nature portfolio

Peer Review File

Genetic, transcriptomic, histological, and biochemical analysis of progressive supranuclear palsy implicates glial activation and novel risk genes

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have conducted an impressive achievement by conducting the largest GWAS of PSP, with the majority of samples being confirmed through autopsy. This is a remarkable achievement. My specific concerns and recommendations are as follows.

Major comments:
1.The authors have conducted extensive work utilizing colocalization, such as prioritizing genes based on the number of colocalized brain regions. Nevertheless, it is recommended to place greater emphasis on specific brain regions affected by PSP to avoid potential false positive results from conducting too many tests in unrelated tissues and brain regions. Additionally, for a more comprehensive and reliable gene prioritization, the integration of results from various analyses is encouraged, extending beyond colocalization alone. An example of such an approach is the utilization of a 'gene prioritization score,' as demonstrated in studies on Alzheimer's disease (see references PMID 30820047, 34493870, and 35379992).

2.I have read the admirable PSP WGS conducted by the same working group in medRxiv. Compared to the WGS paper, this GWAS paper has a larger sample size but identifies fewer loci. What factors contribute to this difference?

Minor comments:
1.Table 1. 'Illumina OEE' and 'llumina OEE' confuse me. What is the difference?

2.Table 2. It is not clear whether the frequencies and ORs belong to REF or ALT alleles.

3.Line 381. It is not clear what are the 16 candidate genes.

4.If permitted by policy, it is advisable to share the summary statistics with the public.

Reviewer #2 (Remarks to the Author):

PSP is the prototypical tau-pathological disorder. Although it has a specific pattern, tau pathology is also encountered in many different neurodegenerative disorders. Studying PSP is essential to disentangle possible shared mechanisms in neurodegeneration. The authors presented the most significant GWAS study to date on PSP patients. The "small" sample size in this study, pointed out as a limitation by the authors, should be balanced with the fact that most of the patients included (>90%) were autopsy-confirmed, which can be regarded as one of the strengths of this study. More than 5,000 autopsy-confirmed subjects, free of specific neurodegenerative disorders, were used as controls. They nominated six susceptibility loci. After bioinformatic analysis and functional fine-mapping, they replicated previous findings in five candidate genes (MAPT, MOBP, STX6, RUNX2, and SLCO1A2) and nominated a novel one (C4A). In half of them, a specific oligodendrocyte gene expression signature was determined, highlighting the potential role of this cell type in PSP pathology. To further explore the new candidate gene C4A effect on PSP, immunohistochemical and biochemical analysis was performed, demonstrating that tau aggregates in oligodendrocytes of tissue from PSP subjects co-localize with C4, as well as higher C4A alpha-chain levels. This discovery reinforces the hypothesis of innate immune responses associated with tauopathies, especially PSP, providing clues to developing new therapeutic targets. The manuscript is clear and well-written, the topic is fascinating, and they used a robust methodology.

The authors included 2,695 subjects with PSP pathology and only 184 with just a clinical diagnosis. As a sensitive analysis, would the results change significantly, excluding those with a clinical diagnosis?

The authors included subjects with comorbid pathological features; they should present the number of subjects with comorbid neurodegenerative pathologies. As a sensitive analysis, would the results differ if we controlled for comorbidity as a confounder?

The authors should explain why they included just the three PCs in the association analysis. What is the percentage of variance explained by three PCs?

The authors should explain why they excluded variants with regression coefficients greater than 5.

Reviewer #3 (Remarks to the Author):

The authors conducted a large-scale GWAS of progressive supranuclear palsy (PSP) using 2779 cases and 5584 controls, and identified six risk loci for PSP. Based on the results of GWAS, the authors performed serial post-GWAS analyses to prioritize potential causal genes. Overall, this is an impressive study. The findings of this study not only provide important insights into the genetic architecture of PSP, but also implicate new therapeutic targets for PSP. My concerns and comments are below:

1. The authors claimed that they conducted the largest GWAS of PSP to date. Considering most of the cases were autopsy-confirmed, including these high-quality cases is a strength of this study. Despite this advantage, I noticed that five out of the six identified risk loci have been reported previously. That is, only one new risk locus was identified in this study, which limits the novelty and impact of this study. The authors need to discuss or acknowledge this in the manuscript.

2. In addition to this study, there were several GWASs of PSP have been published, is it possible to combine these results and perform a GWAS meta-analysis ? This will provide more information about the genetic architecture of PSP.

3. As GWAS is the most important part of this study, detailed information about how to conduct GWAS is needed. Did the cases and controls match well in PCA analysis ? Potential population stratification will affect the GWAS results. The authors need to make sure that cases and controls matched well in PCA, and outlies need to be excluded. In addition, what is about the races of the participants included this study? Whether all individuals were European ancestry ? For clarity, I suggest the authors drawing a flowchart to detail the quality control steps of GWAS, and listing the numbers of cases and controls excluded from each step.

4. The authors conducted several integrative analyses, including colocalization, eQTL and differential gene expression. However, it seems that the authors presented the results of each analysis separately. Combining these results and perform a gene prioritization will provide useful information. If a gene receives support from GWAS, colocalization, eQTL and differential gene expression, then this gene is more likely to be causal. I suggest the authors perform a prioritization analysis, as described previously (PMID: 37945807, PMID: 33589840).

5. The authors carried out differential expression analysis on the 16 candidate genes identified in the significant GWAS loci. This is a nice analysis. However, given colocalization could not tell us if predicted expression of the GWAS identified genes were associated with PSP. I suggest the authors performing a transcriptome-wide association study (TWAS), which will tell us if genetically predicted expression of these genes are associated with PSP. Based on the TWAS, we can know if elevated or decreased expression of a gene is associated with PSP. Differential expression analysis can validate the results of TWAS. By doing so, the authors can integrate different results and draw more reliable conclusions.

Reviewer #4 (Remarks to the Author):

The authors of "Genetic, transcriptomic, histological, and biochemical analysis of progressive supranuclear palsy implicates glial activation and novel risk genes" report some potentially interesting genetic findings in the PSP genetic space. Some comments and suggestions below:

- Data and code availability
Its very disappointing to see that the authors at the Data availability section use a lot of publicly available data, but fail to make their own PSP summary statistics and underlying genotype data available. Similarly with the code availability, just looking at Supplementary Figure 1, seems very complex and important, but "Code availability statement" says => "All software used is publicly available at the URLs or references cited." isn't right. Proper code is needed to see what exact parameters, flows etc was used here.

- Figure 1b, it would be better to color by the 4 cell types. In addition, given the very very modest p-value that does not pass multiple test correction, there isn't enough evidence to say "Integration with cell type-specific epigenomic annotations revealed a unique oligodendrocytic signature that distinguishes PSP from AD and Parkinson's disease."

- Analytical approaches 1) I didn't see any conditional analyses based on the 6 GWAS hits to identify if there are secondary signals in any of the regions 2) a stratified GWAS based on the MAPT is needed here to see if the MAPT is independent of the other 5 signals. 3) is there no clinical PSP cohort available that can be used as replication cohort?

- Differential gene expression section
"The data suggest regionally specific changes of multiple genes identified in loci identified from the GWAS data which may have downstream effects on disease relevant protein expression."
The main issue here that this is bulk data, so could these differences be simply a difference in cell composition between disease and non disease?

- C4A, 1) is the signal the same signal as ALS? Unclear based on currently provided data. 2) What about the association between C4A and Alzheimer?

Minor comments:
"autopsy studies have found PSP pathology in 2-6% of individuals with no PSP diagnosis prior to death, suggesting that it is more prevalent that appreciated in living individuals3-5."
Worth to clarify is no PSP diagnosis means here, NO neurological disease or something else?

"Given the sharing of common tau pathology across multiple neurodegenerative diseases" what other disease have tau pathology?

"Analyses were then performed in GRCh37/hg19 using the INFERNO and SparkINFERNO pipelines which are detailed elsewhere" how was the data converted from hg38 TOPMED to hg19?

"Imputation of C4A and C4B copy number"=> is there any validation possible to include here? Eg pacbio sequencing?

"n=10 for cases vs. n=10 for PSP" => what is the difference here between cases and PSP?

"of PLA2G7, MOBP, MSH5, HLA-DPB1, HLA-DMB, and SLCO1A2 in PSP versus controls (P<0.0025," based on what was this p-value set?

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have conducted an impressive achievement by conducting the largest GWAS of PSP, with the majority of samples being confirmed through autopsy. This is a remarkable achievement. My specific concerns and recommendations are as follows.

Major comments:

1. The authors have conducted extensive work utilizing colocalization, such as prioritizing genes based on the number of colocalized brain regions. Nevertheless, it is recommended to place greater emphasis on specific brain regions affected by PSP to avoid potential false positive results from conducting too many tests in unrelated tissues and brain regions. Additionally, for a more comprehensive and reliable gene prioritization, the integration of results from various analyses is encouraged, extending beyond colocalization alone. An example of such an approach is the utilization of a 'gene prioritization score,' as demonstrated in studies on Alzheimer's disease (see references PMID 30820047, 34493870, and 35379992).

> **We thank the reviewer and have added a column in Table 3 indicating if the significant colocalization results occurred in brain regions vulnerable to pathology in PSP (putamen, substantia nigra, and frontal cortex BA9) and have added language to the text emphasizing this point. We also have added up each criterion and added a score to Table 3. Language was also added on how this was achieved, but in summary we followed the protocol detailed by Fritsche et al. 2016.**

Table 3 | Summary of the computational results

Locus	Fine mapping (credible sets and snps)	Enhancers	Promoters	MPRA	Gene	Coloc*, all brain regions	Coloc*, PSP vulnerable brain regions	Bryois Single Nuc-seq	Bulk RNA-seq logFC** Cortex	Bulk RNA-seq LogFC** Cerebellum	Priority Score
					AL162431.2	1	0	-	-	-	1
1q25.3	2, 4	Astro Oligo	-	-	KIAA1614	6	YES, 2	-	-	-	2
					STX6	11	YES, 3	Oligo	-	0.23	**4**
					ZNF621	-	-	OPC	-	-	1
3p22.1	3, 4	Neurons Oligo	Oligo	Signal observed	**MOBP**	-	-	Oligo, Ext. Neuron	-1.01	-0.90	**3**
					VILL	1	0	-	-	-	1
					RPSA	1	YES, 1	-	-	-	2
					PLA2G7	1	0	-	-	-0.64	2
6p21.1	1,3	Microglia	-	-	SUPT3H	-	-	Oligo	-	-	1
					RUNX2	7	YES, 2	Oligo	-	-	**3**
					TNF	1	YES, 1	-	N/A	N/A	2
					BTNL2	3	0	-	N/A	N/A	1
					TNXB	-	-	-	-	-	1
					HLA-DPB1	-	-	-	-0.50	-	1
6p21.32	-	-	-	-	HLA-DMB	-	-	-	-0.80	-	1
					MSH5	-	-	-	-0.54	-0.69	1
					FLOT1	-	-	-	0.16	-	1
					C4A	-	-	-	-	-	0
					CYP21A1P	-	-	-	-	-	0
12p12.1	1, 1	-	-	-	**SLCO1A2**	-	-	-	-0.69	-0.42	**2**

*Number for coloc indicate the number of brain regions and cell types PP is > 0.5), PSP vulnerable brain regions included were putamen, substantia nigra, and frontal cortex BA9, ** values included on if adjusted p-values were signficantly differentially expressed (after adjustment) in cases vs controls, - indicates no signficant values observed. Bulk RNA-seq fold change values shown only if the differential expression p-value < 0.0025. N/A idicates data not available. Best candidate gene/s is in bold.

2. I have read the admirable PSP WGS conducted by the same working group in medRxiv. Compared to the WGS paper, this GWAS paper has a larger sample size but identifies fewer loci. What factors contribute to this difference?

> We thank the reviewer for the comment and we hypothesize that these novel rare variants in the PSP WGS are due to an aggregate effect based on our inability to infer variants in the reference panel. For this reason, it is common that short read whole genome data will always find more novel variants (specifically, the PSP WGS manuscript has 7,945,112 SNVs/indels vs our 7,230,420 common SNPs). However, we would like to add that there is likely less uncertainty around our estimates due to our larger sample size. Additionally it is not uncommon for signals to fluctuate above and below genome-wide significance. For example, in 2011 we (Naj *et al.*) ran a GWAS of Alzheimer's disease (8,309 LOAD cases, 7,366 cognitively normal controls) and found a locus containing *CD33* to be genome-wide significant, however in 2013 a larger GWAS (Lambert *et al.*) using data generated from IGAP was unable to replicate this result in Stage 2 of the analysis. We believe a similar phenomenon is happening in our work with the *EIF2AK3* locus which is no longer significant here but has been in prior studies. Lastly, the WGS paper identified five genome-wide significant loci ($P<5\times10^{-8}$), whereas here we identify six. We believe the apparent inconsistency is arising from Table 2 in the PSP WGS paper, which included both significant ($P<10^{-8}$) and suggestive loci ($P>10^{-6}$).

Minor comments:

1. Table 1. 'Illumina OEE' and 'Ilumina OEE' confuse me. What is the difference?

> We thank the reviewer and recognize we have made a typo, they should both be Illumina OEE (OmniExpressExome), thus the SNP chips are the same however they were used to genotype different cohorts and different timepoints and were treated independently for QC. To clarify this point we have added "Illumina OEE batch 2".

2. Table 2. It is not clear whether the frequencies and ORs belong to REF or ALT alleles.

> We thank the reviewer for this comment, the odds ratio is the difference between the frequency of the alternate allele between the two groups. Language indicating this has been added to the table.

3. Line 381. It is not clear what are the 16 candidate genes.

> We thank the reviewer and have clarified this in the manuscript and have clarified this point by adding the following line "After re-analyzing the raw data to include relevant covariates, we focused on 16 candidate genes identified in the significant GWAS loci and available in the bulk RNA-seq dataset (*C4A, CYP21A1P, FLOT1, HLA-DPB1, HLA-DMB, KIAA1614, MOBP, MSH5, PLA2G7, RPSA, RUNX2, SLCO1A2, STX6, SUPT3H, VILL, ZNF621*)."

4.If permitted by policy, it is advisable to share the summary statistics with the public.

> We thank the reviewer and the data have been uploaded to NIAGADs for public access, the repository ID is NG00169 and the summary statistics will be openly available in a few weeks.

Reviewer #2 (Remarks to the Author):

PSP is the prototypical tau-pathological disorder. Although it has a specific pattern, tau pathology is also encountered in many different neurodegenerative disorders. Studying PSP is essential to disentangle possible shared mechanisms in neurodegeneration. The authors presented the most significant GWAS study to date on PSP patients. The "small" sample size in this study, pointed out as a limitation by the authors, should be balanced with the fact that most of the patients included (>90%) were autopsy-confirmed, which can be
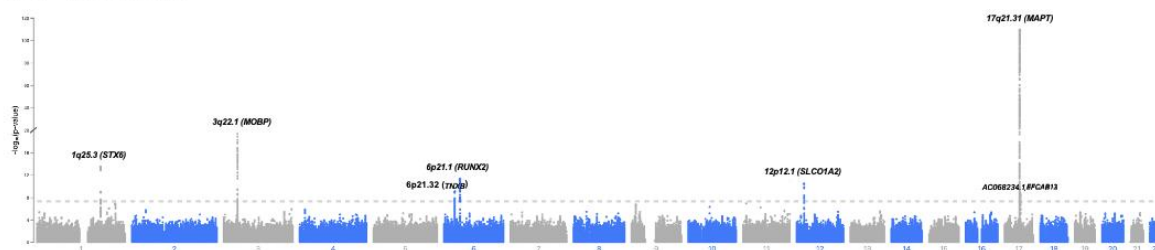
regarded as one of the strengths of this study. More than 5,000 autopsy-confirmed subjects, free of specific neurodegenerative disorders, were used as controls. They nominated six susceptibility loci. After bioinformatic analysis and functional fine-mapping, they replicated previous findings in five candidate genes (MAPT, MOBP, STX6, RUNX2, and SLCO1A2) and nominated a novel one (C4A). In half of them, a specific oligodendrocyte gene expression signature was determined, highlighting the potential role of this cell type in PSP pathology. To further explore the new candidate gene C4A effect on PSP, immunohistochemical and biochemical analysis was performed, demonstrating that tau aggregates in oligodendrocytes of tissue from PSP subjects co-localize with C4, as well as higher C4A alpha-chain levels. This discovery reinforces the hypothesis of innate immune responses associated with tauopathies, especially PSP, providing clues to developing new therapeutic targets. The manuscript is clear and well-written, the topic is fascinating, and they used a robust methodology.

The authors included 2,695 subjects with PSP pathology and only 184 with just a clinical diagnosis. As a sensitive analysis, would the results change significantly, excluding those with a clinical diagnosis?

> We thank the reviewer and have performed this analysis and found the data does not significantly change. Specifically, the number of genome wide significant associations stays the same with only minor changes to the *P*-value. We have added this information as a Supplemental Figure 3 which demonstrates this. The following text was added:

"As a sensitivity analysis, we repeated the analysis excluding the 125 non-autopsy confirmed subjects and did not observe any major differences (Supplementary Figure 3)."
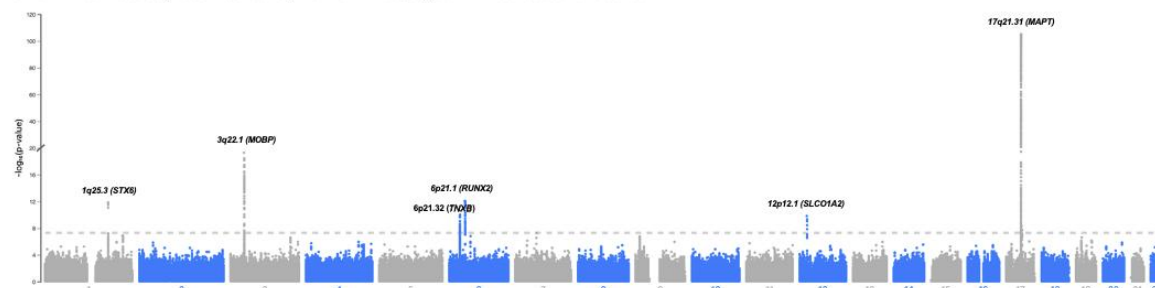
The authors included subjects with comorbid pathological features; they should present the number of subjects with comorbid neurodegenerative pathologies. As a sensitive analysis, would the results differ if we controlled for comorbidity as a confounder?

> We thank the reviewer for this insightful comment. Because we do not have full access to all the neuropathology reports it is not feasible to perform this sensitivity analysis. However, given this is an important point, we opted to pull all the significant SNPs previously identified in GWAS studies of AD and PD and compare those top hits to what we found in our PSP GWAS. The data is now included in Supplementary Tables 1 and 2.

The following line was added to the results "Additionally, we examined signals identified in prior neurodegenerative studies GWAS and found 22 PD SNPs and 13 AD SNP which had a modest association signal (0.05>*P*>0.0005, Supplementary Tables 1 and 2)."

Additionally the following line was added to the discussion "Follow up studies, such as functional genomic analyses, model organism experiments, _stratification based on potential comorbid pathological features_, and a replication cohort are necessary to elucidate how the identified variants affect biological processes related to neurodegeneration."

The authors should explain why they included just the three PCs in the association analysis. What is the percentage of variance explained by three PCs?

> We evaluated the clustering patterns of cases and controls across the first four PCs and observed in relatively standard normal centering of the distribution of PCs starting at PC4, indicating minimal informativeness for population substructure beyond PCs 1-3. We further looked at the eigen vector loadings for the first 10 PCs and observed loadings less than 10 for all PCs after PC3. These evaluations mirror those performed across the majority of GWAS in identifying how many PCs to adjust for as covariates in parsimonious modeling. We have now included PC scatterplots and a scree plot (Supplemental Figure 17) to show the number of informative variants described by PCs.
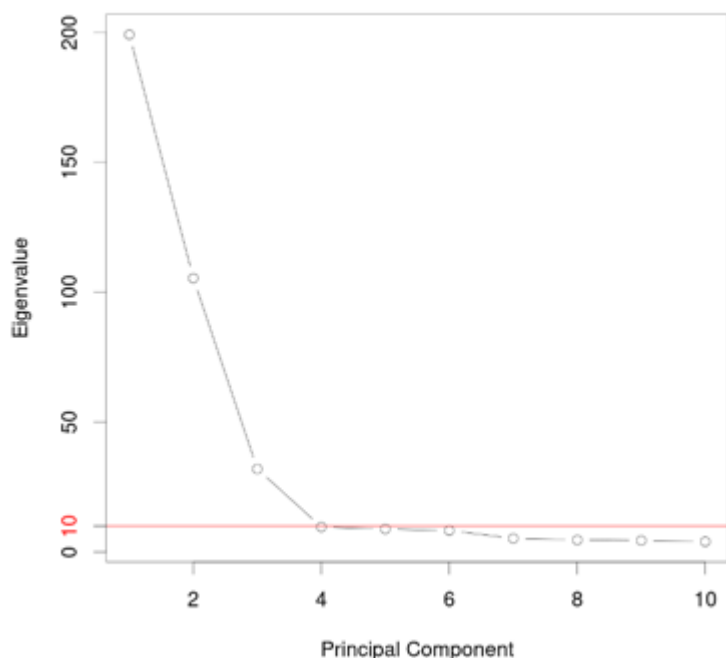
Supplementary Figure 17. Scree Plot for Principal Components Analysis (PCA). The selection threshold for principal components (PCs) to be included in covariate adjustment was a PC eigenvalue>10.

The authors should explain why they excluded variants with regression coefficients greater than 5.

> We thank the reviewer for the comment. A regression coefficient of greater of 5 corresponds to an odds ratio of 148. For context, risk-increasing haplotypes of _APOE_ as the major gene effect in AD report relative risks of ~4, thus an association of the magnitude of OR>100 is unlikely to be a valid measure of association observed in a GWAS of almost any phenotype. A beta coefficient of >5 is likely

indicative of an asymptotic effect due to extremely "small cells" (i.e., fewer than five carriers of a risk allele) or a low-quality variant that had not failed QC. Thus, to be conservative and avoid nominating likely false positives, we applied this threshold.

The following language has been added to the methods:
"After analysis, variants with regression coefficient of |β|>5 and any erroneous estimates (negative standard errors or *P*-values equal to 0 or 1) were excluded from further analysis as these values are likely indicative of asymptotic effects."

Reviewer #3 (Remarks to the Author):

The authors conducted a large-scale GWAS of progressive supranuclear palsy (PSP) using 2779 cases and 5584 controls and identified six risk loci for PSP. Based on the results of GWAS, the authors performed serial post-GWAS analyses to prioritize potential causal genes. Overall, this is an impressive study. The findings of this study not only provide important insights into the genetic architecture of PSP, but also implicate new therapeutic targets for PSP. My concerns and comments are below:

1. The authors claimed that they conducted the largest GWAS of PSP to date. Considering most of the cases were autopsy-confirmed, including these high-quality cases is a strength of this study. Despite this advantage, I noticed that five out of the six identified risk loci have been reported previously. That is, <u>only one new risk locus was identified in this study, which limits the novelty and impact of this study. The authors need to discuss</u> or acknowledge this in the manuscript.

> We thank the reviewer and agree it is important we emphasize this point. Thus we have added the following line to the discussion:

"To this point, the relatively limited availability of genetically- and phenotypically-characterized PSP cases, the global majority of which are incorporated into this the largest study of PSP to-date, still limited this study to the observation of only one novel locus.."

2. In addition to this study, there were several GWASs of PSP [that] have been published, is it possible to combine these results and perform a GWAS meta-analysis? This will provide more information about the genetic architecture of PSP.

> We thank the reviewer for this comment and want to clarify that this current study includes all the samples from the three previous GWAS studies of PSP. Language has been added in the methods to clarify this.

"A full list of the institutions where the material was collected can be found in Supplementary Table 1 and it should be noted a majority of the samples included here were contained in previous studies [8,10,11]."

3. As GWAS is the most important part of this study, detailed information about how to conduct GWAS is needed. Did the cases and controls match well in PCA analysis ? Potential population stratification will affect the GWAS results. The authors need to make sure that cases and controls matched well in PCA and outliers needed to be excluded. In addition, what is about the races of the participants included this study? Whether all individuals were European ancestry ? For clarity, I suggest the authors drawing a flowchart to detail the quality control steps of GWAS, and listing the numbers of cases and controls excluded from each step.

> We thank the reviewer and have added a figure showing the cases and controls do overlap, as well as running additional analysis to confirm that the population was of the expected European ancestry (imposed on 1000 Genomes, Supplementary Figure 16).

 Additionally, we have flowchart showing at what steps samples were excluded (Supplementary Figure 18).

Supplementary Figure 18. Quality Control by Genotyping Platform for Each Dataset. Pre- and Post-QC SNP and sample counts are shown for each genotyping platform subset, and the number of SNPs or of samples removed at each QC step is depicted in italics. A high resolution version of this figure is available in the submission.

The following language was added to the methods:

"Population substructure was rechecked using EIGENSTRAT and plotted and overlaid on 1000 genomes (Supplementary Figure 16)[65,66]."
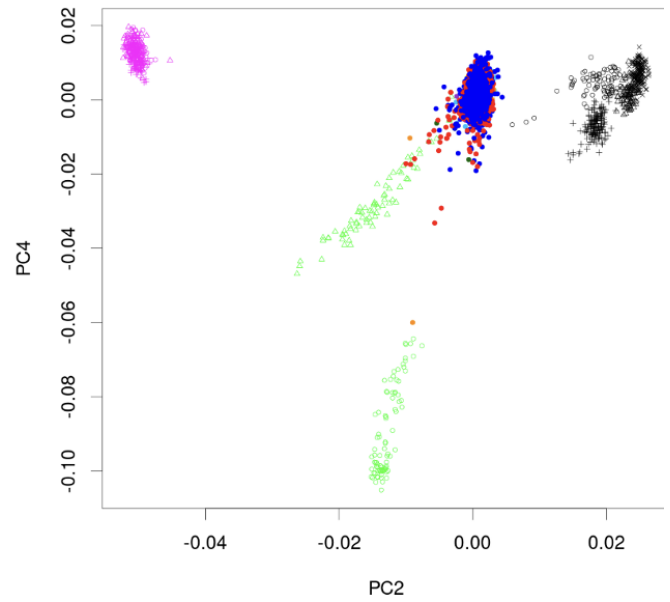
a) PC1 vs PC2
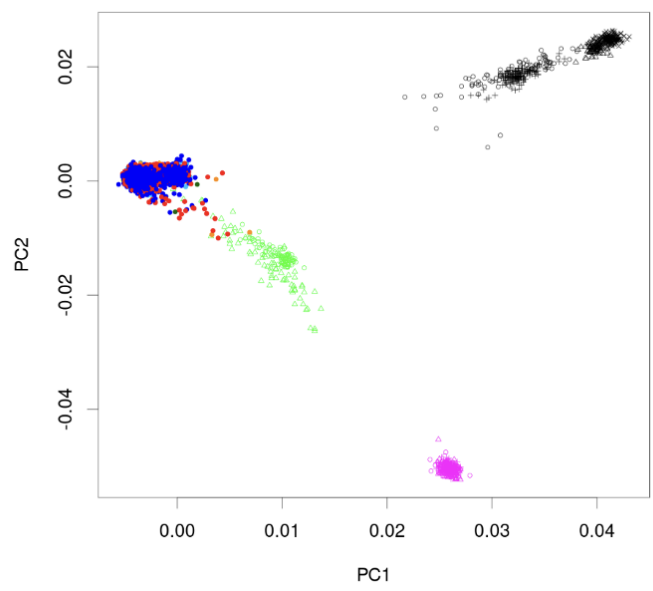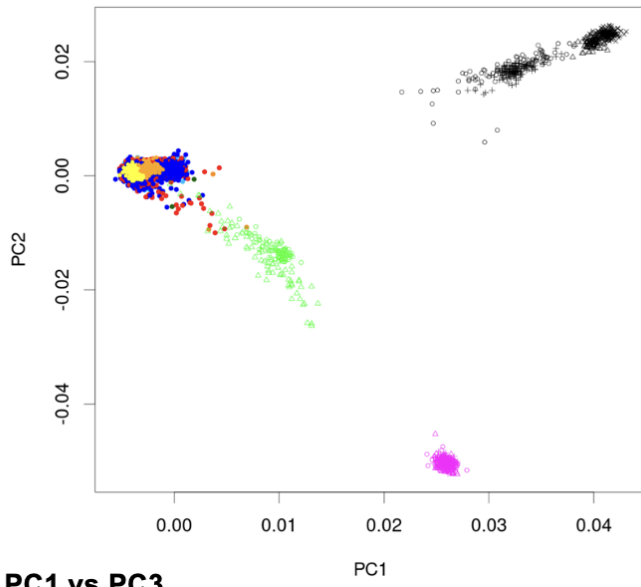
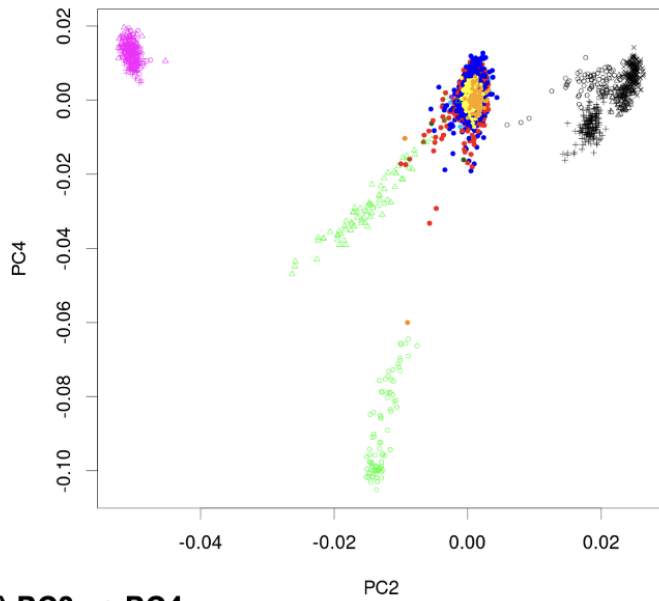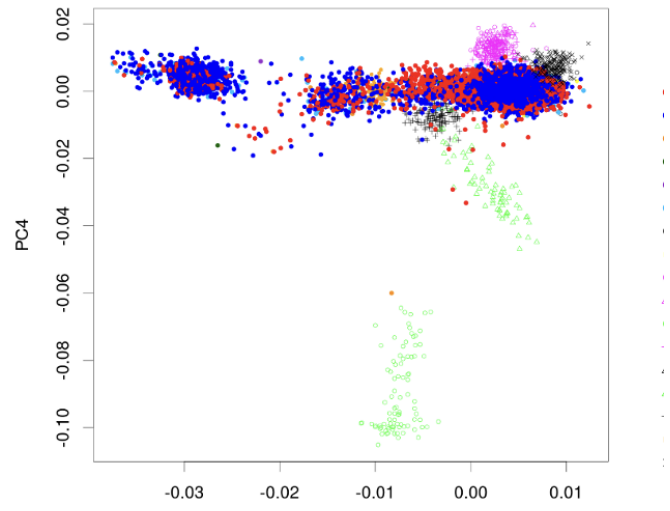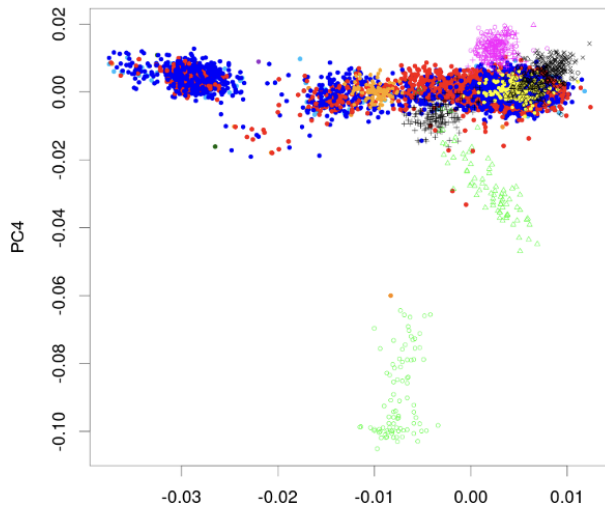**b) PC1 vs PC3**
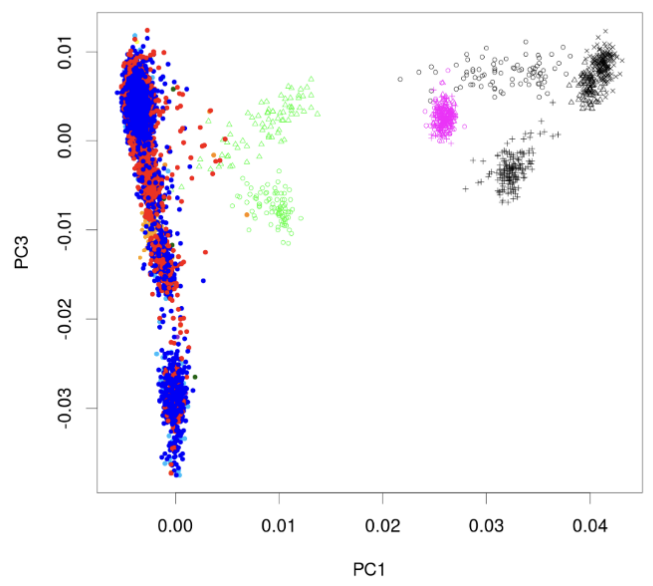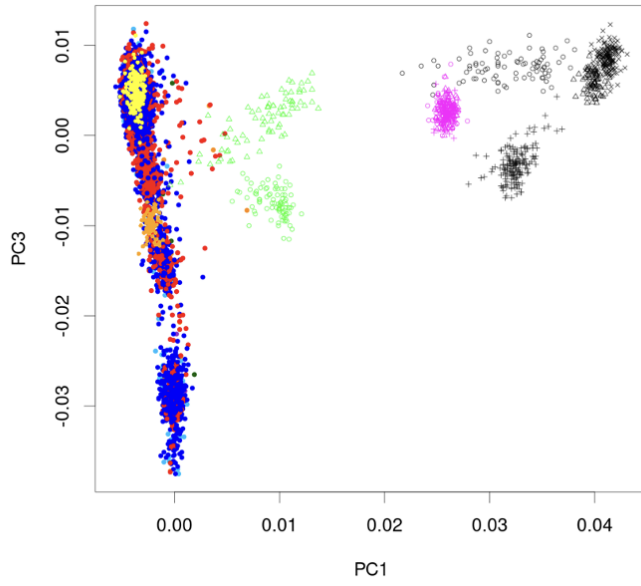
c) PC1 vs PC4
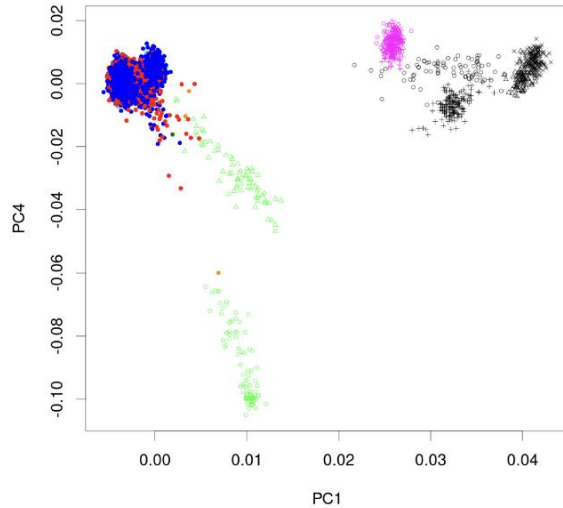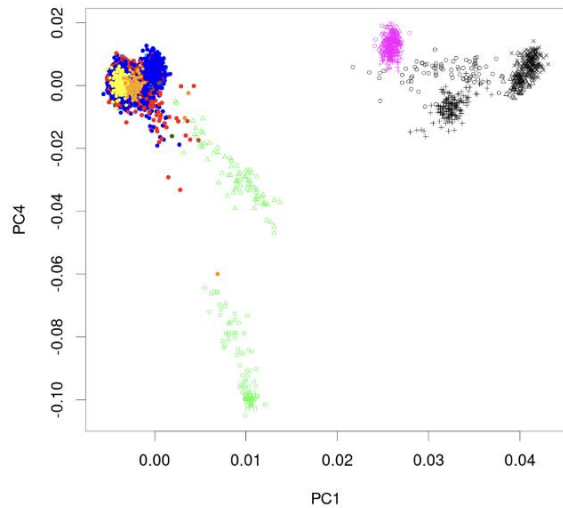
d) PC2 vs PC3
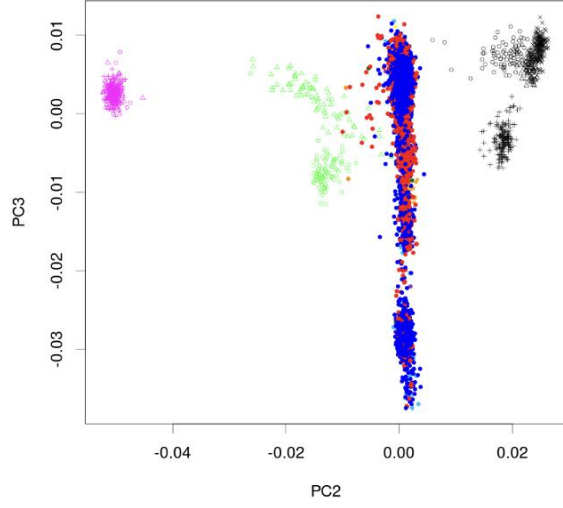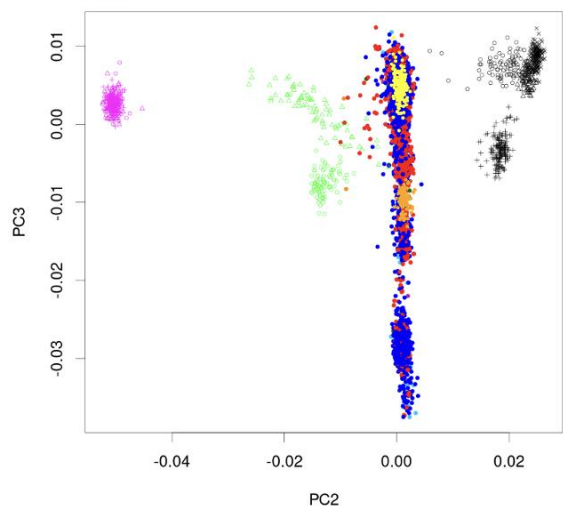
e) PC2 vs PC4

f) PC3 vs PC4

- • UPENN Cases
- • UPENN Controls
- • UCLA Cases
- • UCLA controls
- • MSSM Cases
- • MSSM Controls
- ○ ASW - African ancestry in Southwest USA
- CEU - Utah residents with Northern and Western European ancestry from the CEPH collection
- ○ CHB - Han Chinese in Beijing, China
- △ CHD - Chinese in Metropolitan Denver, Colorado
- ○ GIH - Gujarati Indians in Houston, Texas
- + JPT - Japanese in Tokyo, Japan
- △ LWK - Luhya in Webuye, Kenya
- △ MEX - Mexican ancestry in Los Angeles, California
- + MKK - Maasai in Kinyawa, Kenya
- • TSI - Toscani in Italia
- × YRI - Yoruba in Ibadan, Nigeria

Supplementary Figure 16. Plots of Principal Components against 1,000 Genomes (1kG) Reference Samples. Shown are plots of PC1 vs. PC2 (a), PC1 vs. PC3 (b), PC1 vs. PC4 (c), PC2 vs. PC3 (d), PC2 vs. PC4 (e), and PC3 vs. PC4 (f) for the combined dataset. All population substructure outliers have been excluded. The figure legend including shape representations and coloring are shown after the plots (g). Plots on the left depict case-control samples in the background to allow overlapping reference samples to be distinguishable while plots on the right depict case-control samples in the foreground above reference samples.

4. The authors conducted several integrative analyses, including colocalization, eQTL and differential gene expression. However, it seems that the authors presented the results of each analysis separately. Combining these results and perform a gene prioritization will provide useful information. If a gene receives support from GWAS, colocalization, eQTL and differential gene expression, then this gene is more likely to be causal. I suggest the authors perform a prioritization analysis, as described previously (PMID: 37945807, PMID: 33589840).

> We thank the reviewer for the comment and agree this is an excellent point. This has been added per the comment made from Reviewer 1.

5. The authors carried out differential expression analysis on the 16 candidate genes identified in the significant GWAS loci. This is a nice analysis. However, given colocalization could not tell us if predicted expression of the GWAS identified genes were associated with PSP. I suggest the authors performing a transcriptome-wide association study (TWAS), which will tell us if genetically predicted expression of these genes are associated with PSP. Based on the TWAS, we can know if elevated or decreased expression of a gene is associated with PSP. Differential expression analysis can validate the results of TWAS. By doing so, the authors can integrate different results and draw more reliable conclusions.

> We thank the reviewer for the comment. We have now performed TWAS using two publicly available panels of TWAS models from the prefrontal cortex. We have added the data as a supplemental figure and added language to the results and discussion. We have updated our summary table to include these findings.
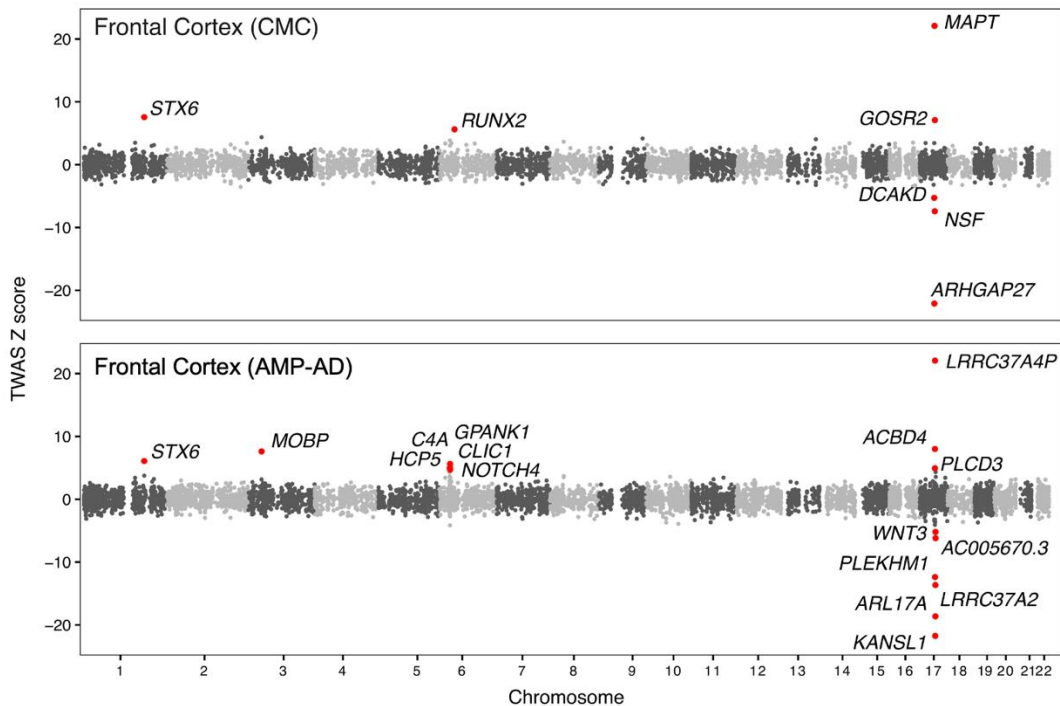
Supplementary Figure 10: Transcriptome-wide Association Study (TWAS) of PSP. Gene expression models were used from two studies of the dorsolateral prefrontal cortex. Upper panel: CommonMind Consortium (n=452). Lower panel: Accelerating Medicines Partnership in Alzheimer's Disease (AMP-AD; n=888). Genes highlighted in red pass Bonferroni-adjusted *P*<0.05.

The following text was added:

Methods:
The GWAS summary stats were first converted to Z-scores using munge_sumstats.py from the LDSC toolkit. Panels of pre-computed TWAS weights from dorsolateral prefrontal cortex samples as part of the CommonMind Consortium (n=452) and the AMP-AD project (n=888) were downloaded from their respective sources. We used an existing 1000 Genomes European LD reference mapped to the hg19 build. TWAS estimates cis-SNP heritability (all SNPs 1Mbp from gene) for each gene then imputes expression in the GWAS to identify associations between gene expression and disease risk. Each gene was given a *z*-score and *P*-value. *P*-values were adjusted for multiple testing within each panel using the Bonferroni method. Genes were called significant at an adjusted *P*<0.05.

Discussion:
Additionally, we ran a transcriptome-wide association study (TWAS) analysis using genetically predicted expression models from two cohorts of dorsolateral prefrontal cortex: the CommonMind Consortium, and the AMP-AD Project. By associating genetically predicted gene expression with our PSP GWAS, we identified increased cortical expression of *STX6*, *RUNX2* and *MOBP* were associated with increased risk of PSP (Bonferroni-adjusted *P*<0.05; Supplementary Figure 4). TWAS prioritized multiple genes in the 6p21.32 and 17q21.32 loci, including *C4A*, *C4B*, and *MAPT*."

Reviewer #4 (Remarks to the Author):

The authors of "Genetic, transcriptomic, histological, and biochemical analysis of progressive supranuclear palsy implicates glial activation and novel risk genes" report some potentially interesting genetic findings in the PSP genetic space. Some comments and suggestions below:

Data and code availability
Its very <u>disappointing to see that the authors at the Data availability section use a lot of publicly available data, but fail to make their own PSP summary statistics and underlying genotype data available</u>. Similarly with the code availability, just looking at Supplementary Figure 1, seems very complex and important, but "Code availability statement" says => "All software used is publicly available at the URLs or references cited." isn't right. Proper code is needed to see what exact parameters, flows etc was used here.

> We thank the reviewer and have uploaded the summary statistics to NIAGADS (NG00169) and will be publicly available in mid June. We added the line "Specific parameters for the code used in this paper can be found at https://github.com/jackhump/PSP_GWAS" to view the specific details of the code. Additional details on the GWAS pipeline have been added in the form of a supplemental figure (18)

Figure 1b, it would be better to color by the 4 cell types. In addition, given the very modest p-value that does not pass multiple test correction, there isn't enough evidence to say "Integration with cell type-specific epigenomic annotations revealed a unique oligodendrocytic signature that distinguishes PSP from AD and Parkinson's disease."

> We thank the reviewer and have modified the figure to be colored by the 4 cell types.
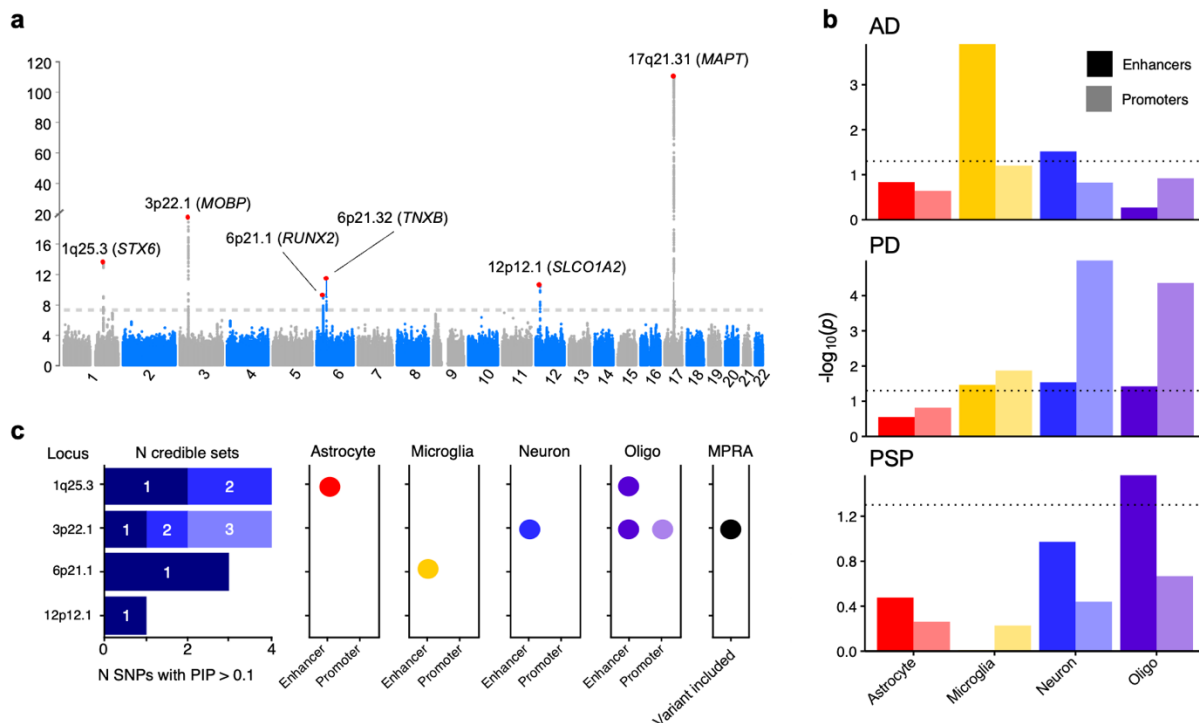
Additionally, we have changed the language of the abstract to "Integration with cell type-specific epigenomic annotations revealed a modest oligodendrocytic signature that might distinguish PSP from AD and Parkinson's disease in subsequent studies with larger sample sizes."

Analytical approaches

1) I didn't see any <u>conditional analyses</u> based on the 6 GWAS hits to identify if there are secondary signals in any of the regions

> We thank the reviewer and have performed these conditional analyses using PLINK and conditions on the lead SNP and have added this analysis to the supplemental figures and added text in the discussion based on the new results. No new independent signals were observed.

"Conditional analysis using the lead SNP in each loci did not reveal any secondary signals in each loci (Supplementary Figures 5-9)."
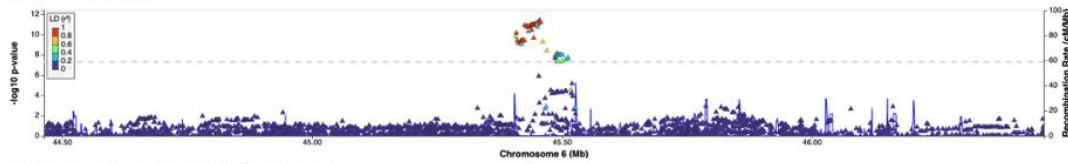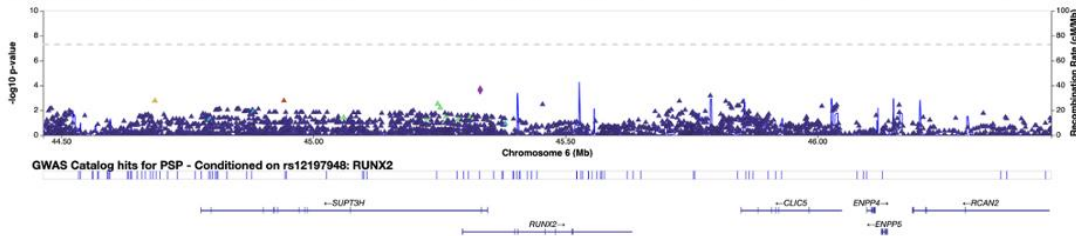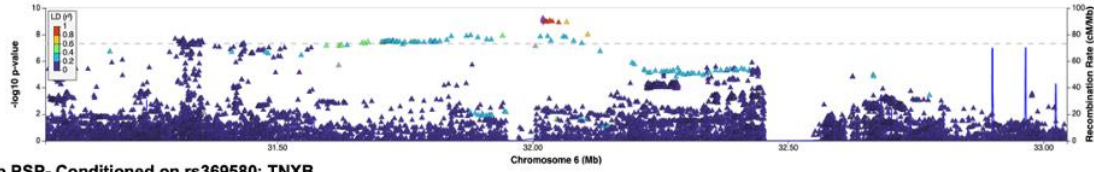
a PSP- Main analysis
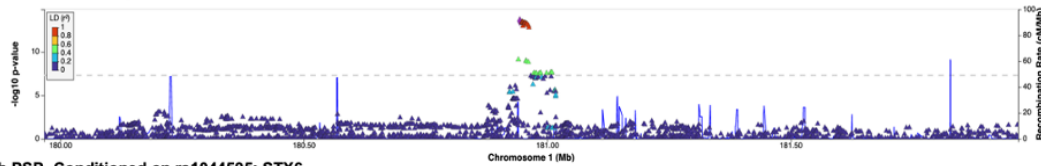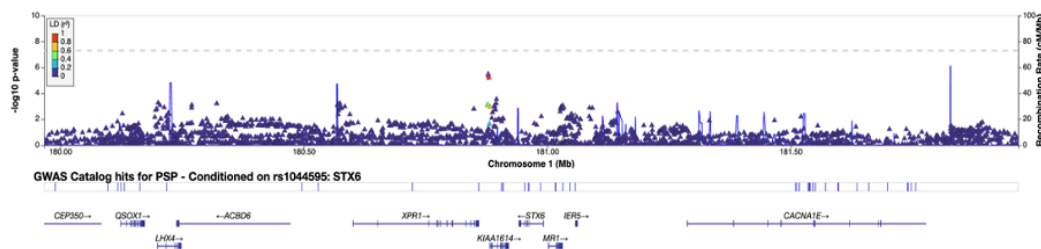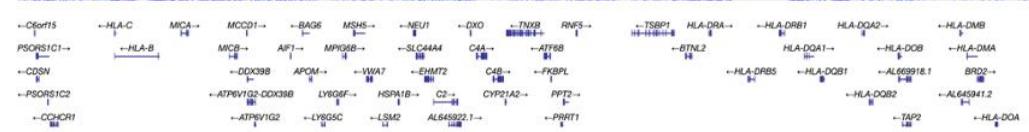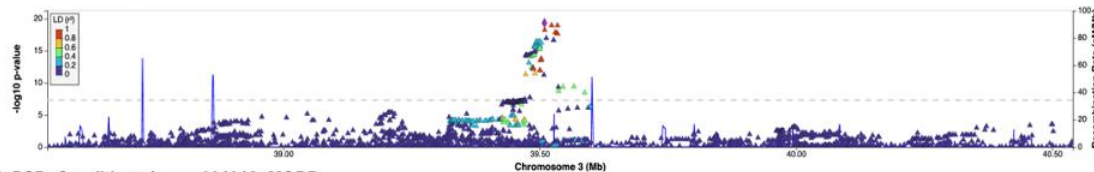

b PSP- Conditioned on rs369580: TNXB
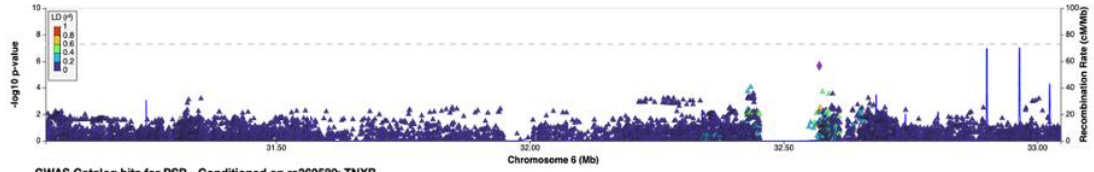
GWAS Catalog hits for PSP - Conditioned on rs369580: TNXB

a PSP- Main analysis
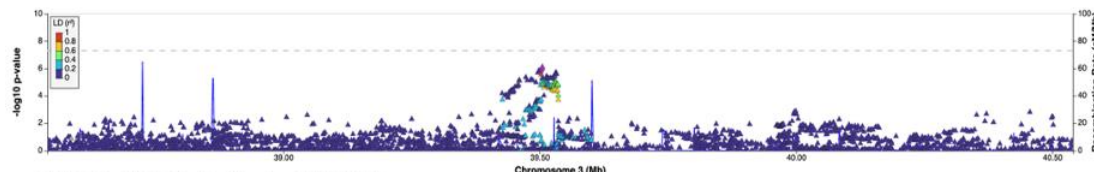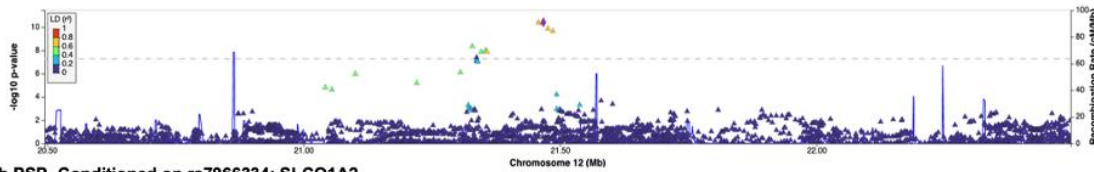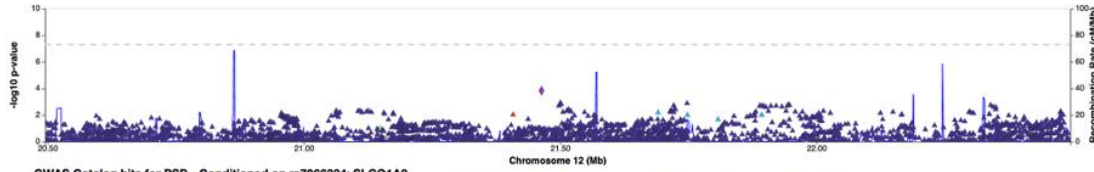

b PSP- Conditioned on rs631312: MOBP

GWAS Catalog hits for PSP - Conditioned on rs631312: MOBP

a PSP- Main analysis

b PSP- Conditioned on rs7966334: SLCO1A2

GWAS Catalog hits for PSP - Conditioned on rs7966334: SLCO1A2

2) a <u>stratified GWAS based on the MAPT</u> is needed here to see if the MAPT is independent of the other 5 signals.

> **We thank the reviewer and have rerun the analysis using a dose dependent covariate for the MAPT H1 and H2 haplotype. The data is now included as a supplemental figure and language has been added in the discussion.**

"The addition of the *MAPT* haplotype status as a covariate reduced the signal observed in the 17q21.31 loci (P=2.28×10^{-17} vs. P=1.94×10^{-110}) but did not modify the five other signals (Supplementary Figure 10)"

Language was added to the methods explaining this analysis:

"Conditional analysis was performed by conditioning the association on the lead SNP in the locus using PLINK and the dose dependent analysis of the *MAPT* sub haplotype (*i.e.,* number of H1 alleles) was performed by adding this variable as a covariate into the main model."
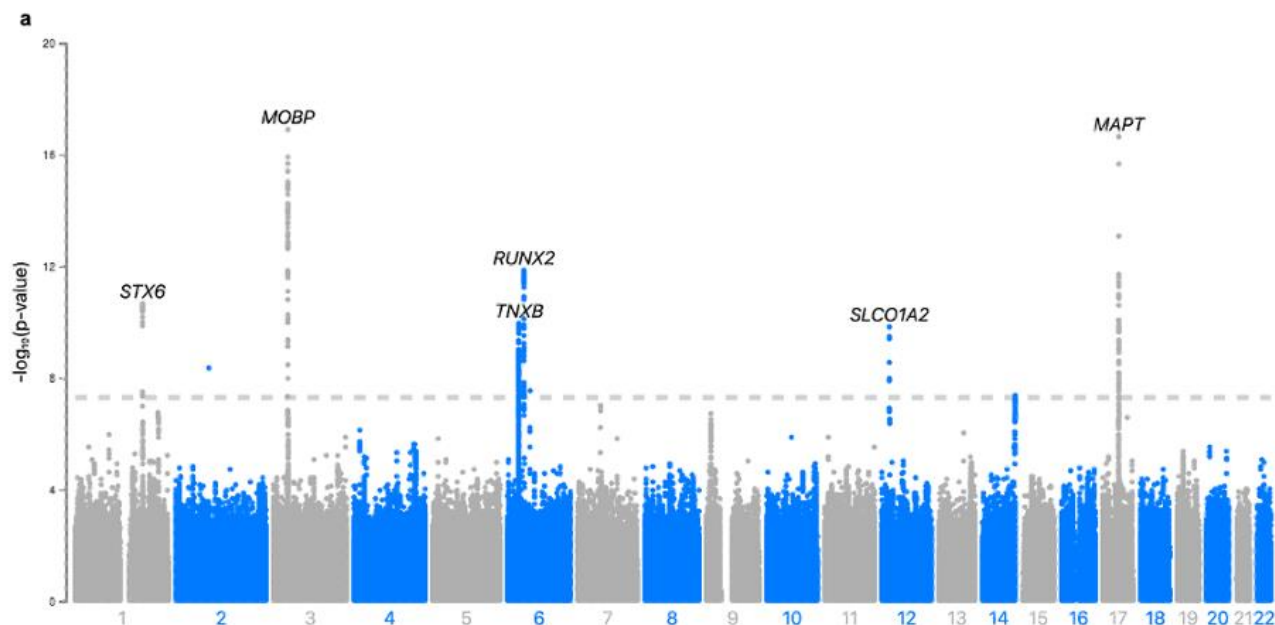
3) Is there no clinical PSP cohort available that can be used as a replication cohort?

> **We have exhausted our resources and were unable to obtain a replication cohort (even clinically defined) given the lack of any external datasets. We are currently in the process of curating this type of cohort for future analysis in a partnership with CurePSP.**

Differential gene expression section
"The data suggest regionally specific changes of multiple genes identified in loci identified from the GWAS data which may have downstream effects on disease relevant protein expression." The main issue here that this is bulk data, so could these differences be simply a difference in cell composition between disease and non disease?

> **We thank the reviewer and agree with this important point and have added the following language to warrant caution in the data interpretation.**

"The data suggest regionally specific changes of multiple genes identified in loci identified from the GWAS data which may have downstream effects on disease relevant protein expression, however

these differences may be attributed to a difference in cell composition between cases and controls and further single cell analysis studies are warranted."

C4A, 1) is the signal the same signal as ALS? Unclear based on currently provided data. 2) What about the association between C4A and Alzheimer?

> We thank the reviewer and have clarified this point. The ALS signal is in the HLA region and language has been modified to reflect this. The lead SNP (rs9275477) is not the same lead SNP found in our PSP GWAS (rs369580). The AD SNP for *C4A* again is not identical to our lead SNP (rs9271171), and again was contained more in the HLA region, however functional validation in an iPSC CRISPR model demonstrated its effect on *C4A* expression. To the reviewers point we did observe a modest signal in rs9275477 in our GWAS (P=0.0041), however rs9271171 was filtered out during QC. To clarify the point the reviewer has made we have added the following line to conclude this section of the discussion:

"In Summary, although the genetic signals (*e.g.,* lead SNP) differ in these genetic studies compared to the PSP genetics presented here, these findings underscore the importance of exploring the role of innate immune interactions and oligodendrocyte pathology in the pathogenesis of multiple neurodegenerative conditions."

Minor comments:
"autopsy studies have found PSP pathology in 2-6% of individuals with no PSP diagnosis prior to death, suggesting that it is more prevalent that appreciated in living individuals3-5."
Worth to clarify is no PSP diagnosis means here, NO neurological disease or something else?

> We thank the reviewer and have clarified this point adding that there is no "neurological" diagnosis in the introduction.

"Given the sharing of common tau pathology across multiple neurodegenerative diseases" what other diseases have tau pathology?

> We thank the reviewer and have added the following line for clarification "tau pathology across multiple neurodegenerative diseases (*i.e.* AD, corticobasal degeneration, chronic traumatic encephalopathy and others)"

"Analyses were then performed in GRCh37/hg19 using the INFERNO and SparkINFERNO pipelines which are detailed elsewhere" how was the data converted from hg38 TOPMED to hg19?

> We thank the reviewer and have clarified this point in the methods

"Data was converted between genome references using LiftOver and all SNPs were preserved."

"Imputation of C4A and C4B copy number" → is there any validation possible to include here? E.g., pacbio sequencing?

> We thank the reviewer and have performed digital droplet PCR on a subset of 4 PSP cases to demonstrate the robustness of the *C4* copy number imputation software. We used a total of 4 subjects, each containing a unique number of copies of *C4A* outlining the spectrum of copies known to occur.

Only one sample was not identified via droplet PCR to have the same copy numbers identified by imputation. This level of accuracy is within the previously reported documentation provided by the group that designed the imputation software. The results are reported in Supplementary Table 4. The following language has been added

"Validation of the imputation was performed using droplet PCR (*n*=4, each sample with a unique number of *C4A* copies, run in duplicate) using methods outlined elsewhere and was found to be within the previously reported accuracy generally (0.70<*r²*<1.00, Supplementary Table 4)[39]."

"n=10 for cases vs. n=10 for PSP" → what is the difference here between cases and PSP?

> We thank the reviewer and have fixed this typo using the following language.

"(*n*=10 for controls vs. *n*=10 for cases, PSP)"

"of PLA2G7, MOBP, MSH5, HLA-DPB1, HLA-DMB, and SLCO1A2 in PSP versus controls (P<0.0025," based on what was this p-value set?

> We thank the reviewer and have clarified this. Specifically, our type 1 error threshold is alpha = 0.05 / 16, which gets us a Bonferroni-corrected *P*<0.0025 value. The following language was added to the methods.

"Because we looked specifically at 16 genes contained within five significant loci, a *P*<0.05/16=0.0025 was considered differentially expressed based on a Bonferroni correction for multiple comparisons."

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have addressed my comments.

Reviewer #2 (Remarks to the Author):

The authors have satisfactorily addressed all the points raised in the initial review.

Reviewer #3 (Remarks to the Author):

The authors have adequately addressed my comments and concerns.

Reviewer #4 (Remarks to the Author):

The authors did a great job with all the revisions very impressive work. My one minor remaining comment is where is the underlying genotype data, is that somewhere available as well? Assume it has to be deposited to NIAGADS as well, due to NIA's mandatory data sharing regulations?

REVIEWER COMMENTS (SECOND RESPONSE)

Reviewer #1 (Remarks to the Author):

The authors have addressed my comments.

> We thank the reviewer for positive feedback

Reviewer #2 (Remarks to the Author):

The authors have satisfactorily addressed all the points raised in the initial review.

> We thank the reviewer for positive feedback

Reviewer #3 (Remarks to the Author):

The authors have adequately addressed my comments and concerns.

> We thank the reviewer for positive feedback

Reviewer #4 (Remarks to the Author):

The authors did a great job with all the revisions very impressive work. My one minor remaining comment is where is the underlying genotype data, is that somewhere available as well? Assume it has to be deposited to NIAGADS as well, due to NIA's mandatory data sharing regulations?

> We thank the reviewer for positive feedback. We are actively working on getting the full genetic dataset available on NIAGADS. However, because we collected tissue used to generate the data from over 45 domestic and international brain banks, and we need to complete a genomic data sharing agreement with each center to comply with NIH data sharing/privacy policy. Portions of the data are already available NIAGADS (NG00037). It is our goal to have the full harmonized dataset available within a reasonable amount of time, albeit this is dependent on the time it takes each center to complete these new forms. In the mean time we have made the summary statistics available.