Supplementary Materials for

Learning Phenotypic Associations for Parkinson's Disease with Longitudinal Clinical Records

Supplementary Methods

Phenotypic association graph generation. Given the study cohort, suppose V is the set of features and features may be either discrete or continuous. Each feature $i \in V$ corresponds to a node on a graph G. We use the PC algorithm to discover the associations among the features. The first step is generating the phenotypic association graph of the clinical features. In this context, we introduce a modified conditional Gaussian likelihood ratio test which can be applied to mix-type features as conditional independence test. For the threshold of p-value in conditional independence test, we choose 0.01 for PPMI, 0.01 for PDBP and 0.05 for BioFIND.

Algorithm 1 Phenotypic Association Graph Generation

INPUT: Vertex Set $\boldsymbol{V}$, Conditional Independence Criterion
OUTPUT: Estimated graph $\boldsymbol{C}$, separation sets $\boldsymbol{S}$
Form the complete undirected graph $\widetilde{\boldsymbol{C}}$ on the vertex set $\boldsymbol{V}$.
$\boldsymbol{l} = -\mathbf{1}, \boldsymbol{C} = \widetilde{\boldsymbol{C}}$
Init $\boldsymbol{S}(\boldsymbol{j}, \boldsymbol{i})$ as empty set for each $\boldsymbol{i}, \boldsymbol{j} \in \boldsymbol{V}$
Repeat
  $\boldsymbol{l} = \boldsymbol{l} + \mathbf{1}$
  Repeat
    Select a (new) ordered pair of nodes $\boldsymbol{i}, \boldsymbol{j}$ that are adjacent in $\boldsymbol{C}$ such that $|\boldsymbol{adj}(\boldsymbol{C}, \boldsymbol{i})\backslash\{\boldsymbol{j}\}| \geq \boldsymbol{l}$
    Repeat
      Choose $\boldsymbol{k} \subseteq \boldsymbol{adj}(\boldsymbol{C}, \boldsymbol{i})\backslash\{\boldsymbol{j}\}$ with $|\boldsymbol{k}| = \boldsymbol{l}$, all nodes in $\boldsymbol{k}$ should
      If $\boldsymbol{i}$ and $\boldsymbol{j}$ are conditional independent given $\boldsymbol{k}$ **Then**
        Delete edge $\boldsymbol{i}, \boldsymbol{j}$ in $\boldsymbol{C}$
        $\boldsymbol{S}(\boldsymbol{i}, \boldsymbol{j}) = \boldsymbol{k}, \boldsymbol{S}(\boldsymbol{j}, \boldsymbol{i}) = \boldsymbol{k}$
      End if
    Until edge $\boldsymbol{i}, \boldsymbol{j}$ is deleted or all $\boldsymbol{k} \subseteq \boldsymbol{adj}(\boldsymbol{C}, \boldsymbol{i})\backslash\{\boldsymbol{j}\}$ with $|\boldsymbol{k}| = \boldsymbol{l}$ have been chosen
  Until each pair $\boldsymbol{i}, \boldsymbol{j}$ that are adjacent in $\boldsymbol{C}$ such that $|\boldsymbol{adj}(\boldsymbol{C}, \boldsymbol{i})\backslash\{\boldsymbol{j}\}| \geq \boldsymbol{l}$ has been chosen
Until each ordered pair $\boldsymbol{i}, \boldsymbol{j}$ that are adjacent in $\boldsymbol{C}$ that $|\boldsymbol{adj}(\boldsymbol{C}, \boldsymbol{i})\backslash\{\boldsymbol{j}\}| < \boldsymbol{l}$

Variable grouping with longitudinal phenotypic association graphs. We use the Louvain community detection algorithm to group clinical variables. Modularity Q is defined to measure the density of inter-community links as compared to intra-community ones. The formula to calculate modularity of a weighted graph is as follows:

$$Q = \frac{1}{2m}\sum_{i,j}[A_{i,j} - \frac{k_i k_j}{2m}]\delta(c_i, c_j)$$

where $A_{i,j}$ is the weight of the link between node i and j, $k_i = \sum_j A_{i,j}$, $m = \frac{1}{2}\sum_{i,j} A_{i,j}$, c_i is the community which node belongs to, $\delta(c_i, c_j) = \begin{cases} 1, if\ c_i = c_j \\ 0, else \end{cases}$. As defined, Q is a scalar ranging from -1 to 1. The Louvain community detection algorithm aims to search for the optimal grouping result by maximizing Q.

In the study, since we want to obtain a grouping result which is consistent over all periods, we first generate phenotypic association graphs for all periods. Then we construct the summarization graph where weight of each edge is the frequency of occurrence of the edge over all periods.

Supplemental Tables

Table S1. The first part of features for analysis, including demographic, medicine, biomarker, neuroimaging, and motor features. Alongside the name of the feature and the corresponding question addressed, data type is denoted by C: continuous; D: discrete, and whether the feature is included in the dataset is denoted by Y: yes; N: no; Y(BL): included only at baseline in PPMI.

Type	Name	Meaning	Type	PPMI	PDBP	BioFIND
Demographic	Gender		D	Y	Y	Y
	Age		C	Y	Y	Y
	Race	Whether is white	D	Y	Y	Y
	Family history	Whether has PD family history	D	Y	Y	Y
	EY	Years of education	C	Y	Y	Y
	Genetic risk	A score indicating the genetic risk of PD, combining different genes related to PD	C	Y	N	Y
	Gene APOE	Another gene related to Alzheimer's, not included in calculation genetic risk score	D	Y	N	N
	Gene SNCA1	mutation of SNCA rs3910105	D	Y	N	N
	Gene SNCA2	mutation of SNCA rs356181	D	Y	N	N
	Gene GBA	mutation of GBA rs76763715	D	Y	N	N
	Gene LRRK2	mutation of LRRK2 rs34637584	D	Y	N	N
Medicine	If-use-med	If using PD medicine	D	Y	N	Y
	If-use-levo	If using Levodopa, the most common PD medicine	D	Y	N	Y
	If-use-da	If using Dopamine Agonist, another common PD medicine	D	Y	N	Y
Motor	MDS-T	Tremor subscore of PD, calculated from MDS-UPDRS	C	Y	Y	Y
	MDS-P	Postural instability and gait disturbance subscore of PD, calculated from MDS-UPDRS	C	Y	Y	Y
	MDS H&Y stage		C	Y	Y	Y
Biomarkers	CSF $A\beta_{1-42}$	cerebrospinal fluid amyloid beta1-42	C	Y(BL)	Y	Y
	CSF t-tau	cerebrospinal fluid total (t)-tau	C	Y(BL)	Y	Y
	CSF p-tau	cerebrospinal fluid phosphorylated tau (P-tau181)	C	Y(BL)	Y	Y
	CSF α-Syn	cerebrospinal fluid α-synuclein	C	Y(BL)	Y	Y
Neuroimaging	DaTScan-c	DaTScan Caudate	C	Y(BL)	N	N
	DaTScan-p	DaTScan Putamen	C	Y(BL)	N	N
	MRI	MRI results	D	Y(BL)	N	N

Abbreviation: MDS-UPDRS=Movement Disorder Society Unified Parkinson Disease Rating Scale. PD=Parkinson's disease. CSF= Cerebrospinal fluid. MRI=magnetic resonance imaging.

Table S2. The second part of features for analysis, including all non-motor clinical features. Alongside the name of the feature and the corresponding question addressed, data type is denoted by C: continuous; D: discrete, and whether the feature is included in the dataset is denoted by Y: yes; N: no; Y(BL): included only at baseline in PPMI.

Type	Name	Meaning	Type	PPMI	PDBP	BioFIND
Non-motor (Not cognitive)	MDS-fatigue	The level of self-reported fatigue, calculated from MDS-UPDRS	C	Y	Y	Y
	MDS-hallucinations	The level of self-reported hallucinations, calculated from MDS-UPDRS	C	Y	Y	Y
	MDS-apathy	The level of self-reported apathy, calculated from MDS-UPDRS	C	Y	Y	Y
	QUIP	Questionnaire for Impulsive-Compulsive Disorders, measurement of impulse control disorders	C	Y	N	N
	RBDSQ	REM Sleep Behavior Disorder Screening Questionnaire, measurement of REM Sleep Behavior Disorder(RBD)	C	Y	Y	Y
	STAI	State-Trait Anxiety Inventory, measurement of anxiety	C	Y	Use the apathy score in MDS-UPDRS Part I	Use the apathy score in MDS-UPDRS Part I
	SCOPA-AUT	Scales for Outcomes in PD-Autonomic, measurement of autonomic functions	C	Y	N	N
	ESS	Epworth Sleepiness Scale, measurement of excessive daytime sleepiness(EDS)	C	Y	Y	N
	GDS	Geriatric Depression Scale, measurement of depression	C	Y	Use the depress score in MDS-UPDRS Part I	Use the depress score in MDS-UPDRS Part I
Non-motor (cognitive)	MoCA	Montreal Cognitive Assessment, measurement of general cognition	C	Y	Y	Y
	SDMT	Symbol Digit Modalities Test, measurement of processing speed/ attention	C	Y	N	N
	LNS	Letter-Number Sequencing, measurement of executive function/working memory	C	Y	N	N
	BJLO	Benton Judgment-of-Line-Orientation, measurement of visuospatial function	C	Y	N	N
	Semantic Frequency	Number of words generated for animals, vegetables, fruit, measurement of executive function/working memory	C	Y	N	N
	HVLT total recall	Hopkins Verbal Learning Test-Revised (HVLT-R) immediate and delayed free recall and recognition: verbal memory, measurement of verbal memory	C	Y	N	N
	HVLT delay recall		C	Y	N	N
	HVLT Disr. recognition		C	Y	N	N

Abbreviation: MDS-UPDRS=Movement Disorder Society Unified Parkinson Disease Rating Scale. PD=Parkinson's disease.

Table S3. MRI Features used in PPMI-baseline.

Type	Region
surface area, cortical volume	bankssts, caudal anterior cingulate, caudal middle frontal, cuneus, entorhinal, fusiform, inferior parietal, inferior temporal, isthmus cingulate, lateral occipital, lateral orbitofrontal, lingual, medial orbitofrontal, middle temporal, parahippocampal, paracentral, pars opercularis, pars orbitalis, pars triangularis, pericalcarine, posterior cingulate, postcentral, precuneus, precentral, rostral anterior cingulate, rostral middle frontal, superior frontal, superior parietal, superior temporal, supramarginal, frontal pole, temporal pole, transverse temporal, insula
White matter volume	bankssts, caudal middle frontal, cuneus, entorhinal, fusiform, inferior parietal, inferior temporal, lateral occipital, lateral orbitofrontal, lingual, middle temporal, parahippocampal, paracentral, pars opercularis, pars orbitalis, pars triangularis, pericalcarine, postcentral, precuneus, precentral, rostral middle frontal, superior frontal, superior parietal, superior temporal, supramarginal, temporal pole, transverse temporal, insula

Supplemental Figures

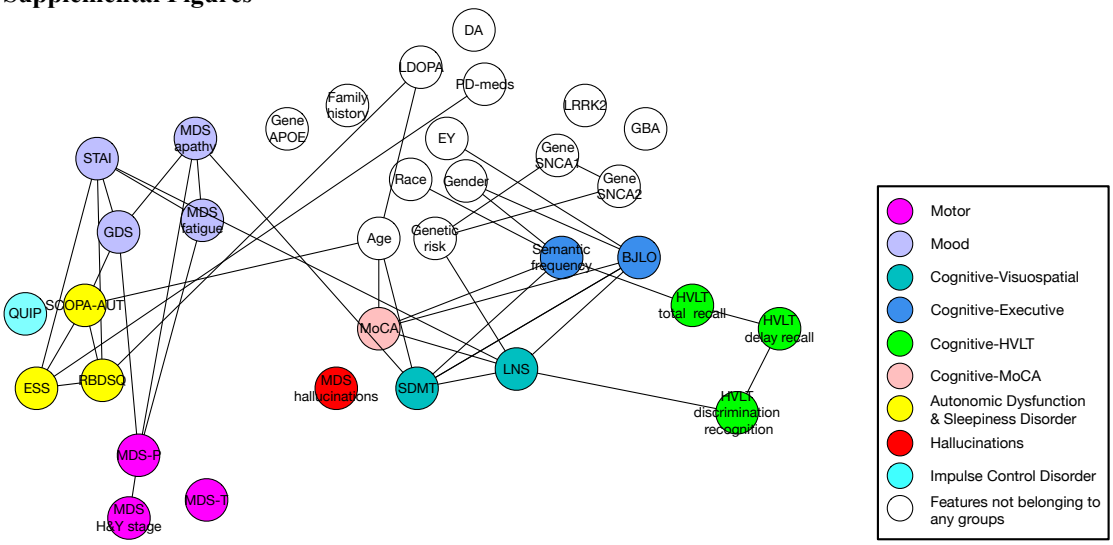

Figure S1. The feature-level phenotypic association graph at 1-year follow-up of PPMI cohort. Each node represents a feature and colors of the nodes represent which group the specific feature belongs to. Nodes in white are those features which do not participate in grouping. MDS=Movement Disorder Society. MDS-T=MDS Tremor score. MDS-P=MDS Postural Instability and Gait Difficulty score. H&Y=Hoehn and Yahr. GDS=Geriatric Depression Scale. STAI=State Trait Anxiety Inventory. HVLT=Hopkin's Verbal Learning Test. MoCA=Montreal Cognitive Assessment. SDMT=Symbol Digit Modalities Test. LNS=Letter Number Sequencing. BJLO=Benton Judgment of Line Orientation. QUIP=Questionnaire for Impulsive-Compulsive Disorders in Parkinson's Disease. SCOPA-AUT=SCales for Outcomes in PArkinson's disease-AUTomotic symptoms. ESS=Epworth Sleepiness Scale. RBDSQ=REM Sleep Behavior Disorder Screening Questionnaire. EY=Years of education. CSF= Cerebrospinal fluid. PD-meds=If using PD medicines. DA=If using Dopamine Agonist, LDOPA=If using Levodopa.

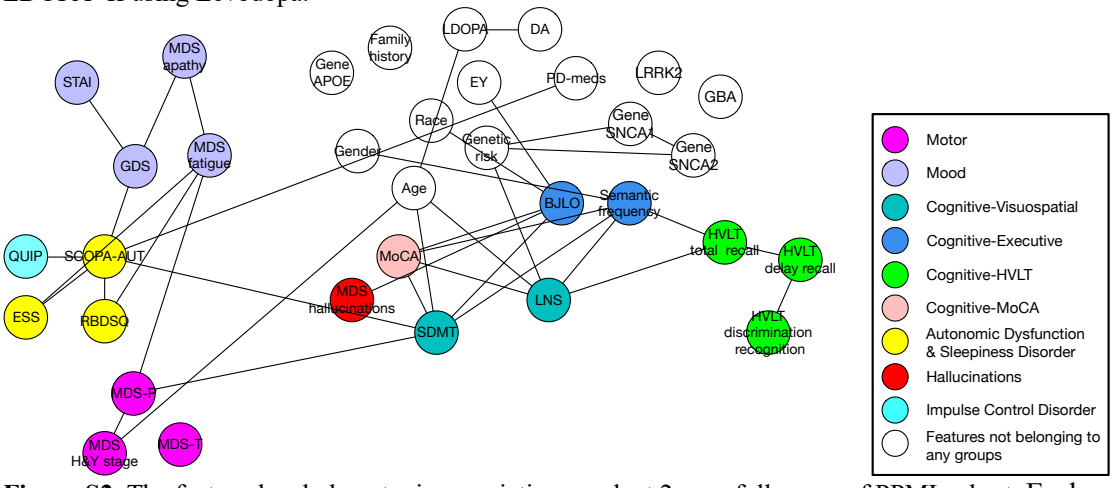

Figure S2. The feature-level phenotypic association graph at 2-year follow-up of PPMI cohort. Each node represents a feature and colors of the nodes represent which group the specific feature belongs to. Nodes in white are those features which do not participate in grouping. MDS=Movement Disorder Society. MDS-T=MDS Tremor score. MDS-P=MDS Postural Instability and Gait Difficulty score. H&Y=Hoehn and Yahr. GDS=Geriatric Depression Scale. STAI=State Trait Anxiety Inventory. HVLT=Hopkin's Verbal Learning Test. MoCA=Montreal Cognitive Assessment. SDMT=Symbol Digit Modalities Test. LNS=Letter Number Sequencing. BJLO=Benton Judgment of Line Orientation. QUIP=Questionnaire for Impulsive-Compulsive Disorders in Parkinson's Disease. SCOPA-AUT=SCales for Outcomes in PArkinson's disease-AUTomotic symptoms. ESS=Epworth Sleepiness Scale. RBDSQ=REM Sleep Behavior Disorder Screening Questionnaire. EY=Years of education. CSF= Cerebrospinal fluid. PD-meds=If using PD medicines. DA=If using Dopamine Agonist, LDOPA=If using Levodopa.

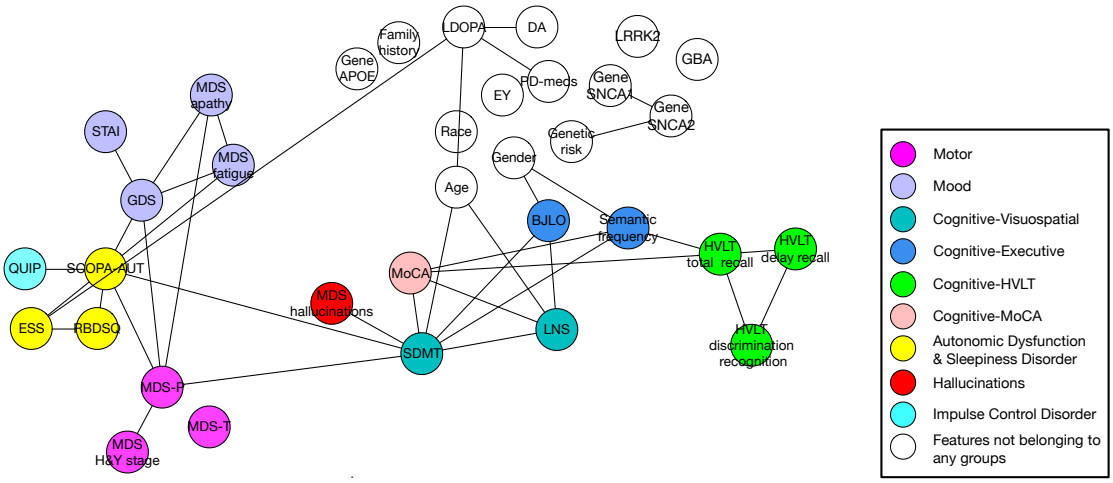

Figure S3. The feature-level phenotypic association graph at 3-year follow-up of PPMI cohort. Each node represents a feature and colors of the nodes represent which group the specific feature belongs to. Nodes in white are those features which do not participate in grouping. MDS=Movement Disorder Society. MDS-T=MDS Tremor score. MDS-P=MDS Postural Instability and Gait Difficulty score. H&Y=Hoehn and Yahr. GDS=Geriatric Depression Scale. STAI=State Trait Anxiety Inventory. HVLT=Hopkin's Verbal Learning Test. MoCA=Montreal Cognitive Assessment. SDMT=Symbol Digit Modalities Test. LNS=Letter Number Sequencing. BJLO=Benton Judgment of Line Orientation. QUIP=Questionnaire for Impulsive-Compulsive Disorders in Parkinson's Disease. SCOPA-AUT=SCales for Outcomes in PArkinson's disease-AUTomotic symptoms. ESS=Epworth Sleepiness Scale. RBDSQ=REM Sleep Behavior Disorder Screening Questionnaire. EY=Years of education. CSF= Cerebrospinal fluid. PD-meds=If using PD medicines. DA=If using Dopamine Agonist, LDOPA=If using Levodopa.

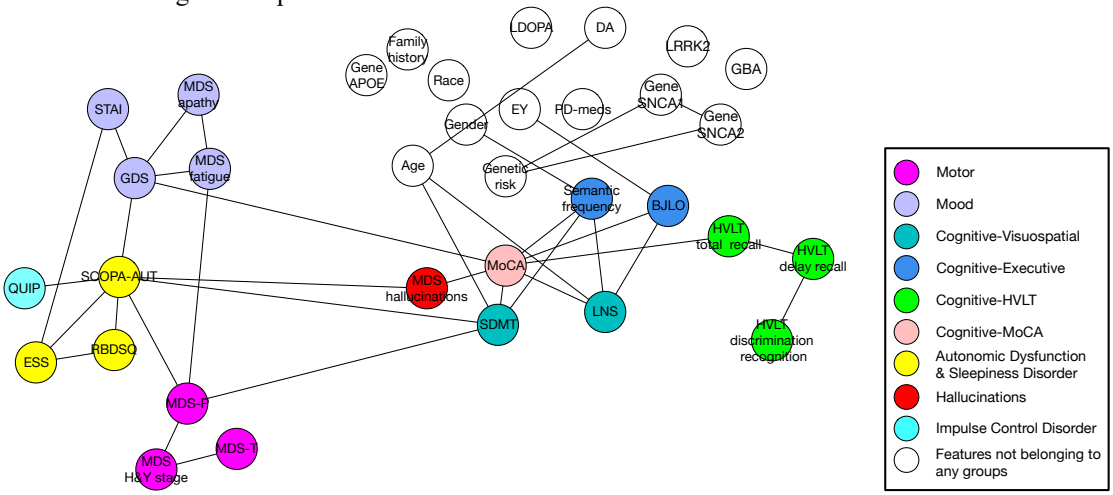

Figure S4. The feature-level phenotypic association graph at 4-year follow-up of PPMI cohort. Each node represents a feature and colors of the nodes represent which group the specific feature belongs to. Nodes in white are those features which do not participate in grouping. MDS=Movement Disorder Society. MDS-T=MDS Tremor score. MDS-P=MDS Postural Instability and Gait Difficulty score. H&Y=Hoehn and Yahr. GDS=Geriatric Depression Scale. STAI=State Trait Anxiety Inventory. HVLT=Hopkin's Verbal Learning Test. MoCA=Montreal Cognitive Assessment. SDMT=Symbol Digit Modalities Test. LNS=Letter Number Sequencing. BJLO=Benton Judgment of Line Orientation. QUIP=Questionnaire for Impulsive-Compulsive Disorders in Parkinson's Disease. SCOPA-AUT=SCales for Outcomes in PArkinson's disease-AUTomotic symptoms. ESS=Epworth Sleepiness Scale. RBDSQ=REM Sleep Behavior Disorder Screening Questionnaire. EY=Years of education. CSF= Cerebrospinal fluid. PD-meds=If using PD medicines. DA=If using Dopamine Agonist, LDOPA=If using Levodopa.

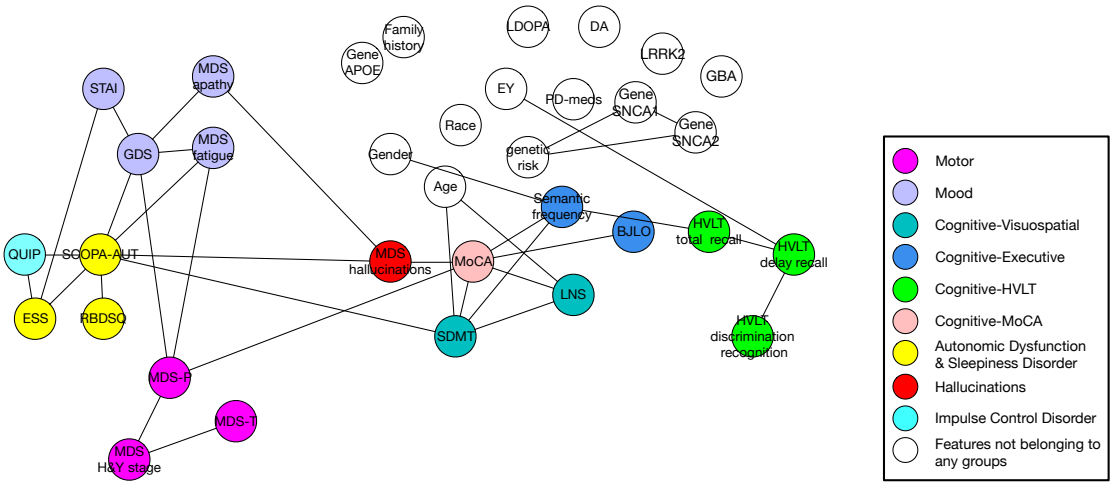

Figure S5. The feature-level phenotypic association graph at 5-year follow-up of PPMI cohort. Each node represents a feature and colors of the nodes represent which group the specific feature belongs to. Nodes in white are those features which do not participate in grouping. MDS=Movement Disorder Society. MDS-T=MDS Tremor score. MDS-P=MDS Postural Instability and Gait Difficulty score. H&Y=Hoehn and Yahr. GDS=Geriatric Depression Scale. STAI=State Trait Anxiety Inventory. HVLT=Hopkin's Verbal Learning Test. MoCA=Montreal Cognitive Assessment. SDMT=Symbol Digit Modalities Test. LNS=Letter Number Sequencing. BJLO=Benton Judgment of Line Orientation. QUIP=Questionnaire for Impulsive-Compulsive Disorders in Parkinson's Disease. SCOPA-AUT=SCales for Outcomes in PArkinson's disease-AUTomotic symptoms. ESS=Epworth Sleepiness Scale. RBDSQ=REM Sleep Behavior Disorder Screening Questionnaire. EY=Years of education. CSF= Cerebrospinal fluid. PD-meds=If using PD medicines. DA=If using Dopamine Agonist, LDOPA=If using Levodopa.

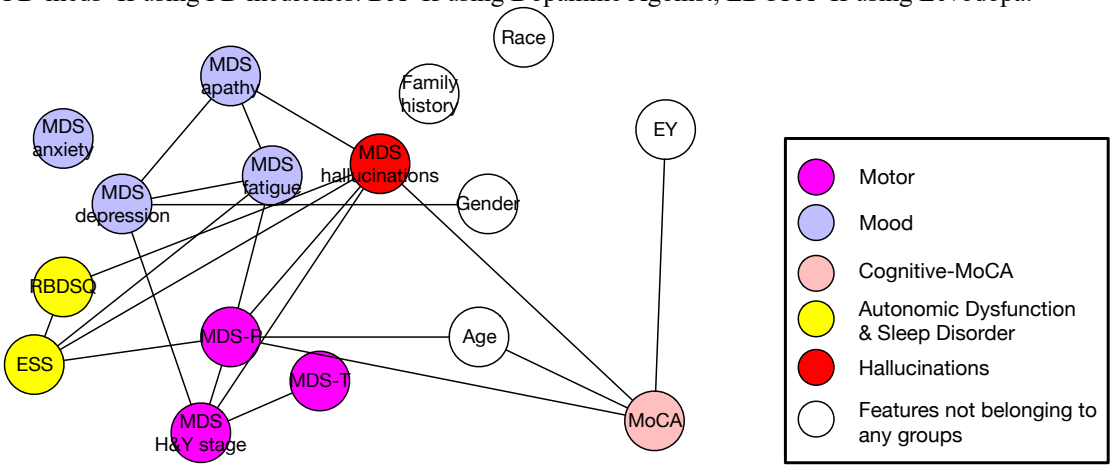

Figure S6. The feature-level phenotypic association graph of at 1-year follow-up of PDBP study. Each node represents a feature and colors of the nodes represent which group the specific feature belongs to. Nodes in white are those features which do not participate in grouping. MDS=Movement Disorder Society. MDS-T=MDS Tremor score. MDS-P=MDS Postural Instability and Gait Difficulty score. H&Y=Hoehn and Yahr. GDS=Geriatric Depression Scale. MoCA=Montreal Cognitive Assessment. ESS=Epworth Sleepiness Scale. RBDSQ=REM Sleep Behavior Disorder Screening Questionnaire. EY=Years of education.

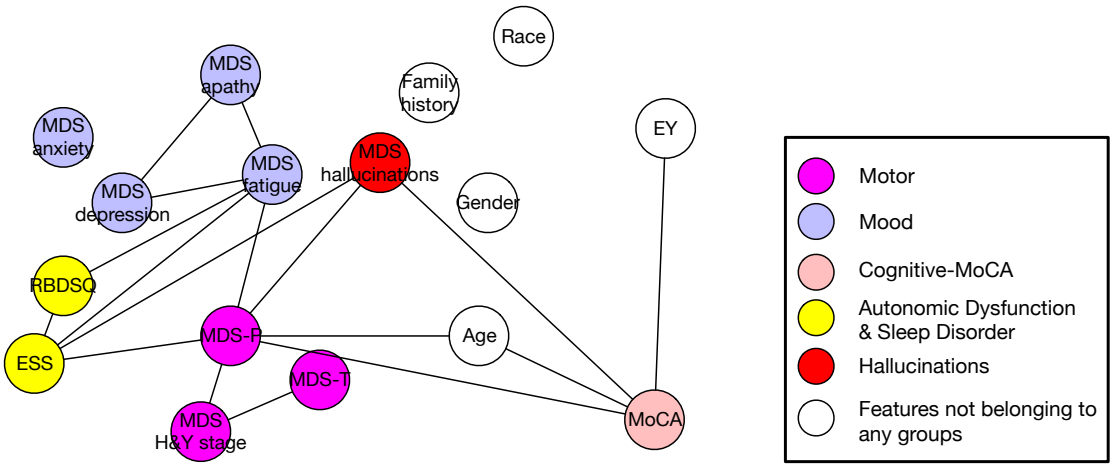

Figure S7. The feature-level phenotypic association graph of at 1-year follow-up of PDBP study. Each node represents a feature and colors of the nodes represent which group the specific feature belongs to. Nodes in white are those features which do not participate in grouping. MDS=Movement Disorder Society. MDS-T=MDS Tremor score. MDS-P=MDS Postural Instability and Gait Difficulty score. H&Y=Hoehn and Yahr. GDS=Geriatric Depression Scale. MoCA=Montreal Cognitive Assessment. ESS=Epworth Sleepiness Scale. RBDSQ=REM Sleep Behavior Disorder Screening Questionnaire. EY=Years of education.