$)									
rs56251816	19	17750888	A	G	0.22	0.35	6.57×10^{-08}	FCHO1/MAP1S	
rs12817984	12	53410523	T	G	0.16	-0.37	8.91×10^{-08}	SP1	SP1[a*]
rs4712314	6	17833813	G	T	0.51	0.27	2.37×10^{-07}	KIF13A	
rs74651308	12	21323155	G	A	0.07	0.51	2.86×10^{-07}	SLCO1A2	
rs111593852	7	138449166	C	T	0.02	0.87	3.75×10^{-07}	TRIM24	
rs367364	6	32052169	C	T	0.13	-0.37	7.07×10^{-07}	TNXB	CYP21A1P[c*]
rs839764	1	43367703	T	A	0.41	0.27	7.94×10^{-07}	ELOVL1	TIE1[a*]
rs12026659	1	221976623	G	A	0.21	0.31	9.48×10^{-07}	DUSP10	

Chr, chromosome; Ref, reference allele; Alt, alternative allele; AF, allele frequency.

[*]Represents the SNV regulates multiple genes, and the gene with the smallest *P*-value was shown here (eQTL/sQTL for the brain region was obtained through GTEx).

[a]SNVs with significant eQTL hits.

[b]SNVs with significant sQTL hits.
[c]SNVs with both eQTL and sQTL hits.

Table 3. Allele Frequency of *APOE* ε4 SNV (rs429358) and ε2 SNV (rs7412)

Studies	rs429358		rs7412	
	AF (Case)	AF (Control)	AF (Case)	AF (Control)
PSP WGS (This study)	0.1279	0.1742	0.0844	0.0414
PSP GWAS[25]	0.1159	0.1366	0.0826	0.0794
1000 Genomes Project[26]		0.1512		0.0771
ExAC European (non-Finnish)[27]		0.2078		0.1060
gnomAD V4 European (non-Finnish)[28]		0.1506		0.0783
TOPMed Freeze 8 NFE (Non-Finnish European)		0.1501		0.0752
ADSP R3 Non-Hispanic White[29]	0.3139 (AD as cases)	0.1803	0.0244 (AD as cases)	0.0406

Rare SNVs/indels and network analysis

The heritability of PSP for common SNVs and indels (MAF > 0.01) was estimated to be 20%, while common plus rare SNVs/indels was estimated to be 23% from our analysis using GCTA-LDMS[30] (**Methods**). Therefore, we performed aggregated tests for rare SNVs and indels, and identified *ZNF592* (SKAT-O FDR=0.043, burden test FDR=0.041) with an of OR = 1.08 (95% CI: 1.008-1.16) (**Fig. 2**, **Table 4**, **Table S5**) for protein truncating or damaging missense variants (**Methods**). There was no genomic inflation with a λ=1.07 (**Fig. 2**). Risk in *ZNF592* was imparted by 16 unique variants, with one splice donor and 15 damaging missense variants (**Table S5**). *ZNF592* has not been previously associated with PSP but showed moderate RNA expression in the cerebellum compared to other tissues from GTEx (**Fig. S8**). There were no significant genes identified when evaluating protein-truncating variants (PTVs) only or when restricting to loss of function intolerant genes.

Considering that genes do not operate along, but rather within signaling pathways and networks, we and others have shown that better understanding of disease mechanisms can be achieved through gene network analysis[31–33]. Therefore, we scrutinized rare variants within a network framework, focusing on co-expression network analysis performed in PSP post mortem brain that had previously identified a brain co-expression module, C1, which was conserved at the protein interaction level and enriched for common variants in PSP[34]. We found this C1 neuronal module was significantly enriched with PSP rare variants (P = 0.006, OR [95% CI] = 1.31 [1.01-1.70], **Table 4**; **Table S3**). Genes from the C1 module were more likely to be loss of function intolerant compared to the background of all brain expressed genes (**Methods**) (**Fig. S8**). To ensure that this was association not spurious, we performed permutation testing using random gene modules of brain expressed genes with the same number of genes as C1. The C1 module remains significant (Permutation P = 0.078). Exploring GTEx, we found that C1 genes are highly expressed in brain tissues including the cerebellum, frontal cortex, and basal ganglia (**Fig. S8**), consistent with regions affected in this disorder.

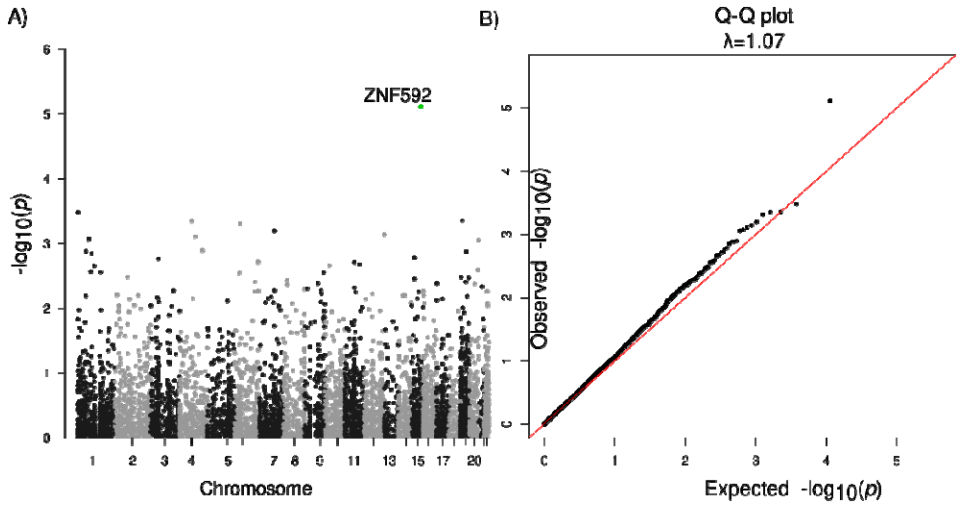


Fig. 2: Association analysis of rare SNVs/indels.

A. Manhattan plot for genes with protein truncating variants or damaging missense variants. **B.** Q-Q plot of gene *P*-values with protein truncating variants or damaging missense variants.

263 **Table 4. Association analysis of ZNF592 and the C1 module.**

Gene	Variants	Total MAC	Case MAC	Control MAC	Fraction Case	Fraction Control	OR (95% CI)	SKAT-O		Burden	
								FDR	P	FDR	P
ZNF592	16	19	8	11	0.0023	0.0018	1.08 (1.01-1.16)	0.044	7.60×10^{-6}	0.041	7.30×10^{-06}
Module	**Variants**	**Total MAC**	**Case MAC**	**Control MAC**	**Fraction Case**	**Fraction Control**	**OR (95% CI)**	**Permutation test**	**P**	**Permutation test**	**P**
C1	180	234	101	133	0.029	0.022	1.31 (1.01-1.70)	0.19	0.048	0.078	0.006

SVs associated with PSP

Seven high-confident SVs achieved genome-wide significance with PSP (**Table 5**, **Fig. S9**), including three deletions tagging the H2 haplotype. The most significant signal is a 238 bp deletion in *MAPT* intron 9 (**Fig. S10A**, chr17:46009357-46009595, $P = 3.14 \times 10^{-50}$, AF = 0.16) that has been reported on the H2 haplotype[35,36] and is in LD ($r^2 = 0.99$) with the lead SNV, rs62057121 (chr17:45823394, $P = 7.45 \times 10^{-78}$, $\beta = -1.32$, MAF = 0.15), in the *MAPT* region. Adding to this, two other deletions, one spanning 314 bp (**Fig. S10B**, chr17:46146541-46146855, AF = 0.19) and the other covering 323 bp (**Fig. S10C**, chr17:46099028-46099351, AF = 0.22), both are Alu elements and in LD ($r^2 > 0.8$) with the top signal (the 238 bp deletion). This observation indicates that transposable elements may play an important role in the evolution of H1/H2 haplotype structure.

Beyond the identified SVs in the H1/H2 region, we uncovered a significant deletion (chr14:105864208-105916743, $P = 4.74 \times 10^{-14}$, AF = 0.01) within the immunoglobulin heavy locus (*IGH*), which is a complex SV region (**Fig. S11**) related to antigen recognition. Moreover, a 619 bp deletion (chr6:149762615-149763234, $P = 8.60 \times 10^{-12}$, AF = 0.55; **Fig. S10D**) in *PCMT1* displayed increased risk of PSP with an odds ratio of 4.19. The odds ratio increased to 8.38 when comparing 1,244 individuals with homozygous deletions in *PCMT1* with the rest of sample set. *PCMT1* encodes a type □ class of protein carboxyl methyltransferase enzyme that is highly expressed in the brain[37] and is able to ameliorate $A\beta_{25-35}$ induced neuronal apoptosis[38,39]. Additionally, we found a deletion between *CYP2F1* and *CYP2A13* (chr19:41102802-41104285, AF = 0.17) and an insertion in *SMCP* (chr1:152880979-152880979, AF = 0.74) which were also significant (**Table 5**). The 1.5 kb deletion (chr19:41102802-41104285) almost completely overlaps the SINE-VNTR-Alus (SVA) transposon

285 region annotated by RepeatMasker[40].

286 **Table 5. Significant structural variants from association analysis ($P < 5 \times 10^{-8}$).**

Name	N	AF	beta	P	AF (case)	AF (control)	Odds Ratio	Fisher's P	Gene
chr17:46009357-46009595:DEL*	4357	0.16	-1.22	3.14×10^{-50}	0.054	0.23	0.19	5.80×10^{-118}	MAPT
chr17:46146541-46146855:DEL*	3697	0.19	-1.12	2.13×10^{-39}	0.079	0.25	0.26	1.58×10^{-83}	KANSL1
chr17:46099028-46099351:DEL*	3699	0.22	-1.07	3.88×10^{-37}	0.11	0.28	0.33	2.05×10^{-66}	KANSL1
chr14:105864208-105916743:DEL	4378	0.010	-1.53	4.74×10^{-14}	0.0053	0.014	0.39	1.33×10^{-04}	IGH
chr6:149762615-149763234:DEL	3811	0.55	0.50	8.60×10^{-12}	0.75	0.42	4.19	6.00×10^{-182}	PCMT1
chr19:41102802-41104285:DEL	2921	0.17	0.64	7.46×10^{-09}	0.21	0.14	1.59	5.95×10^{-11}	CYP2A13
chr1:152880979-152880979:INS	2872	0.74	0.67	2.37×10^{-08}	0.79	0.71	1.62	1.46×10^{-13}	SMCP

*Represents SVs with DNA samples available and PCR validated

SVs in H1/H2 haplotype region

The H1/H2 region stands out as the pivotal genetic risk factor for PSP[17,41]. The H2 haplotype exhibits a reduced odds ratio of 0.19, as we observed the allele frequency of the 238 bp H2-tagging deletion is 23% in PSP and only 5% in control ($P < 2.2 \times 10^{-16}$). Moreover, our analysis pointed out five common (MAF > 0.01) and 12 rare deletions and duplications in the region (**Table 6**), ranging from 88 bp to 47 kb. Additionally, one common and four rare high-confidence insertions were reported in the region.

Of the five common deletions and duplications (**Fig. S12**), three show genome-wide significant association with the disease (**Table 2**); four are located in regions with transposable elements (SVA, L1, or Alu) and in LD (r^2 from 0.63 to 0.92) with the 238 bp H2-tagging deletion (**Methods**). This further highlights the important role of transposable elements in shaping the landscape of H1/H2 region.

Among the 12 rare deletions and duplications (**Fig. S13**), five are located in potentially functional regions, such as splice sites, exons, and transcription factor binding sites (**Table 6**). Particularly, one deletion (chr17:45993882-45993970) in exon 9 of *MAPT* was identified in a PSP patient, adding to previous reports of exonic deletions in the *MAPT* in frontotemporal dementia, such as deletion of exon 10[42] and exons 6-9[43] in *MAPT*. Using the SKAT-O test (N = 4,432), the 12 rare CNVs displayed a significantly higher burden in PSP than controls ($P = 0.01$, OR = 1.64).–

Table 6. High-confident structural variants in the H1/H2 haplotype region

Name	Size	N	AF	AF (PSP)	AF (Control)	Gene	Annotation
chr17:46099028-46099351:DEL[a]*	323	3,699	0.24	0.11	0.28	KANSL1	intron
chr17:46146541-46146855:DEL[a]*	314	3,697	0.21	0.08	0.25	KANSL1	intron
chr17:46237619-46238142:DEL[a]	523	3,686	0.19	0.09	0.22	MAPK8IP1P1	intergenic
chr17:46009357-46009595:DEL[a]*	238	4,357	0.19	0.05	0.23	MAPT	intron
chr17:46277789-46282210:DEL	4,421	4,233	0.12	0.03	0.15	ARL17B	intron
chr17:46113802-46113802:INS	311	2,464	0.31	0.32	0.32	KANSL1	intron

Name	Size	N	N (Carriers)	N (PSP)	N (Control)	Gene	Annotation
chr17:46811121-46811289:DEL[a]	168	2,614	36	15	21	WNT3	intron
chr17:45847702-45851880:DEL[a]	4,178	4,427	31	17	14	MAPT-AS1	splicing
chr17:46837153-46839088:DEL[a]	1,935	4,415	12	8	4	WNT9B	intron
chr17:45918825-45920861:DEL[a]	2,036	4,422	1	0	1	MAPT	intron
chr17:45916681-45920693:DEL	4,012	4,430	3	0	3	MAPT	intron
chr17:45570198-45572012:DEL	1,814	4,243	3	2	1	AC091132.4	intron
chr17:45334194-45381549:DEL[a]	47,355	4,430	1	0	1	AC003070.2	transcript ablation
chr17:45311955-45312258:DEL	303	4,365	2	0	2	MAP3K14	intron
chr17:45894637-45914976:DUP[a]	20,339	4,260	1	1	0	MAPT-AS1	transcript amplification
chr17:45993882-45993970:DEL[a]	88	4,283	1	1	0	MAPT	splicing
chr17:45665996-45666370:DEL[a]	374	4,412	1	1	0	LINC02210-CRHR1	TFBS ablation
chr17:45879141-45881180:DEL	2,039	4,431	1	1	0	MAPT-AS1	intron
chr17:45741582-45741582:INS	315	4,420	10	4	6	LINC02210-CRHR1	intergenic
chr17:45929579-45929579:INS	453	3,025	5	1	4	MAPT	intron
chr17:46754483-46754483:INS	330	3,692	12	2	10	NSF	intron

AF, allele frequency; N, number of individuals with non-missing genotypes.

*High-quality SVs that were included in association analysis.

[a]Represents SVs with DNA samples available and PCR validated.

306

Discussion

Through comprehensive analysis of whole genome sequence, we identified SNVs, indels and SVs contributing to the risk of PSP. For common SNVs, previously reported regions, including *MAPT*, *MOBP*, STX6, *SLCO1A2*, *DUSP10*, and *SP1*[7,11,12] were replicated in our analysis and novel loci in *APOE*, *FCHO1*/*MAP1S*, *KIF13A*, *TRIM24*, *ELOVL1*, and *TNXB* were discovered. *EIF2AK3* which was significantly associated with PSP in a previous GWAS[17] did not reach significance in our study. The SNV with the lowest *P* around *EIF2AK3* was rs13003510 ($P = 8.30 \times 10^{-5}$, $\beta = 0.22$, MAF = 0.3).

The *APOE* ε4 haplotype was of particular interest as it is a common risk factor for AD, explaining more than a 1/3 of population attributable risk[44,45]. Typically, individuals with one copy of the *APOE* ε4 allele (rs429358-C and rs4712-G) have approximately a threefold increased risk of developing AD, while those with two copies of the allele have an approximately a 12-fold increase in risk[46]. In striking contrast, the ε4 tagging allele rs429358 was protective in PSP and the ε2 tagging allele rs7412 was deleterious. This observation is particularly intriguing since both AD and PSP have intracellular aggregated tau as a prominent neuropathologic feature. Notably, both ε2 allele and ε4 allele have been associated with tau pathology burden in the brain of mice models[19,47], which raises the question of distinct tau species in 4R-PSP versus 3R-4R-AD. It is also notable that the ε2 allele is also associated with increased risk for age-related macular degeneration (AMD), and the ε4 allele was associated with decreased risk[48,49]. These results demonstrate that the same variant may have opposite effects in different degenerative diseases. This is especially important, given the advent of gene editing as a therapeutic modality, and programs focused on changing *APOE* ε4 to ε2. Although

this therapy would likely decrease risk for AD, our results indicate that it would increase risk for PSP, in addition to AMD. From this standpoint, caution is warranted in germ-line genome editing until the broad spectrum of phenotypes associated with human genetic variation is understood.

Burden association tests are an highly valuable for addressing sample size limitations in analyzing rare variats[50]. Indeed, burden testing allowed us to identify *ZNF592*, a classical C2H2 zinc finger protein (ZNF)[51,52], as a candidate risk gene. ZNF proteins have been causative or strongly associated with large numbers of neurodevelopmental disease[53,54] and neurodegenerative disease including Parkinson's disease[55] and Alzheimer's disease[56,57]. *ZNF592* was initially thought to be responsible for autosomal recessive spinocerebellar ataxia 5 from a consanguineous family with neurodevelopmental delay including cerebellar ataxia and intellectual disability due to a homozygous G1046R substitution[58]. However, further analysis of this family identified *WDR73* to be the most likely causative gene, consistent with Galloway-Mowat syndrome, although *ZNF592* may have contributed to the phenotype[59].

We also extended classical gene-based burden analysis to consider rare risk burden in the context of a gene set defined by co-expression networks[34,60]. We leveraged combined previous proteomic and transcriptomic analysis of post-mortem brain from patients afflicted with PSP, and showed that rare variants enrich in the C1 neuronal module, which was the same module enriched with common variants[34]. This, along with our recent work identifying a neuronally-enriched transcription factor network centered around SP1 disrupted by PSP common genetic risk, suggests that although PSP neuropathologically is defined by tufted astrocytes and oligodendroglial coiled bodies[61–63], initial causal drivers of PSP appear to be primarily neuronal.

In analysis of SVs, we found deletions in *PCMT1* and *IGH* were significantly associated with PSP. The *IGH* deletions are in a complex region on chromosome 14 that encodes immunoglobins recognizing foreign antigens. The size of the *IGH* deletion varies across individuals (**Fig. S9**). In addition, the *IGH* deletions can be accompanied by other deletions, duplications, and inversions (**Fig. S9**). These combined make the experimental validation of the deletion challenging. The *PCMT1* deletion is common (AF = 0.55) with an odds ratio of 8.38 for PSP in homozygous individuals.

There were limitations to this study. Not all PSP were pathologically confirmed, although pathological confirmation was available in a significant subset (of the 1,718 PSP individuals, 1,441 were autopsy-confirmed and 277 were clinically-diagnosed). Additionally, the majority of control samples in this study were from ADSP and were initially collected as controls for AD studies. As ADSP is a dataset composed of multiple cohorts from diverse sources, it is imperative to ensure that any observed allele frequency differences between controls and cases can be attributed to the disease itself rather than sample selection biases arising from technical artifacts or batch effects. To mitigate the risk of false reports, we meticulously examined the allele frequencies of both cases and controls, especially in relation to novel and significant signals.

This work represents an important first step; future work is necessary to further delineate the rare genetic risk in PSP harbored in coding and noncoding regions. These results may come to fruition as additional genomic analytical methods are developed, sample size increased, and orthogonal genomic data are integrated. While PSP is rare, it is the most common primary tauopathy, and studying this disease is critical to understanding common pathological mechanisms across tauopathies. Further work to include individuals with diverse ancestry background will also improve

our understanding of genetic architecture of the disease.

Methods

Study subjects

We performed WGS at 30x coverage (**Table S7**) for 1,834 PSP cases and 128 controls from the PSP-NIH-CurePSP-Tau, PSP-CurePSP-Tau, PSP-UCLA, and AMPAD-MAYO cohorts included in ADSP (ng00067) and used 3,008 controls from ADSP[64]. Control subjects were self-identified as non-Hispanic white. WGS data is available on NIAGADS[65]. We removed related subjects (IBD>0.25), five clinically diagnosed PSP who were not found to have PSP on autopsy, and non-Europeans (subjects that were eight standard deviations away from the 1000 Genomes Project European samples[26,66] using the first six principal components), resulting in 1,718 individuals with PSP and 2,944 control subjects. Of the 1,718 PSP individuals, 1,441 were autopsy-confirmed and 277 were clinically-diagnosed (**Table 1**).

Given that our dataset included external controls from ADSP collected for Alzheimer's disease studies, there was a potential selection biases for *APOE* ε4 and ε2 in controls. We broke down the allele frequencies of *APOE* ε4 and ε2 by cohorts (**Table S2**) and reviewed the study design of each cohort. The ADSP-FUS1-APOEextremes study used an age extremes sampling approach stratified by APOE genotype, comparing younger onset AD cases against older cognitively normal controls: the controls were APOE ε 4/ ε 4 controls with age-at-last-assessment ≥ 75 years, APOE ε 3/ ε 4 controls with age-at-last-assessment ≥ 80 years, or APOE ε 3/ ε 3 controls with age-at-last-assessment ≥ 85 years[64]. The ADSP-FUS1-StEPAD1 study aims to identify and

characterize novel genetic variants that promote resilience to AD pathology in the presence of the APOE4 allele: controls from ADSP-FUS1-StEPAD1 were protected APOE4 carriers have normal cognition at older age[64]. The CacheCounty study selects "AD resilient individuals" and define them as individuals who are at least 75 years old, cognitively normal, and carry at least one APOE ε4 allele[67].

Common SNVs/indels analysis

Only biallelic variants were included in common SNVs/indels analysis. Variants were removed if they were monomorphic, did not pass variant quality score recalibration (VQSR), had an average read depth ≥ 500, or if all calls have DP<10 & GQ<20. Individual calls with a DP<10 or GQ<20 were set to missing. Indels were left aligned using the GRCh38 reference[68,69]. Common variants (MAF > 0.01) with a missing rate < 0.1, 0.25 < ABHet < 0.75 and HWE (in control) > 1×10^{-5} were kept for analysis, leaving 7,945,112 SNVs/indels for analysis. Genetic relatedness matrix was obtained using KING[70]. Principal components were obtained by PC-AiR[71] which accounts for sample relatedness. Linear mixed model implemented in R Genesis[72] were used for association. Sex and PC1-5 were adjusted in the linear mixed model. Age was not adjusted as more than half (1,159 of 1,718) of PSP cases had age missing. After association, variants with a $P < 1 \times 10^{-6}$ were reported. For SNVs/indels without supporting evidence from nearby SNVs/indels in LD, we removed possible spurious calls with FS (Phred-scaled *P*-value using Fisher's exact test to detect strand bias) > 4, VQSLOD (Log odds of being a true variant versus being false under the trained gaussian mixture model) < 15, or located in regions of genome showing discrepancy from Telomere-to-Telomere Consortium[73] and GRCh38. Fine-mapping of the H1/H2 region were analyzed using SuSie[74]. We ran

the analysis several times assuming the number of maximum causal variants were from 2 to 10. The only variant (rs242561) robust to the choice of maximum causal variants was reported. To avoid potential confounding effects (particularly for *APOE* alleles), we also performed association analysis (**Table S4**) for suggestive and genome-wide significant signals when excluding subjects from the three cohorts with selection bias against APOE alleles along with cohorts with less than 10 subjects (NACC-Genentech, FASe-Families-WGS, KnightADRC-WGS).

Rare SNVs/indels analysis

Multi-allelic variants were split into biallelic variants. Variants where ALT=*, representing a spanning deletion, were removed. Biallelic and multiallelic variants were concatenated, and duplicated variants were removed. Variants were removed if they were monomorphic, did not pass VQSR, had an average read depth ≥ 500, or if all calls have DP<10 & GQ<20. Individual calls with a DP<10 or GQ<20 were set to missing. Indels were left aligned using the GRCh38 reference[68,69]. Then, variants with a missing rate > 0.1 or a $P_{HWE} < 1 \times 10^{-7}$ in controls were removed, resulting in 91,863,622 variants. We calculated the heritability of PSP using GCTA-LDMS[30] for common SNVs/indels (MAF > 0.01) and common plus rare SNVs/indels. A prevalence of 5 PSP cases per 100,000 individuals (0.00005) was used in the GCTA-LDMS analysis.

For aggregated tests of rare variants, we considered rare protein truncating variants (PTVs) and PTVs/damaging missense variants. Variant were annotated with ANNOVAR (version 2020-06-07)[75] and Variant Effect Predictor (VEP, version 104.3)[76]. PTVs were in protein coding genes (Ensembl,version 104)[77] and had VEP consequence as stop gained, splice acceptor, splice donor or frameshift. Damaging missense variants were in protein coding genes (Ensembl version

104)[77] and had a VEP consequence as missense, CADD score ≥ 15, and PolyPhen-2 HDIV of probably damaging. Rare variants were selected based on a MAF < 0.01% from gnomAD and a MAF < 1% in our dataset. The number of alternative allele variants in protein truncating variants (PTV) and PTV/damaging missense variants was similar across sequencing centers and when evaluated for loss of function intolerant genes (observed/expected score upper confidence interval < 0.35[78]) (**Fig. S14**)

After LD clumping with a r^2 cutoff of 0.2, we applied the bigsnpr R package to perform PCA using variants with MAF > 1%. We tested if genes with PTVs or PTVs/missense variants were associated with PSP using the sequence kernel association test-optimized (SKAT-O)[79] (SKAT R package version 2.0.1)[80]. We used a linear kernel and weighed each variant by the maximum external database MAF where lower MAF would have higher weight. We normalized variant MAFs, where $MAF_{norm} = MAF_{ext}/MAF_{max}$, where MAF_{ext} is the external database MAF from gnomAD[78], and $MAF_{max} = 0.0001$. The variant weight is defined by the MAF_{norm} on the β (1,4) distribution. Thus, variant weight is high at very low MAF_{norm} and spread across the range of 1 to 4. Covariates included sex, PC1-3, and H1/H2 haplotype. *P*-values were FDR corrected for the number of genes with a total minor allele count (MAC) ≥ 10. In addition to SKAT-O, we performed gene burden testing (SKATBinary method='burden'). As SKAT-O does not calculate an odds ratio, we calculated the odds ratio of significant genes using logistic regression with the same covariates as SKAT-O and burden testing, and the same variant weights. We also considered only PTVs or PTVs/missense variants in loss of function intolerant genes (observed/expected score upper confidence interval < 0.35[78]).

We evaluated the C1 module, a gene set, which was previously shown to be composed of neuronal genes and enriched for common variants in PSP[34]. We performed a permutation test (N=1000) of random gene set modules from brain expressed genes that contained the same number of genes as C1. From the human protein atlas (www.proteinatlas.org)[81], brain expressed genes were defined as the union of unique proteins from the cerebral cortex, basal ganglia and midbrain (N=15,638). We calculated SKAT-O *P*-values from these random gene modules to determine the null distribution. We calculated the unadjusted odds ratio of significant genes or gene sets by summing the number of alternate alleles in the gene set among the total number alleles in cases and controls。

Normalized quantification (TPM) gene expression across tissues was obtained from Genotype-Tissue Expression (GTEx)[82]. The expression of ZNF592 and C1 module (summarized as an eigengene[83]) were plotted.

SV detection and filtering

For each sample, SVs were called by Manta[84] (v1.6.0) and Smoove[85] (v0.2.5) with default parameters. Calls from Manta and Smoove were merged by Svimmer [86] to generate a union of two call sets for a sample. Then, all individual sample VCF files were merged together by Svimmer as input to Graphtyper2 (v2.7.3)[86] for joint genotyping. SV calls after joint-genotyping are comparable across the samples, therefore, can be used directly in genome-wide association analysis[86]. A subset of SV calls was defined as high-quality calls[86]. Details of SV calling pipeline were in our previous study[87].

There are regions in the human genome that tend to have anomalous, or high signal in WGS experiments[88]. SVs that reside in those regions can be unreliable and should be reported. Specifically,

we compiled problematic regions in the genome from the following sources: (1) the ENCODE blacklist: a comprehensive set of regions that could result in erroneous signal[89]; (2) the 1000 Genome masks: regions of the genome that are more or less accessible to next generation sequencing methods using short reads; (3) the set of assembly gaps defined by UCSC; (4) the set of segmental duplications defined by UCUC; (5) the low-complexity regions, satellite sequences and simple repeats defined by RepeatMasker[40]. For each individual SV reported, Samplot[90] or IGV[91] were used to keep only high-confident CNVs and inversions that are supported by read depth or split reads; for insertions, we kept high-confident insertions that are high-quality and not in the masked regions.

SV analysis

For SV association, more strict sample filtering was applied: outlier samples with too many (larger than median + 4*MAD) CNV/insertion calls or too little (smaller than median - 4*MAD) high-quality CNV/insertion calls were removed. There were 4,432 samples (1,703 cases and 2,729 controls) remaining for PSP SV association analysis. Due to more false positives being picked up, the genomic inflation would be high (λ = 1.89, **Fig. S9**) if all SVs were included in the analysis. Therefore, we restricted our analysis to high-quality SVs only, making the genomic inflation drop to 1.27 (**Fig. S9**). The 14,792 high-quality common SVs (MAF > 0.1) with call rate > 0.5 were included in the analysis. Mixed model implemented in R Genesis[72] were used for association. Sex, PCR information, SV PCs 1-5, and SNV PCs 1-5 were adjusted in the mixed model. After association, we manually inspect deletions, duplications, and inversions by Samplot[90] or IGV[91] to keep only those with support from read depth, split read or insert size. For insertions, those not on masked regions were reported.

For SVs inside the H1/H2 region, all SVs those that are not high-quality are included. Then, we removed SVs with missing rate > 0.5 and manual inspect deletions, duplications, and inversions by Samplot[90] or IGV[91] to keep only those with support from read depth, split read or insert size. For insertions, those high-quality ones not on masked regions were kept for analysis. LD between SVs was calculated using PLINK (V1.90 beta)[92].Rare SV burden on H1/H2 region was evaluated by SKAT-O[79] adjusting for gender and PCs 1-5. As SKAT-O does not calculate an odds ratio, we calculated the odds ratio using logistic regression with the same covariates.

Declarations

Ethics approval and consent to participate

Consent for publication

Not applicable.

Availability of data and materials

NIAGADS Data Sharing Service (https://dss.niagads.org/)
https://github.com/whtop/PSP-Whole-Genome-Sequencing-Analysis

Competing interests

Laura Molina-Porcel received income from Biogen as a consultant in 2022. Gesine Respondek is employed by Roche (Hoffmann-La Roche, Basel, Switzerland) since 2021. Her affiliation whilst completing her contribution to this manuscript was München Technische Universität München, German Center for Neurodegenerative Diseases (DZNE), Munich. Thomas G Beach is a consultant for Aprinoia Therapeutics and a Scientific Advisor and stock option holder for Vivid Genomics. Huw Morris is employed by UCL. In the last 12 months he reports paid consultancy from Roche, Aprinoia, AI Therapeutics and Amylyx; lecture fees/honoraria - BMJ, Kyowa Kirin, Movement Disorders Society. Huw Morris is a co-applicant on a patent application related to C9ORF72 - Method for diagnosing a neurodegenerative disease (PCT/GB2012/052140). Giovanni Coppola is currently an employee of Regeneron Pharmaceuticals. Alison Goate serves on the SAB for Genentech and Muna Therapeutics.

Funding

This work was supported by NIH 5UG3NS104095, the Rainwater Charitable Foundation, and CurePSP. HW and PLC are supported by RF1-AG074328, P30-AG072979, U54-AG052427 and U24-AG041689. TSC is supported by NIH K08AG065519 and the Larry L Hillblom Foundation 2021-A-005-SUP. KF was supported by CurePSP 685-2023-06-Pathway and K01 AG070326. MG is supported by P30 AG066511. BFG and KLN are supported by P30 AG072976 and R01 AG080001. TGB and GES are supported by P30AG072980. IR is supported by 2R01AG038791-06A, U01NS100610, R25NS098999, U19 AG063911-1 and 1R21NS114764-01A1. OR is support by U54 NS100693. DG is supported by P30AG062429. ALB is supported by U19AG063911, R01AG073482, R01AG038791, and R01AG071756. BLM is supported by P01 AG019724, R01 AG057234 and P0544014. VMV is supported by P01-AG-066597, P01-AG-017586. HRM is supported by CurePSP, PSPA, MRC, and Michael J Fox Foundation. RDS is supported by CurePSP, PSPA, and Reta Lila Weston Trust. JFC is supported by R01 AG054008, R01 NS095252, R01 AG060961, R01 NS086736, R01 AG062348, P30 AG066514, the Rainwater Charitable Foundation / Tau Consortium, Karen Strauss Cook Research, and Scholar Award, Stuart Katz & Dr. Jane Martin. AMG is supported by the Tau Consortium and U54-NS123746. YYL is supported by U54-AG052427; U24-AG041689. LSW is supported by U01AG032984, U54AG052427, and U24AG041689. DHG is supported by 3UH3NS104095, Tau Consortium. WPL is supported by RF1-AG074328; P30-AG072979; U54-AG052427; U24-AG041689. Cases from Banner Sun Health Research Institute were supported by the NIH (U24 NS072026, P30 AG19610 and P30AG072980), the Arizona Department of Health Services (contract 211002, Arizona Alzheimer's Research Center),

the Arizona Biomedical Research Commission (contracts 4001, 0011, 05-901 and 1001 to the Arizona Parkinson's Disease Consortium) and the Michael J. Fox Foundation for Parkinson's Research. The Mayo Clinic Brain Bank is supported through funding by NIA grants P50 AG016574, CurePSP Foundation, and support from Mayo Foundation.

Acknowledgements

This project is supported by CurePSP, courtesy of a donation from the Morton and Marcine Friedman Foundation. We are indebted to the Biobanc-Hospital Clinic-FRCB-IDIBAPS for samples and data procurement. The acknowledgement of PSP cohorts is listed below, whereas the acknowledgement of ADSP cohorts for control samples can be found in the supplementary materials. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from: https://gtexportal.org/home/datasets the GTEx Portal on 1/27/2022. We also thank to Drs. Murray Grossman and Hans Kretzschmar for their valuable contribution to this work.

AMP-AD (sa000011) data: Mayo RNAseq Study- Study data were provided by the following sources: The Mayo Clinic Alzheimer's Disease Genetic Studies, led by Dr. Nilufer Ertekin-Taner and Dr. Steven G. Younkin, Mayo Clinic, Jacksonville, FL using samples from the Mayo Clinic Study of Aging, the Mayo Clinic Alzheimer's Disease Research Center, and the Mayo Clinic Brain Bank. Data collection was supported through funding by NIA grants P50 AG016574, R01 AG032990, U01 AG046139, R01 AG018023, U01 AG006576, U01 AG006786, R01 AG025711, R01 AG017216,

R01 AG003949, NINDS grant R01 NS080820, CurePSP Foundation, and support from Mayo Foundation. Study data includes samples collected through the Sun Health Research Institute Brain and Body Donation Program of Sun City, Arizona. The Brain and Body Donation Program is supported by the National Institute of Neurological Disorders and Stroke (U24 NS072026 National Brain and Tissue Resource for Parkinson's Disease and Related Disorders), the National Institute on Aging (P30 AG19610 Arizona Alzheimer's Disease Core Center), the Arizona Department of Health Services (contract 211002, Arizona Alzheimer's Research Center), the Arizona Biomedical Research Commission (contracts 4001, 0011, 05-901 and 1001 to the Arizona Parkinson's Disease Consortium) and the Michael J. Fox Foundation for Parkinson's Research.

PSP-NIH-CurePSP-Tau (sa000015) data: This project was funded by the NIH grant UG3NS104095 and supported by grants U54NS100693 and U54AG052427. Queen Square Brain Bank is supported by the Reta Lila Weston Institute for Neurological Studies and the Medical Research Council UK. The Mayo Clinic Florida had support from a Morris K. Udall Parkinson's Disease Research Center of Excellence (NINDS P50 #NS072187), CurePSP and the Tau Consortium. The samples from the University of Pennsylvania are supported by NIA grant P01AG017586.

PSP-CurePSP-Tau (sa000016) data: This project was funded by the Tau Consortium, Rainwater Charitable Foundation, and CurePSP. It was also supported by NINDS grant U54NS100693 and NIA grants U54NS100693 and U54AG052427. Queen Square Brain Bank is supported by the Reta Lila Weston Institute for Neurological Studies and the Medical Research Council UK. The Mayo Clinic Florida had support from a Morris K. Udall Parkinson's Disease Research Center of Excellence (NINDS P50 #NS072187), CurePSP and the Tau Consortium. The samples from the University of

589 Pennsylvania are supported by NIA grant P01AG017586. Tissues were received from the Victorian
590 Brain Bank, supported by The Florey Institute of Neuroscience and Mental Health, The Alfred and
591 the Victorian Forensic Institute of Medicine and funded in part by Parkinson's Victoria and MND
592 Victoria. We are grateful to the Sun Health Research Institute Brain and Body Donation Program of
593 Sun City, Arizona for the provision of human biological materials (or specific description, e.g. brain
594 tissue, cerebrospinal fluid). The Brain and Body Donation Program is supported by the National
595 Institute of Neurological Disorders and Stroke (U24 NS072026 National Brain and Tissue Resource
596 for Parkinson's Disease and Related Disorders), the National Institute on Aging (P30 AG19610
597 Arizona Alzheimer's Disease Core Center), the Arizona Department of Health Services (contract
598 211002, Arizona Alzheimer's Research Center), the Arizona Biomedical Research Commission
599 (contracts 4001, 0011, 05-901 and 1001 to the Arizona Parkinson's Disease Consortium) and the
600 Michael J. Fox Foundation for Parkinson's Research. Biomaterial was provided by the Study Group
601 DESCRIBE of theClinical Research of the German Center for Neurodegenerative Diseases (DZNE).
602 PSP UCLA (sa000017) data: Thank to the AL-108-231 investigators, Adam L Boxer, Anthony E
603 Lang, Murray Grossman, David S Knopman, Bruce L Miller, Lon S Schneider, Rachelle S Doody,
604 Andrew Lees, Lawrence I Golbe, David R Williams, Jean-Cristophe Corvol, Albert Ludolph,
605 David Burn, Stefan Lorenzl, Irene Litvan, Erik D Roberson, Günter U Höglinger, Mary Koestler,
606 Cliff ord R Jack Jr, Viviana Van Deerlin, Christopher Randolph, Iryna V Lobach, Hilary W Heuer,
607 Illana Gozes, Lesley Parker, Steve Whitaker, Joe Hirman, Alistair J Stewart, Michael Gold, and
608 Bruce H Morimoto.

609

Authors' contribution

Study design: TSC, DD, GUH, GDS, DHG, and WPL. Sample collection, brain biospecimens, and neuropathological examinations: TSC, CM, LM, AR, PPDD, NLB, MG, LDK, JCVS, ED, BFG, KLN, CT, JGdY, ARG, TM, WHO, GR, TA, SR, PP, AB, AD, ILB, TGC, GES, LNH, IL, RR, OR, DG, ALB, BLM, WWS, VMVD, CLW, HM, JH, RdS, JFC, AMG, GC, and DHG. Genotype or phenotype acquisition: HW, TSC, VP, LVB, KF, AN, LSW, GDS, DHG, and WPL. Variant detection and variant quality check: HW, TSC, VP, LVB, KF, YYL, and WPL. Statistical analyses and interpretation of results: HW, TSC, KF, AN, GDS, DHG, and WPL. Experimental validation: BAD and PLC. Draft of the manuscript: HW, TSC, GDS, DHG, and WPL. All authors read, critically revised, and approved the manuscript.

References

1. Hauw, J.-J. *et al.* Preliminary NINDS neuropathologic criteria for Steele-Richardson-Olszewski syndrome (progressive supranuclear palsy). *Neurology* **44**, 2015–2015 (1994).
2. Stamelou, M. *et al.* Evolving concepts in progressive supranuclear palsy and other 4-repeat tauopathies. *Nat. Rev. Neurol.* **17**, 601–620 (2021).
3. Hoglinger, G. U. *et al.* Clinical Diagnosis of Progressive Supranuclear Palsy: The Movement Disorder Society Criteria. *Mov. Disord. Off. J. Mov. Disord. Soc.* **32**, 853–864 (2017).
4. Lukic, M. J. *et al.* Long-Duration Progressive Supranuclear Palsy: Clinical Course and Pathological Underpinnings. *Ann. Neurol.* **92**, 637–649 (2022).
5. Ali, F. *et al.* Sensitivity and specificity of diagnostic criteria for progressive supranuclear palsy. *Mov. Disord.* **34**, 1144–1153 (2019).
6. Wen, Y., Zhou, Y., Jiao, B. & Shen, L. Genetics of progressive supranuclear palsy: a review. *J. Park. Dis.* **11**, 93–105 (2021).
7. Höglinger, G. U. *et al.* Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. *Nat. Genet.* **43**, 699–705 (2011).
8. Borroni, B., Agosti, C., Magnani, E., Di Luca, M. & Padovani, A. Genetic bases of Progressive Supranuclear Palsy: the MAPT tau disease. *Curr. Med. Chem.* **18**, 2655–2660 (2011).
9. Rademakers, R., Cruts, M. & Van Broeckhoven, C. The role of tau (MAPT) in frontotemporal dementia and related tauopathies. *Hum. Mutat.* **24**, 277–295 (2004).
10. Cooper, Y. A. *et al.* Functional regulatory variants implicate distinct transcriptional networks in dementia. *Science* **377**, eabi8654 (2022).

11. Sanchez-Contreras, M. Y. *et al.* Replication of progressive supranuclear palsy genome-wide association study identifies SLCO1A2 and DUSP10 as new susceptibility loci. *Mol. Neurodegener.* **13**, 1–10 (2018).

12. Chen, J. A. *et al.* Joint genome-wide association study of progressive supranuclear palsy identifies novel susceptibility loci and genetic correlation to neurodegenerative diseases. *Mol. Neurodegener.* **13**, 1–11 (2018).

13. Jabbari, E. *et al.* Genetic determinants of survival in progressive supranuclear palsy: a genome-wide association study. *Lancet Neurol.* **20**, 107–116 (2021).

14. Jabbari, E. *et al.* Variation at the TRIM11 locus modifies progressive supranuclear palsy phenotype. *Ann. Neurol.* **84**, 485–496 (2018).

15. Wang, X. *et al.* A polymorphic antioxidant response element links NRF2/sMAF binding to enhanced MAPT expression and reduced risk of Parkinsonian disorders. *Cell Rep.* **15**, 830–842 (2016).

16. Anaya, F., Lees, A. & Silva, R. Tau gene promoter rs242557 and allele-specific protein binding. *Transl. Neurosci.* **2**, (2011).

17. Höglinger, G. U. *et al.* Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. *Nat. Genet.* **43**, 699–705 (2011).

18. Sawa, A. *et al.* Apolipoprotein E in progressive supranuclear palsy in Japan. *Mol. Psychiatry* **2**, 341–342 (1997).

19. Zhao, N. *et al.* APOE ε2 is associated with increased tau pathology in primary tauopathy. *Nat. Commun.* **9**, 4388 (2018).

662 20. Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580 (2013).

663 21. Xie, R. *et al.* Microtubule-associated protein 1S (MAP1S) bridges autophagic components with
664 microtubules and mitochondria to affect autophagosomal biogenesis and degradation. *J. Biol.*
665 *Chem.* **286**, 10367–10377 (2011).

666 22. Shi, L. *et al.* Pilot study: molecular risk factors for diagnosing sporadic Parkinson's disease
667 based on gene expression in blood in MPTP-induced rhesus monkeys. *Oncotarget* **8**, 105606
668 (2017).

669 23. Pan, M. *et al.* Tripartite Motif Protein Family in Central Nervous System Diseases. *Cell. Mol.*
670 *Neurobiol.* 1–23 (2023).

671 24. Kutkowska-Kaźmierczak, A. *et al.* Dominant ELOVL1 mutation causes neurological disorder
672 with ichthyotic keratoderma, spasticity, hypomyelination and dysmorphic features. *J. Med. Genet.*
673 **55**, 408–414 (2018).

674 25. Farrell, K. *et al.* Genetic, transcriptomic, histological, and biochemical analysis of progressive
675 supranuclear palsy implicates glial activation and novel risk genes. *bioRxiv* 2023–11 (2023).

676 26. Consortium, 1000 Genomes Project. *A global reference for human genetic variation*. *Nature* vol.
677 526 68 (Nature Publishing Group, 2015).

678 27. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**,
679 285–291 (2016).

680 28. Chen, S. *et al.* A genome-wide mutational constraint map quantified from variation in 76,156
681 human genomes. *bioRxiv* 2022–03 (2022).

682 29. Lee, W.-P. *et al.* Association of Common and Rare Variants with Alzheimer's Disease in over

683 13,000 Diverse Individuals with Whole-Genome Sequencing from the Alzheimer's Disease
684 Sequencing Project. *medRxiv* 2023–09 (2023).
685 30. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing
686 heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).
687 31. Swarup, V. *et al.* Identification of evolutionarily conserved gene networks mediating
688 neurodegenerative dementia. *Nat. Med.* **25**, 152–164 (2019).
689 32. Swarup, V. *et al.* Identification of conserved proteomic networks in neurodegenerative dementia.
690 *Cell Rep.* **31**, (2020).
691 33. Parikshak, N. N., Gandal, M. J. & Geschwind, D. H. Systems biology and gene networks in
692 neurodevelopmental and neurodegenerative disorders. *Nat. Rev. Genet.* **16**, 441–458 (2015).
693 34. Swarup, V. *et al.* Identification of Conserved Proteomic Networks in Neurodegenerative
694 Dementia. *Cell Rep.* **31**, 107807 (2020).
695 35. Baker, M. *et al.* Association of an extended haplotype in the tau gene with progressive
696 supranuclear palsy. *Hum. Mol. Genet.* **8**, 711–715 (1999).
697 36. Wang, H., Wang, L.-S., Schellenberg, G. & Lee, W.-P. The role of structural variations in
698 Alzheimer's disease and other neurodegenerative diseases. *Front. Aging Neurosci.* (2023).
699 37. Mizobuchi, M., Murao, K., Takeda, R. & Kakimoto, Y. Tissue-specific expression of isoaspartyl
700 protein carboxyl methyltransferase gene in rat brain and testis. *J. Neurochem.* **62**, 322–328 (1994).
701 38. Wu, X. *et al.* Neural stem cell-conditioned medium upregulated the PCMT1 expression and
702 inhibited the phosphorylation of MST1 in SH-SY5Y cells induced by Aβ 25-35. *Biocell* **46**, 471
703 (2022).

704 39. Shi, L. *et al.* PCMT1 ameliorates neuronal apoptosis by inhibiting the activation of MST1 after
705 subarachnoid hemorrhage in rats. *Transl. Stroke Res.* **8**, 474–483 (2017).
706 40. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015
707 <http://www.repeatmasker.org>.
708 41. Chen, J. A. *et al.* Joint genome-wide association study of progressive supranuclear palsy
709 identifies novel susceptibility loci and genetic correlation to neurodegenerative diseases. *Mol.*
710 *Neurodegener.* **13**, 41 (2018).
711 42. Rizzu, P. *et al.* High prevalence of mutations in the microtubule-associated protein tau in a
712 population study of frontotemporal dementia in the Netherlands. *Am. J. Hum. Genet.* **64**, 414–421
713 (1999).
714 43. Rovelet-Lecrux, A. *et al.* Partial deletion of the MAPT gene: A novel mechanism of FTDP-17.
715 *Hum. Mutat.* **30**, E591–E602 (2009).
716 44. Farrer, L. A. *et al.* Effects of age, sex, and ethnicity on the association between apolipoprotein E
717 genotype and Alzheimer disease: a meta-analysis. *Jama* **278**, 1349–1356 (1997).
718 45. Selkoe, D. J. & Podlisny, M. B. Deciphering the genetic basis of Alzheimer's disease. *Annu. Rev.*
719 *Genomics Hum. Genet.* **3**, 67–99 (2002).
720 46. Bertram, L., McQueen, M. B., Mullin, K., Blacker, D. & Tanzi, R. E. Systematic meta-analyses
721 of Alzheimer disease genetic association studies: the AlzGene database. *Nat. Genet.* **39**, 17–23
722 (2007).
723 47. Shi, Y. *et al.* ApoE4 markedly exacerbates tau-mediated neurodegeneration in a mouse model of
724 tauopathy. *Nature* **549**, 523–527 (2017).

48. Rasmussen, K. L., Tybjærg-Hansen, A., Nordestgaard, B. G. & Frikke-Schmidt, R. Associations of Alzheimer disease–protective APOE variants with age-related macular degeneration. *JAMA Ophthalmol.* **141**, 13–21 (2023).

49. Klaver, C. C. *et al.* Genetic association of apolipoprotein E with age-related macular degeneration. *Am. J. Hum. Genet.* **63**, 200–206 (1998).

50. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).

51. Cassandri, M. *et al.* Zinc-finger proteins in health and disease. *Cell Death Discov.* **3**, 1–12 (2017).

52. Fedotova, A. A., Bonchuk, A. N., Mogila, V. A. & Georgiev, P. G. C2H2 Zinc Finger Proteins: The Largest but Poorly Explored Family of Higher Eukaryotic Transcription Factors. *Acta Naturae* **9**, 47–58 (2017).

53. Bu, S., Lv, Y., Liu, Y., Qiao, S. & Wang, H. Zinc Finger Proteins in Neuro-Related Diseases Progression. *Front. Neurosci.* **15**, 760567 (2021).

54. Al-Naama, N., Mackeh, R. & Kino, T. C2H2-Type Zinc Finger Proteins in Brain Development, Neurodevelopmental, and Other Neuropsychiatric Disorders: Systematic Literature-Based Analysis. *Front. Neurol.* **11**, 32 (2020).

55. Shin, J.-H. *et al.* PARIS (ZNF746) Repression of PGC-1α Contributes to Neurodegeneration in Parkinson's Disease. *Cell* **144**, 689–702 (2011).

56. Li, R., Strohmeyer, R., Liang, Z., Lue, L.-F. & Rogers, J. CCAAT/enhancer binding protein delta (C/EBPdelta) expression and elevation in Alzheimer's disease. *Neurobiol. Aging* **25**, 991–999

(2004).

57. Ko, C.-Y. *et al.* CCAAT/enhancer binding protein delta (CEBPD) elevating PTX3 expression inhibits macrophage-mediated phagocytosis of dying neuron cells. *Neurobiol. Aging* **33**, 422.e11–25 (2012).

58. Nicolas, E. *et al.* CAMOS, a nonprogressive, autosomal recessive, congenital cerebellar ataxia, is caused by a mutant zinc-finger protein, ZNF592. *Eur. J. Hum. Genet. EJHG* **18**, 1107–1113 (2010).

59. Vodopiutz, J. *et al.* WDR73 Mutations Cause Infantile Neurodegeneration and Variable Glomerular Kidney Disease. *Hum. Mutat.* **36**, 1021–1028 (2015).

60. Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008–1021 (2013).

61. Takahashi, M., Weidenheim, K. M., Dickson, D. W. & Ksiezak-Reding, H. Morphological and biochemical correlations of abnormal tau filaments in progressive supranuclear palsy. *J. Neuropathol. Exp. Neurol.* **61**, 33–45 (2002).

62. Gg, K. *et al.* Distribution patterns of tau pathology in progressive supranuclear palsy. *Acta Neuropathol. (Berl.)* **140**, (2020).

63. Roemer, S. F. *et al.* Rainwater Charitable Foundation criteria for the neuropathologic diagnosis of progressive supranuclear palsy. *Acta Neuropathol. (Berl.)* **144**, 603–614 (2022).

64. Beecham, G. W. *et al.* The Alzheimer's Disease Sequencing Project: study design and sample selection. *Neurol. Genet.* **3**, (2017).

65. Kuzma, A. *et al.* NIAGADS: The NIA Genetics of Alzheimer's Disease Data Storage Site.

767 *Alzheimers Dement.* **12**, 1200–1203 (2016).

768 66. Lowy-Gallego, E. *et al.* Variant calling on the GRCh38 assembly with the data from phase three
769 of the 1000 Genomes Project. *Wellcome Open Res.* **4**, 50 (2019).

770 67. for the Alzheimer's Disease Neuroimaging Initiative *et al.* Linkage, whole genome sequence,
771 and biological data implicate variants in RAB10 in Alzheimer's disease resilience. *Genome Med.*
772 **9**, 100 (2017).

773 68. Genome Reference Consortium. GRCh38 reference 000001405.15.
774 https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_f
775 or_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz.

776 69. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies
777 demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).

778 70. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies.
779 *Bioinformatics* **26**, 2867–2873 (2010).

780 71. Conomos, M. P., Miller, M. B. & Thornton, T. A. Robust inference of population structure for
781 ancestry prediction and correction of stratification in the presence of relatedness. *Genet.*
782 *Epidemiol.* **39**, 276–293 (2015).

783 72. Gogarten, S. M. *et al.* Genetic association testing using the GENESIS R/Bioconductor package.
784 *Bioinformatics* **35**, 5346–5348 (2019).

785 73. Mc Cartney, A. M. *et al.* Chasing perfection: validation and polishing strategies for
786 telomere-to-telomere genome assemblies. *Nat. Methods* **19**, 687–695 (2022).

787 74. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the

788 "Sum of Single Effects" model. *PLoS Genet.* **18**, e1010299 (2022).

789 75. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from

790 high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

791 76. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol*. **17**, 122 (2016).

792 77. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Res*. **50**, D988–D995 (2022).

793 78. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456

794 humans. *Nature* **581**, 434–443 (2020).

795 79. Lee, S. *et al.* Optimal Unified Approach for Rare-Variant Association Testing with Application to

796 Small-Sample Case-Control Whole-Exome Sequencing Studies. *Am. J. Hum. Genet.* **91**, 224–237

797 (2012).

798 80. Seunggeun Lee, Zhangchen Zhao, Larisa Miropolsky, & Michael Wu. SKAT: SNP-Set

799 (Sequence) Kernel Association Test. (2020).

800 81. Sjöstedt, E. *et al.* An atlas of the protein-coding genes in the human, pig, and mouse brain.

801 *Science* **367**, eaay5947 (2020).

802 82. Melé, M. *et al.* Human genomics. The human transcriptome across tissues and individuals.

803 *Science* **348**, 660–665 (2015).

804 83. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis.

805 *BMC Bioinformatics* **9**, 559 (2008).

806 84. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer

807 sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).

808 85. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for

809 structural variant discovery. *Genome Biol.* **15**, 1–19 (2014).

810 86. Eggertsson, H. P. *et al.* GraphTyper2 enables population-scale genotyping of structural variation
811 using pangenome graphs. *Nat. Commun.* **10**, 1–8 (2019).

812 87. Wang, H. *et al.* Structural Variation Detection and Association Analysis of
813 Whole-Genome-Sequence Data from 16,905 Alzheimer's Diseases Sequencing Project Subjects.
814 *medRxiv* (2023).

815 88. Scherer, S. W. *et al.* Challenges and standards in integrating surveys of structural variation. *Nat.*
816 *Genet.* **39**, S7–S15 (2007).

817 89. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE blacklist: identification of
818 problematic regions of the genome. *Sci. Rep.* **9**, 1–5 (2019).

819 90. Belyeu, J. R. *et al.* Samplot: a platform for structural variant visual validation and automated
820 filtering. *Genome Biol.* **22**, 1–13 (2021).

821 91. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV):
822 high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192
823 (2013).

824 92. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage
825 analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

826