BRAIN
ORIGINAL ARTICLE

Genetic risk factor clustering within and across neurodegenerative diseases

Mathew J. Koretsky,[1,2] Chelsea Alvarado,[1,2,3] Mary B. Makarious,[2,4] Dan Vitale,[1,2,3] Kristin Levine,[1,2,3] Sara Bandres-Ciga,[1] Anant Dadu,[1,3,5] Sonja W. Scholz,[6,7] Lana Sargent,[1] Faraz Faghri,[1,2,3] Hirotaka Iwaki,[1,2,3] Cornelis Blauwendraat,[1,2] Andrew Singleton,[1,2] Mike Nalls[1,2,3] and Hampton Leonard[1,2,3,8]

See Jain *et al.* (https://doi.org/10.1093/brain/awad337) for a scientific commentary on this article.

Overlapping symptoms and co-pathologies are common in closely related neurodegenerative diseases (NDDs). Investigating genetic risk variants across these NDDs can give further insight into disease manifestations.

In this study we have leveraged genome-wide single nucleotide polymorphisms and genome-wide association study summary statistics to cluster patients based on their genetic status across identified risk variants for five NDDs (Alzheimer's disease, Parkinson's disease, amyotrophic lateral sclerosis, Lewy body dementia and frontotemporal dementia). The multi-disease and disease-specific clustering results presented here provide evidence that NDDs have more overlapping genetic aetiology than previously expected and how neurodegeneration should be viewed as a spectrum of symptomology. These clustering analyses also show potential subsets of patients with these diseases that are significantly depleted for any known common genetic risk factors suggesting environmental or other factors at work.

Establishing that NDDs with overlapping pathologies share genetic risk loci, future research into how these variants might have different effects on downstream protein expression, pathology and NDD manifestation in general is important for refining and treating NDDs.

1 Center for Alzheimer's Disease and Related Dementias, National Institutes of Health, Bethesda, MD 20892, USA
2 Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD 20892, USA
3 Data Tecnica International LLC, Washington, DC 20037, USA
4 UCL Movement Disorders Centre, University College London, London, WC1E 6BT, UK
5 Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA
6 Neurodegenerative Diseases Research Unit, National Institute of Neurological Disorders and Stroke, Bethesda, MD 20892, USA
7 Department of Neurology, Johns Hopkins University, Baltimore, MD 21287, USA
8 DZNE, Tuebingen 72076, Germany

Correspondence to: Mathew Koretsky
Center for Alzheimer's and Related Dementias
9000 Rockville Pike, T44, Bethesda, MD 20892, USA
E-mail: koretskymj@nih.gov

Keywords: dementia; single-nucleotide polymorphism; genome-wide association study; unsupervised; machine learning

Received November 18, 2022. Revised April 11, 2023. Accepted April 26, 2023. Advance access publication May 16, 2023

Introduction

Neurodegenerative diseases (NDDs), such as Alzheimer's disease (AD), Parkinson's disease (PD), amyotrophic lateral sclerosis (ALS), Lewy body dementia (LBD) and frontotemporal dementia (FTD), collectively affect more than 40 million people worldwide.[1] This number is only expected to grow due to their mid to late-life onset combined with an ageing population.[1] Genome-wide association studies (GWAS) have been used to identify common genetic variants linked to a higher risk of developing certain NDDs to uncover pathways that can lead to more advanced and targeted treatments.[2,3]

Established risk variants for one NDD may play a role in the genetic makeup of several others. Overlapping risk variants across NDDs, even when subgenome-wide significant in GWAS studies of a specific disease, may give insights into how disease can manifest across the spectrum of symptoms and co-pathologies shared by multiple NDDs.[4,5] Evidence of pleiotropic effects have been described previously in *GRN* for AD, PD, LBD and ALS, *GBA* for PD and LBD, and *C9orf72* for FTD and ALS.[6-8] Mutations in *MAPT* and *APOE* have also been linked to a range of NDDs and pathologies.[9,10] Additionally, it is possible to investigate how high and low genetic risk may manifest within a single disease, partitioning individuals with an NDD into classes across a quantitative range of genetic influence.

Here, we used genome-wide single nucleotide polymorphism (SNP) and GWAS summary statistics data for five NDDs (AD, PD, ALS, LBD and FTD) to cluster patients based on their genetic status across identified risk variants for each disease. The multi-disease clusters presented here establish data-driven evidence of potential shared disease aetiology that may explain overlapping symptomatology on the molecular level.[11] The single-disease cluster analysis supports the idea that NDD-specific risk data can inform both genetically different subtypes within the same disease and identify patients that may have the disease due to environmental, epigenetic or other risk factors. This work seeks to refine NDD phenotypes and could help to differentiate NDDs for diagnosis and clinical trial enrolment.

Materials and methods

Samples

Supplementary Fig. 1 summarizes the workflow and data used for this project. The samples were obtained from public domain whole-genome sequencing (WGS) cohorts across the aforementioned NDDs. Existing genotype calls from the WGS cohorts were used and no additional genotype calling was performed. AD samples were obtained from the Alzheimer's Disease Sequencing Project (ADSP), Alzheimer's Disease Neuroimaging Initiative (ADNI), Mayo RNAseq Study (MayoRNAseq), Mount Sinai Brain Bank (MSBB), and the Religious Orders Study/Memory and Aging Project (ROSMAP).[12-14] PD samples were obtained from Accelerating Medicines Partnership Parkinson's disease (AMP-PD).[15] FTD, LBD and ALS data were all obtained from DementiaSeq.[16] The total number of subjects across all cohorts was 23 885, of which 13 190 were cases (Supplementary Table 1). Only samples of genetically determined European ancestry were used. For this analysis, 1000 cases for each disease were randomly sampled to ensure even representation, resulting in a final sample size of 5000 cases (Supplementary Table 1). GWAS summary statistics were obtained for each disease for use in the final SNP selection. The GWAS summary level data used include Schwartzentruber *et al.*[17] (AD), Nalls *et al.*[18] (PD), Nicolas *et al.*[19] (ALS), Chia *et al.*[20] (LBD) and Ferrari *et al.*[21] (FTD).

Genetic data quality control

Data from cohorts were not all on the same build; thus, data from cohorts using the hg19 build were lifted over to hg38.[22] Summary statistics were lifted over as needed. Quality control (QC) was performed at both the individual cohort and combined cohort levels with Global Parkinson's Genetics Program (GP2) pipelines (https://github.com/GP2code/) using PLINK (1.9 and 2). Sample level QC included genotype missingness (<0.02) as well as a duplicate removal [genetic relatedness matrix (GRM) cut-off of 0.95] and first cousin or closer relatedness pruning (GRM cut-off of 0.125). ADNI, ROSMAP and AMP-PD also underwent genetic sex confirmation due to the availability of data for the X chromosome. Variant level QC included call rate pruning (<0.05) and pruning SNPs with a minor allele frequency (MAF) < 0.05 for exclusion (Supplementary Table 2). After common SNPs were identified across the cohorts (explained in the 'Systematic review' section), the merged genotype data underwent an additional duplicate and relatedness check. The merged data were then passed through an ancestry prediction and pruning method to ensure all samples were of European descent (Supplementary Table 3). Ancestry was defined using reference panels from the 1000 Genomes Project and an Ashkenazi Jewish Population.[23,24] Fifty principal components were fit on 39 302 overlapping SNPs between the reference panel and the merged data. A classifier that was previously trained on the reference panel principal components was applied to the merged data principal components, which returned the predicted ancestry label of each individual in the merged data. Samples predicted as non-European were removed.

Systematic review

Prior to clustering across the NDDs, we narrowed the number and scope of SNPs. Seventeen million autosomal SNPs sequenced across all cohorts were identified and used to merge the individual cohort data. Disease-specific GWAS data were then used to filter for SNPs that reached genome-wide significance (i.e. $P < 5 \times 10^{-8}$) in any one of the relevant GWAS studies. Using this SNP set, the merged data underwent munging and additional population substructure adjustment in GenoML.[25,26] Munging consists of pruning provided genotype data for linkage disequilibrium (LD) by removing any highly correlated genotypes in the sample series ($r^2 > 0.3$ within a sliding window of 1 Mb as minimum exclusion criteria). LD clumping and pruning was performed at random and was therefore not biased towards associations from any one disease. The adjustment process removes the effect of population substructure, which is further described by Makarious *et al.*[26] The process required creating principal component analysis (PCA) loadings using the 5000 downsampled cases. Unlinked genome-wide SNPs outside of GWAS regions of interest were used to generate the 10 PCA loadings that approximate population substructure. The resulting 10 PCA loadings were used as covariates and regressed against the final SNP candidates using ordinary least squares regression. The resulting residual minor allele dosages were then *z*-normalized and used as the final output for model training at clustering. This process limits the effect of European population substructure from the genotypes before the clustering analysis is performed, mirroring the way in which GWAS uses covariates for population substructure within ancestry groups to reduce genomic inflation.[26] After munging, the final SNP set consisted of 338 GWAS significant and population substructure adjusted SNPs not in LD with each other.

Statistical analyses

Supplementary Fig. 2 summarizes the dimensionality reduction and clustering analyses performed. To effectively visualize and cluster the adjusted SNPs, Unified Manifold Approximation and Projection (UMAP) was chosen for dimensionality reduction. UMAP is a non-linear approach that is widely used in the field of population genetics.[27] Using UMAP, 338 SNPs were reduced to three dimensions for each of the 5000 cases. Unsupervised clustering of the individuals was performed on the reduced data using the mean shift algorithm, as it is a deterministic algorithm that does not require the number of clusters to be specified, unlike more popular approaches such as K-means clustering that require an *a priori* number of clusters to be defined.[28]

UMAP employs two fine-tuning hyperparameters, a and b, that impact the resulting embedding more specifically than minimum distance and spread. UMAP is a flexible algorithm that can be used on many different types and sizes of data; fine-tuning these hyperparameters enhances the model adjustment to the SNP data. Performing a grid search of a and b values from 0.25 to 3, with a step size of 0.25, the UMAP to mean shift pipeline was fitted and applied on a 70:30 (training:testing) split. To determine the best combination of hyperparameters, logistic regression was used with cluster membership as the input to predict an individual's disease status for each NDD (AD, PD, ALS, LBD and FTD). The chosen evaluation metric was the average area under the receiving operating characteristic (ROC) curve (AUC) across the disease-cluster regressions. The hyperparameters with the highest average AUC across the 144 tested combinations were then identified and used throughout the analysis ($a = 2.75$, $b = 0.75$).

UMAP is a stochastic algorithm; different runs can produce different results despite the input data and hyperparameters being the same. This can cause mean shift to identify a variable number of clusters in different iterations. To investigate this phenomenon, the UMAP to mean shift analysis was run on 15 different 70:30 (training:testing) splits for 100 iterations (i.e. 1500 iterations total), recording the number of clusters identified and the sample counts per cluster. Across the different splits and iterations, mean shift consistently identifies the main cluster that contains the majority (>4000 out of 5000 individuals) of the samples (Supplementary Table 4). From there, we applied an iterative clustering approach. Tracking samples across iterations, any sample consistently grouped into the main cluster was identified and labelled as a member of Cluster 0 (C0). Conversely, any sample that was never grouped into this main cluster was labelled as a member of Cluster 2 (C2). All the remaining samples that were not always grouped into the main cluster, according to the UMAP embedding, were labelled as a member of Cluster 1 (C1). This process effectively addresses the variability caused by the stochastic nature of UMAP while capturing the clustering information provided by mean shift. More information on the iterative clustering approach can be found in the Supplementary material, 'Methods' section.

A z-test for proportions was performed to determine which multi-disease clusters were significantly enriched with certain NDDs compared to others. Next logistic regressions were used to see how cluster membership relates to NDD status as a complement to the previously described enrichment analysis. Cluster memberships were regressed on the set of 338 adjusted SNPs to identify any potential SNPs associated with increased likelihood of membership in a particular cluster, in part as a positive control, for loci with established pleiotropic associations (*GBA*, *GRN*, *LRRK2*, *MAPT*, *C9orf72* and *APOE*). The Shapley values of the SNP predictors were then calculated, which is a popular game theory approach that helps explain and interpret how important each feature in a model is to the prediction of the dependent variable, in our case, cluster membership.[29] SNPs most important in determining cluster membership were tracked across the clustering iterations to ensure consistency and a phenome-wide association study (PheWAS) look-up was performed to further clarify the biological impact of the clusters.[30] Finally, for each sample in the dataset, a polygenic risk score (PRS) was calculated for all five NDDs with the beta values from the same GWAS summary statistics used to determine the SNP set. From there, logistic regression was run to see how well the disease-specific PRSs determine cluster membership. Note that the PRSs were z-score normalized to simplify the interpretation of the logistic regression output.

The same UMAP hyperparameter combination was used to perform the dimensionality reduction for each NDD in the disease-specific cluster analysis. Similar to the multi-disease analysis, when the UMAP to mean shift pipeline was applied to the 1000 samples for each NDD separately, the main cluster that contained a majority (>500 out of 1000 individuals) was formed consistently across iterations and diseases. Therefore, the iterative clustering approach from the multi-disease analysis was once again used. To account for the increased variability that comes with running the pipeline on a reduced sample size, individuals were grouped into C0 or C2 (C1 if only two clusters were identified in a disease subset) if they were consistently inside or outside the main cluster for at least 12 of the 15 70:30 (training:testing) splits (Supplementary material, 'Methods' section). The previously calculated PRS were re-normalized using the mean and standard deviations (SD) from the set of 1000 samples for each NDD. A *T*-test was run to see where the PRSs differentiated significantly from the mean of 0 that they were standardized to within the NDD-specific clusters.

To validate the methodology used and compare results, the clustering analysis was rerun under four different conditions. First, the iterative clustering approach was compared to agglomerative hierarchical clustering on the same set of 5000 cases. Similar to mean shift, this popular tree-based approach does not require an *a priori* number of clusters to be defined.[31] Next, the iterative clustering analysis was run with the *APOE* locus removed to ensure the clusters are not directly driven by E4 status due to its importance in determining NDD and AD risk. Next, the analysis was performed on a downsampled set of 500 cases for each disease to test the robustness of the results to sample size. Finally, a set of 1000 controls were included in the analysis to see how results are affected by the presence of a negative control. Note that for the downsampled and control-included analyses, QC and processing was exactly the same for the 2500 and 6000 samples, respectively. In addition, linkage disequilibrium score regression (LDSC) was run on the GWAS summary statistics to compare the genetic overlap seen in the clustering results to a more traditional method.

Data and code availability

All samples for this analysis were obtained from public domain WGS cohorts. A repository containing all code for processing and analysis is publicly available to facilitate replication (https://github.com/NIH-CARD/NDD_risk_variant_clustering). In addition, an interactive website has been developed where researchers can further explore the described cluster memberships and results (https://nih-card-ndd-risk-variant-clustering-app-25rr5g.streamlitapp.com/).

Results

Multi-disease clustering

Using the iterative clustering approach, C0 contained 2863 samples, C1 contained 2074 samples and C2 contained 63 samples (Fig. 1A and Supplementary Fig. 3). Regressions of each disease status per sample against cluster membership revealed that C0 was most significantly enriched with ALS [odds ratio (OR) = 1.631, $P = 4.66 \times 10^{-8}$, beta = 0.489, standard error (SE) = 0.090], C1 with AD (OR 1.637, $P = 9.20 \times 10^{-9}$, beta 0.493, SE 0.086), and C2 with FTD (OR = 3.063, $P = 6.50 \times 10^{-5}$, beta = 1.119, SE = 0.280). C2 was enriched with PD compared to the overall disease distributions but not to the point of significance in the regressions (Table 1). After multiple test corrections, none of the clusters were significantly enriched with LBD.

The PRS regressions revealed that the only PRS that was significantly associated with membership in all clusters was AD. C0 and C2 had negative associations (defining non-AD driven clusters), while C1 had a positive association. The trend of C0 having significant negative associations and C1 having significant positive associations continued for the PD, ALS and LBD PRS, while no other PRS was shown to be significantly associated with C2. The FTD PRS was not significantly associated with membership in any of the clusters. For a summary of these results, refer to Table 2. The PRS distributions by cluster are displayed in Fig. 2.

Based on the Shapley values for individual SNPs, important variants determining membership in C0 and C1 were localized to *APOC1* (rs72654445) and *CEACAM16/AS1* (rs112952132 and rs111278137) for C2 (Fig. 1B–D). Two SNPs in *NECTIN2* (rs41290102 and rs79701229) were of high importance for differentiating C0 and C1 from C2. All specified variants, aside from rs41290102, show significant associations with low density lipoprotein cholesterol levels as well as high cholesterol in the Open Targets

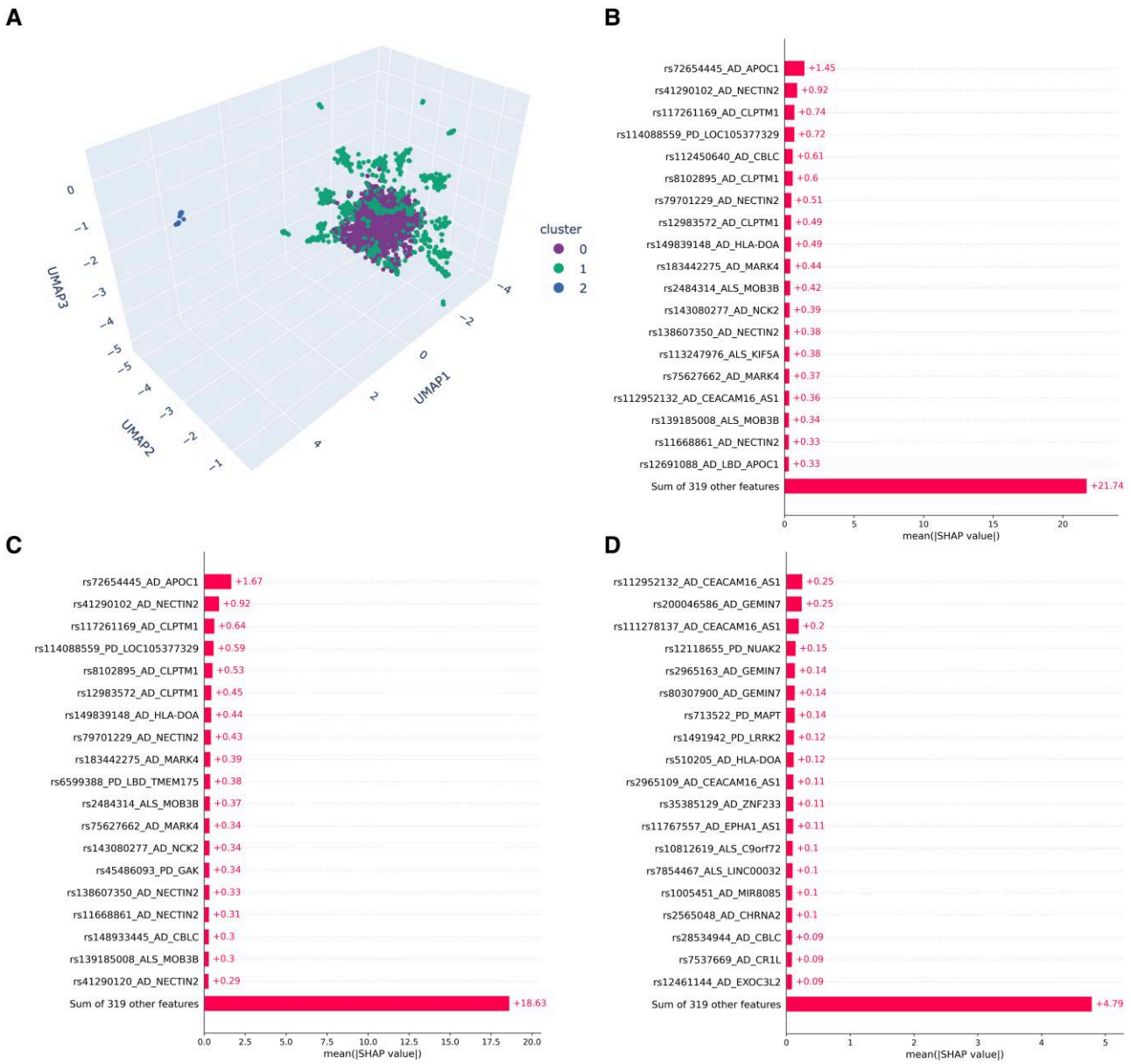

Figure 1 Multi-disease cluster membership. (**A**) Multi-disease clusters. Shapley values of single nucleotide polymorphisms (SNPs) most impacting the defining of (**B**) Cluster 0, (**C**) Cluster 1 and (**D**) Cluster 2.

Table 1 Disease association summary statistics and frequency per multi-disease cluster

Disease	Multi-disease cluster membership	OR	BETA	SE	P	% samples with disease
AD	Cluster 0	0.646	−0.436	0.086	3.51×10^{-7}	0.165[a]
	Cluster 1	1.637	0.493	0.086	9.20×10^{-9}	0.252[a]
	Cluster 2	0.245	−1.408	0.596	0.018	0.079
PD	Cluster 0	1.141	0.132	0.086	0.125	0.210
	Cluster 1	0.859	−0.152	0.087	0.080	0.186
	Cluster 2	1.308	0.269	0.322	0.404	0.238[a]
ALS	Cluster 0	1.631	0.489	0.090	4.66×10^{-8}	0.226[a]
	Cluster 1	0.646	−0.437	0.090	1.15×10^{-6}	0.167[a]
	Cluster 2	0.245	−1.408	0.596	0.018	0.079[a]
LBD	Cluster 0	0.836	−0.179	0.084	0.033	0.190
	Cluster 1	1.198	0.180	0.084	0.032	0.213
	Cluster 2	1.004	4.30×10^{-3}	0.341	0.990	0.206
FTD	Cluster 0	1.021	0.021	0.085	0.802	0.208
	Cluster 1	0.895	−0.111	0.086	0.196	0.182[a]
	Cluster 2	3.063	1.119	0.280	6.50×10^{-5}	0.397[a]

AD = Alzheimer's disease; PD = Parkinson's disease; ALS = amyotrophic lateral sclerosis; LBD = Lewy body dementia; FTD = frontotemporal dementia; OR = odds ratio; SE = standard error.
[a]P-value < 0.05 for the frequency increase or decrease in a certain disease status per cluster compared to the null estimate of 20%.

Table 2 Polygenic risk score association summary statistics per cluster for both the multi-disease and single-disease clustering analyses

Disease/PRS	Multi-disease clustering					Disease-specific clustering				
	Cluster	OR	Beta	SE	P	Cluster	OR	Beta	SE	P
AD	Cluster 0	0.804	−0.218	0.034	2.31×10^{-10}	Cluster 0	0.918	−0.086	0.077	0.266
	Cluster 1	1.331	0.286	0.035	2.32×10^{-16}	Cluster 1	1.117	0.111	0.078	0.154
	Cluster 2	0.146	−1.921	0.249	1.31×10^{-14}	Cluster 2	0.550	−0.598	0.425	0.160
PD	Cluster 0	0.838	−0.177	0.035	3.95×10^{-7}	Cluster 0	0.826	−0.192	0.080	0.017
	Cluster 1	1.209	0.190	0.035	6.38×10^{-8}	Cluster 1	1.211	0.192	0.080	0.017
	Cluster 2	0.829	−0.187	0.144	0.192	–	–	–	–	–
ALS	Cluster 0	0.858	−0.154	0.034	7.00×10^{-6}	Cluster 0	0.577	−0.549	0.093	3.69×10^{-9}
	Cluster 1	1.171	0.158	0.034	4.00×10^{-6}	Cluster 1	1.732	0.549	0.093	3.69×10^{-9}
	Cluster 2	0.950	−0.051	0.142	0.721	–	–	–	–	–
LBD	Cluster 0	0.653	−0.427	0.036	1.31×10^{-32}	Cluster 0	0.627	−0.466	0.081	7.36×10^{-9}
	Cluster 1	1.549	0.437	0.036	5.22×10^{-34}	Cluster 1	1.594	0.466	0.081	7.36×10^{-9}
	Cluster 2	0.872	−0.137	0.148	0.354	–	–	–	–	–
FTD	Cluster 0	0.977	−0.023	0.034	0.498	Cluster 0	0.988	−0.012	0.077	0.872
	Cluster 1	1.029	0.029	0.034	0.402	Cluster 1	1.039	0.038	0.078	0.625
	Cluster 2	0.920	−0.084	0.132	0.528	Cluster 2	0.777	−0.252	0.232	0.278

AD = Alzheimer's disease; PD = Parkinson's disease; ALS = amyotrophic lateral sclerosis; LBD = Lewy body dementia; FTD = frontotemporal dementia; OR = odds ratio; PRS = polygenic risk score; SE = standard error.

PheWAS look up (Supplementary Table 5). Variants belonging to the genes *APOC1*, *CBLC*, *CEACAM16/AS1*, *CLPTM1* and *NECTIN2* were identified as significant drivers in at least one multi-disease cluster based on their mean absolute Shapley value. All these top variants are associated with AD and cluster within 1 Mb of the *APOE* locus on chromosome 19 and likely driven by a connection to variable E4 allele risk.[32] Variants associated with the specified loci only account for 20.99%, 20.97% and 8.90% of the differentiation between C0, C1 and C2, respectively. In addition, of the top 20 variants associated with each cluster, 13 were consistent across clustering iterations for C0 and C1, and nine were consistent for C2 (Supplementary Table 6).

The individual SNP regressions reveal variants localized with *APOE*, *GBA* and *LRRK2* are significantly associated with all clusters (Supplementary Table 7). Additionally, variants localized to *MAPT* (rs713522) and *C9orf72* (rs17696570) showed significant associations with C0 and C1. The *GBA* variant (rs76763715) shows a strong positive association with C1 (OR = 8.020, $P = 0.008$, beta = 2.082, SE = 0.782) and a very strong negative association with C2 (OR = 8.33×10^{-4}, $P = 0.023$, beta = −7.091, SE = 3.110). One of two variants that determine *APOE4* status was in the set of 338 SNPs used for clustering (rs7412). Given its importance in determining NDD and AD risk, it should be noted that the variant itself is not significantly associated with membership in any of the clusters (Supplementary Table 8).

Disease-specific clustering

For the NDD-specific clusters, only AD and FTD had a group of samples that were consistently outside the main cluster across all iterations (i.e. a presence of C2), containing 14 and 25 samples, respectively (Supplementary Fig. 4). Interestingly, across all single-disease clustering analyses, the AD PRS is significantly associated with differentiating subsets of samples (Table 2 and Supplementary Table 9).

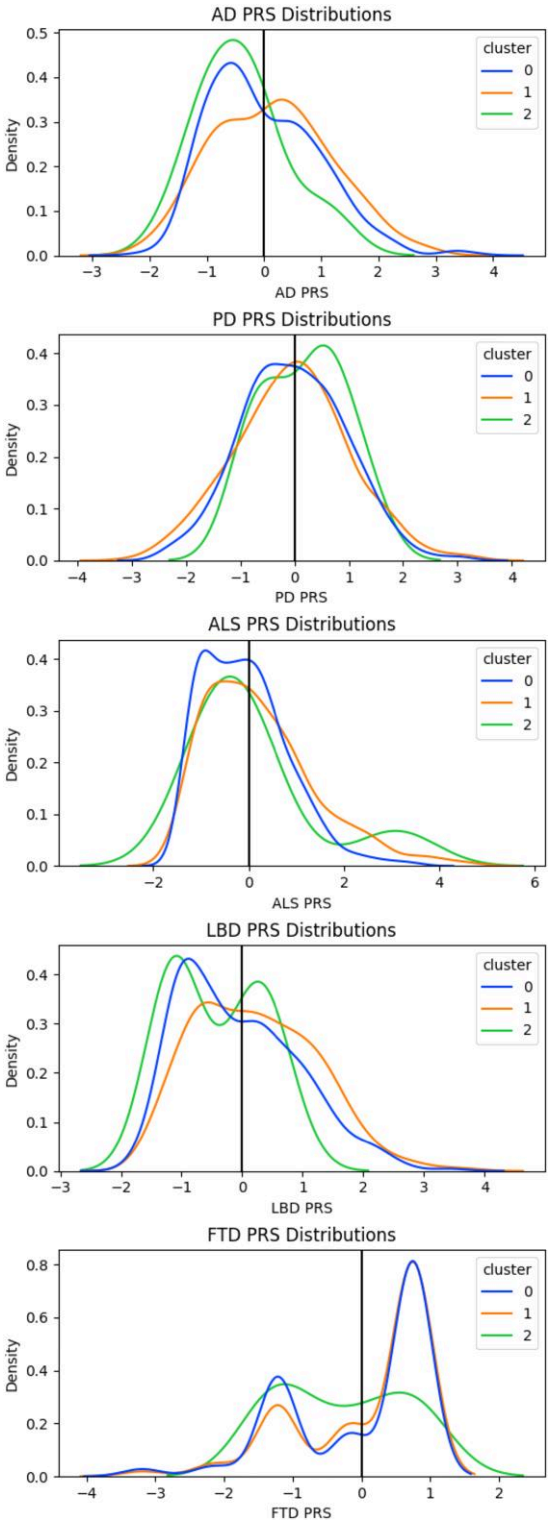

Figure 2 Standardized PRS distributions per multi-disease cluster. AD = Alzheimer's disease; PD = Parkinson's disease; ALS = amyotrophic lateral sclerosis; LBD = Lewy body dementia; FTD = frontotemporal dementia; PRS = polygenic risk score.

In the AD-specific analysis, C1 showed significant genetic risk enrichment for AD. This same cluster also demonstrated significant enrichment of ALS and LBD genetic risk factors. In the PD-specific analysis, C0 was significantly enriched for AD genetic risk, and C1 for LBD genetic risk. In ALS, there are two clearly defined clusters, one significantly enriched for ALS genetic risk and the other depleted for genetic risk. The ALS cluster that has increased ALS genetic risk shows a significant decrease in AD genetic risk loading. For LBD, C0 shows a significant decrease in both AD and LBD genetic risk. In contrast, C1 shows a significant increase in AD and LBD genetic risk, suggesting a less genetically influenced form of the disease is likely. FTD disease-specific clusters were complicated, as the FTD genetic risk was not significantly associated with any of the three clusters identified. What is of interest is that one FTD cluster is significantly enriched for AD, PD, ALS and LBD genetic risk, while another cluster is significantly depleted with regard to genetic risk for PD, ALS and LBD.

Comparative analyses

Supplementary Table 10 compares the multi-disease cluster counts from the iterative clustering approach with the four conditional analyses described previously. Hierarchical clustering also identifies three clusters that show a similar pattern to the iterative clustering approach (Supplementary Fig. 5). Supplementary Fig. 6 contains a dendrogram to visualize the hierarchical relationship between the clusters formed. The hierarchical clusters reveal a much larger C0 (4073 cases compared to 2863 cases) and a much smaller C1 (857 cases compared to 2074 cases). However, when looking at the disease enrichments per cluster the only changes in significance are the presence of PD in C1 and ALS in C2 (Supplementary Table 11).

The proportion of samples in C2 increased for the analysis with the *APOE* locus removed when compared to the original analysis (Supplementary Fig. 7). Despite this, the regression of disease status against cluster membership shows C0 is still most enriched with ALS (OR = 1.752, $P = 2.3 \times 10^{-8}$, beta = 0.561, SE = 0.090), C1 with AD (OR = 1.701, $P = 6.1 \times 10^{-10}$, beta = 0.531, SE = 0.086), and C2 with FTD (OR = 1.855, $P = 0.021$, beta = 0.618, SE = 0.269). C2 is significantly enriched with LBD, but not to the point of significance in the regressions, revealing that the increase in C2 size is largely due to LBD cases (Supplementary Table 12). The PRS regressions revealed a similar trend, with PRS for AD, PD, ALS and LBD having significant negative associations with membership in C0 and significant positive associations with membership in C1 (Supplementary Table 13). This shows a similar pattern to the original analysis and a distinct group of samples that are consistently depleted for genetic risk.

The proportion of samples in each cluster is very similar for the downsampled analysis when compared to the 5000 sample analysis (Supplementary Fig. 8). The regression of disease status against cluster membership once again shows C0 is most enriched with ALS (OR = 1.731, $P = 1.5 \times 10^{-5}$, beta = 0.549, SE = 0.127), C1 with AD (OR = 1.735, $P = 6.0 \times 10^{-6}$, beta = 0.551, SE = 0.121), and C2 with FTD (OR = 2.351, $P = 0.042$, beta = 0.855, SE = 0.421). Again, none of the clusters show significant enrichment of LBD (Supplementary Table 14). The same trend in PRS regressions was observed, with PRS for AD, PD, ALS and LBD having significant negative associations with membership in C0 and significant positive associations with membership in C1 (Supplementary Table 15).

Of the 1000 controls included as a negative control, 991 are grouped into C0 (Supplementary Fig. 9). Because of this, C0 is most significantly enriched with controls (OR = 108.690, $P = 6.1 \times 10^{-25}$, beta = 4.688, SE = 0.450) and is significantly depleted for all NDDs except ALS (Supplementary Table 16). It continues to be the case that C1 is most enriched with AD (OR = 2.097, $P = 1.5 \times 10^{-18}$, beta = 0.740, SE = 0.084) and C2 with FTD (OR = 108.690, $P = 6.1 \times 10^{-25}$,

beta = 4.688, SE = 0.450). Once again, C0 shows significant negative associations with PRS for each NDD besides FTD, which is expected as nearly every control is contained within the cluster (Supplementary Table 17).

The LDSC analysis showed that AD and PD ($r = 0.197$, SE = 0.084, $P = 0.019$), AD and LBD ($r = 0.385$, SE = 0.188, $P = 0.040$), and PD and LBD ($r = 0.599$, SE = 0.166, $P = 3.0 \times 10^{-4}$) were the only pairs of NDDs to have significant genetic correlations (Supplementary Table 18). This is less genetic overlap than is seen in the NDD-specific clustering analysis, where significant associations were seen between the AD PRS and cluster membership in all studied NDDs, as well as the ALS PRS and cluster membership for AD and FTD (Supplementary Table 9).

Discussion

Utility in disease subtyping and clinical trials

The clustering results support the idea that closely related NDDs have more overlapping genetic aetiology than previously expected, using data-driven approaches to show how in many cases, neurodegeneration should be viewed as a spectrum of symptomology and risk factors, not discrete units. This is evident in that we have shown that all three of the multi-disease clusters have members from each studied NDD, and each of these clusters shows a significant association with variants from loci that have been previously linked to multiple NDDs. We have also shown these clusters to be robust to changes in SNP selection, sample size, the inclusion of the *APOE* locus, and the presence of controls. Synthesizing the results of our disease-specific and multi-disease clustering, we note that major differentiating factors between patients seem to be a general lack of genetic risk or a mix of disease risk enrichments at varying degrees (corresponding to Cluster 0). This exemplifies the need for future studies of environmental and epigenetic risk factors shared across NDDs. We also suggest that repeating these analyses in a large set of harmonized pathology-derived data would provide downstream insights on shared mechanisms in the brains of affected patients within and across diseases.

The overlapping deviations between the disease-specific clusters and the PRS for various NDDs provide evidence that neurodegeneration lies on a spectrum.[33] While it is possible that our participant-level analysis may be more optimistic compared to summary statistics-based methods, these overlaps show that there are more genetic associations between NDDs than previously revealed through traditional methods, such as LDSC. There may be groups of patients diagnosed with one NDD but have a high genetic risk for another. For example, the PRS for LBD, a disease that is already known to be closely related to PD and AD in terms of both clinical and pathological manifestations, has significant associations with cluster membership in all of the other presented NDDs.[34-36] Overlaps like these show the need for further research into refining phenotypes for the diagnosis of NDDs, as well as closer monitoring of individuals post-diagnosis to see if changes occur that may cause the need for reconsideration of treatments. Understanding that NDDs with overlapping pathologies tend to share genetic risk loci, in diagnosis and clinical trial enrolment, it will be important to determine how the variants are most strongly associated with each NDD. More importantly, understanding how the variants might have subtly different effects on downstream protein expression (i.e. tau pathology, α-synuclein expression) and pathology that influence disease manifestation will be valuable for precision clinical trials.

Limitations

The limitations of this research include a lack of diversity, insufficient clinical data across sample series, case imbalances between diseases that limited total sample sizes, a lack of rare variant inclusion, and a lack of information on insertions, deletions and expansions.[37] In particular, the FTD PRS estimates suffered from the small GWAS sample size and that may have impacted the results. Given the limited availability of non-European samples, it is difficult to appropriately model any effects ancestral differences may introduce. Similarly, the imbalance between disease sample sizes was a significant limitation. The imbalance resulted in the use of 1000 cases from each disease. The decision to use 1000 cases from each disease resulted from the ALS, the smallest cohort, only having 1105 cases. The clustering model would benefit from having more samples in order to make it more robust to outliers and to potentially identify any other potential clusters not captured in the currently sampled cohorts. The number of cases per disease could have been increased to 2000 if ALS and FTD were removed from the analysis, however, the scope of the results would have been limited since LBD has previously been shown to be closely related to AD and PD both clinically and pathologically. In addition, we have shown the results to be robust to downsampling to 500 cases per disease, which mitigates some concern about the sample size used in the clustering analysis. The lack of broad and uniform clinical/phenotypic data across cohorts limited analyses and translational conclusions. The only common phenotype data common across cohorts was sex and European ancestry. Age collection across cohorts varied with no common collection point (i.e. age at onset, age at death, etc.) and differing age measures from precise ages to age range bins. Additionally, the quality of phenotypes and the impact of 'proxy-cases' or self-reported cases in some large biobank studies may impact the overlap across diseases as, ideally, all phenotypes would be corroborated by imaging or pathology. Other clinical traits that would have been useful for further analyses include family history, disease severity, and medication status. The lack of rare variant inclusion implies that the clustering model may not identify acute genetic differences between NDDs. This lends to these clusters being quite broad and focused on sporadic manifestations of these NDDs, likely not establishing contrasts that could be attributed to early-onset familial cases. Last, because the clusters formed here are based on individual SNPs they do not account for insertions, deletions or expansions. For example, pathology overlaps are seen between ALS and FTD cases due to repeat expansions in *C9orf72*.[8] This limits the amount of genetic overlap between diseases that can be captured by the clusters.

Conclusion

This report used data-driven methods to define the spectrum of neurodegenerative disease interconnectivity. These connections between diseases are based on shared genetic risk factors and the interplay between these risk factors, often recapitulated in symptomology and pathology. Using these data, we can better understand potential fine-grained diagnoses that incorporate more variability than previous discrete classifications of neurodegenerative diseases.

Acknowledgements

Data used in the preparation of this article were obtained from the AMP-PD Knowledge Platform. For up-to-date information on the

study, visit https://www.amp-pd.org. AMP-PD—a public-private partnership—is managed by the FNIH and funded by Celgene, GSK, the Michael J. Fox Foundation for Parkinson's Research, the National Institute of Neurological Disorders and Stroke, Pfizer, and Verily. We would like to thank AMP-PD for the publicly available whole-genome sequencing data, including cohorts from the Fox Investigation for New Discovery of Biomarkers (BioFIND), the Parkinson's Progression Markers Initiative (PPMI), and the Parkinson's Disease Biomarkers Program (PDBP). The Parkinson's Disease Biomarker Program (PDBP) consortium is supported by the National Institute of Neurological Disorders and Stroke (NINDS) at the National Institutes of Health. A full list of PDBP investigators can be found at https://pdbp.ninds.nih.gov/policy. Harvard Biomarker Study (HBS) is a collaboration of HBS investigators (full list of HBS investigators found at https://www.bwhparkinsoncenter.org/biobank) and funded through philanthropy and NIH and Non-NIH funding sources. The HBS Investigators have not participated in reviewing the data analysis or content of the manuscript. The DementiaSeq data were obtained from dbGap (accession number: phs001963.v2.p1).

The results published here are in whole or in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.org). Data generation was supported by the following NIH grants: P30AG10161, P30AG72975, R01AG15819, R01AG17917, R01AG036836, U01AG46152, U01AG61356, U01AG046139, P50 AG016574, R01 AG032990, U01AG046139, R01AG018023, U01AG006576, U01AG00 6786, R01AG025711, R01AG017216, R01AG003949, R01NS080820, U24NS072026, P30AG19610, U01AG046170, RF1AG057440, and U24AG061340, and the Cure PSP, Mayo and Michael J Fox foundations, Arizona Department of Health Services and the Arizona Biomedical Research Commission. We thank the participants of the Religious Order Study and Memory and Aging projects for the generous donation, the Sun Health Research Institute Brain and Body Donation Program, the Mayo Clinic Brain Bank, and the Mount Sinai/JJ Peters VA Medical Center NIH Brain and Tissue Repository. Data and analysis contributing investigators include Nilüfer Ertekin-Taner, Steven Younkin (Mayo Clinic, Jacksonville, FL), Todd Golde (University of Florida), Nathan Price (Institute for Systems Biology), David Bennett, Christopher Gaiteri (Rush University), Philip De Jager (Columbia University), Bin Zhang, Eric Schadt, Michelle Ehrlich, Vahram Haroutunian, Sam Gandy (Icahn School of Medicine at Mount Sinai), Koichi Iijima (National Center for Geriatrics and Gerontology, Japan), Scott Noggle (New York Stem Cell Foundation), Lara Mangravite (Sage Bionetworks).

This study used the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health (http://hpc.nih.gov).

Funding

This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging (NIA), National Institutes of Health, Department of Health and Human Services; project number ZO1 AG000535, as well as the National Institute of Neurological Disorders and Stroke (1ZIANS003154).

Competing interests

C.A., K.L., H.L., H.I., D.V., F.F. and M.N.'s participation in this project was part of a competitive contract awarded to Data Tecnica International LLC by the National Institutes of Health to support open science research. M.N. also currently serves on the scientific advisory board for Clover Therapeutics and is an advisor to Neuron23 Inc.

Supplementary material

Supplementary material is available at *Brain* online.

References

1. Wood LB, Winslow AR, Strasser SD. Systems biology of neurodegenerative diseases. *Integr Biol*. 2015;7:758-775.
2. Uffelmann E, Huang QQ, Munung NS, *et al*. Genome-wide association studies. *Nature Reviews Methods Primers*. 2021;1:1-21.
3. Visscher PM, Wray NR, Zhang Q, *et al*. 10 Years of GWAS discovery: Biology, function, and translation. *Am J Hum Genet*. 2017; 101:5-22.
4. Beach TG, Adler CH. Importance of low diagnostic accuracy for early Parkinson's disease. *Mov Disord*. 2018;33:1551-1554.
5. Gaugler JE, Ascher-Svanum H, Roth DL, Fafowora T, Siderowf A, Beach TG. Characteristics of patients misdiagnosed with Alzheimer's disease and their medication use: An analysis of the NACC-UDS database. *BMC Geriatr*. 2013;13:137.
6. Nalls MA, Blauwendraat C, Sargent L, *et al*. Evidence for *GRN* connecting multiple neurodegenerative diseases. *Brain Commun*. 2021;3:fcab095.
7. Vieira SRL, Schapira AHV. Glucocerebrosidase mutations: A paradigm for neurodegeneration pathways. *Free Radic Biol Med*. 2021;175:42-55.
8. Balendra R, Isaacs AM. C9orf72-mediated ALS and FTD: Multiple pathways to disease. *Nat Rev Neurol*. 2018;14:544-558.
9. Strang KH, Golde TE, Giasson BI. MAPT Mutations, tauopathy, and mechanisms of neurodegeneration. *Lab Invest*. 2019;99:912-928.
10. Yin Y, Wang Z. Apoe and neurodegenerative diseases in aging. *Adv Exp Med Biol*. 2018;1086:77-92.
11. Gan L, Cookson MR, Petrucelli L, La Spada AR. Converging pathways in neurodegeneration, from genetics to mechanisms. *Nat Neurosci*. 2018;21:1300-1309.
12. Beecham GW, Bis JC, Martin ER, *et al*. The Alzheimer's disease sequencing project: Study design and sample selection. *Neurol Genet*. 2017;3:e194.
13. Mueller SG, Weiner MW, Thal LJ, *et al*. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin N Am*. 2005;15: 869-877. xi - xii.
14. Hodes RJ, Buckholtz N. Accelerating medicines partnership: Alzheimer's disease (AMP-AD) knowledge portal aids Alzheimer's drug discovery through open data sharing. *Expert Opin Ther Targets*. 2016;20:389-391.
15. Iwaki H, Leonard HL, Makarious MB, *et al*. Accelerating medicines partnership: Parkinson's disease. Genetic resource. *Mov Disord*. 2021;36:1795-1804.
16. DEMENTIA-SEQ: Genome sequencing in Lewy Body Dementia, Frontotemporal Dementia, and neurologically healthy controls. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001963.v1.p1
17. Schwartzentruber J, Cooper S, Liu JZ, *et al*. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nat Genet*. 2021;53: 392-402.
18. Nalls MA, Blauwendraat C, Vallerga CL, *et al*. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: A meta-analysis of genome-wide association studies. *Lancet Neurol*. 2019;18:1091-1102.

19. Nicolas A, Kenna KP, Renton AE, *et al*. Genome-wide analyses identify KIF5A as a novel ALS gene. *Neuron*. 2018;97: 1268-1283.e6.
20. Chia R, Sabir MS, Bandres-Ciga S, *et al*. Genome sequencing analysis identifies new loci associated with Lewy body dementia and provides insights into its genetic architecture. *Nat Genet*. 2021;53:294-303.
21. Ferrari R, Hernandez DG, Nalls MA, *et al*. Frontotemporal dementia and its subtypes: A genome-wide association study. *Lancet Neurol*. 2014;13:686-699.
22. LiftOver. Published July 15, 2015. https://genome.sph.umich.edu/wiki/LiftOver#Lift_genome_positions
23. Siva N. 1000 Genomes project. *Nat Biotechnol*. 2008;26:256.
24. Bray SM, Mulle JG, Dodd AF, Pulver AE, Wooding S, Warren ST. Signatures of founder effects, admixture, and selection in the ashkenazi Jewish population. *Proc Natl Acad Sci U S A*. 2010; 107:16222-16227.
25. Makarious MB, Leonard HL, Vitale D, *et al*. GenoML: Automated machine learning for genomics. *arXiv*. Published online 4 March 2021. https://doi.org/10.48550/arXiv.2103.03221
26. Makarious MB, Leonard HL, Vitale D, *et al*. Multi-modality machine learning predicting Parkinson's disease. *NPJ Parkinsons Dis*. 2022;8:35.
27. McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*. Published online 9 February 2018. https://doi.org/10.48550/arXiv.1802.03426
28. Cheng Y. Mean shift, mode seeking, and clustering. *IEEE Trans Pattern Anal Mach Intell*. 1995;17:790-799.
29. Rozemberczki B, Watson L, Bayer P, *et al*. The shapley value in machine learning. *arXiv*. Published online 11 February 2022. https://doi.org/10.48550/arXiv.2202.05594
30. Ochoa D, Hercules A, Carmona M, *et al*. The next-generation open targets platform: Reimagined, redesigned, rebuilt. *Nucleic Acids Res*. 2023;51(D1):D1353-D1359.
31. Murtagh F, Contreras P. Algorithms for hierarchical clustering: An overview. *WIREs Data Mining Knowl Discov*. 2012;2:86-97.
32. Porcellini E, Carbone I, Ianni M, Licastro F. Alzheimer's disease gene signature says: Beware of brain viral infections. *Immun Ageing*. 2010;7:16.
33. Ruffini N, Klingenberg S, Schweiger S, Gerber S. Common factors in neurodegeneration: A meta-study revealing shared patterns on a multi-omics scale. *Cells*. 2020;9:2642.
34. Azar M, Chapman S, Gu Y, Leverenz JB, Stern Y, Cosentino S. Cognitive tests aid in clinical differentiation of Alzheimer's disease versus Alzheimer's disease with Lewy body disease: Evidence from a pathological study. *Alzheimers Dement*. 2020; 16:1173-1181.
35. McKeith IG, Boeve BF, Dickson DW, *et al*. Diagnosis and management of dementia with Lewy bodies: Fourth consensus report of the DLB consortium. *Neurology*. 2017;89:88-100.
36. Irwin DJ, Hurtig HI. The contribution of tau, amyloid-Beta and alpha-synuclein pathology to dementia in Lewy body disorders. *J Alzheimers Dis Parkinsonism*. 2018;8:444.
37. Uitterlinden AG. Diversity in human genetics studies accelerates discovery and improves health care. *Nat Rev Cardiol*. 2022; 19:289-290.