Accelerating Parkinson's Disease Diagnosis using Multi-omics and Artificial Intelligence

Ruifeng Hu[1], Jie Yuan[1], Clemens R Scherzer[1], and Xianjun Dong[1]

[1]Neurogenomics Lab and Precision Neurology Program of Harvard Medical School and Brigham and Women's Hospital, Boston, MA 02115, USA

Background and aims

Developing diagnosis biomarkers in blood for clinical use is a difficult process that requires the evaluation of multiple, large cohorts. Also, for progressive neurodegenerative diseases, early and accurate diagnosis is key to effectively developing and using new interventions. The Accelerating Medicine Partnership in Parkinson's Disease (AMP-PD) consortium provides an unparalleled opportunity to rapidly achieve these previously elusive goals. Our aim is to utilize the multi-omics data and the cutting-edge artificial intelligence–approaches, including traditional machine learning and more advanced deep learning methods, to accelerate the Parkinson's disease (PD) diagnosis.

Data resources

There are 10,247 participants from 7 cohorts in AMP-PD release version 2.0. Among these participants, 8461 longitudinal RNA-seq samples from 3274 participants are available. There are 9901 participants have the whole genome sequencing data. To find early diagnosis biomarkers, we first only focused on the baseline data analysis.

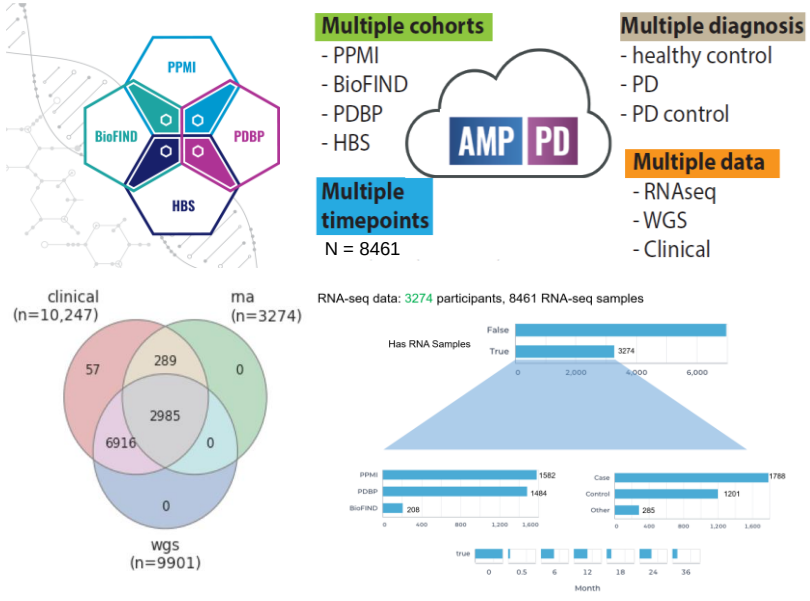

Figure 1. AMP-PD data summary. Multi-omics data including clinical, transcriptomic, genomics data are available.

Methods and data processing

The differentially expressed genes (DEGs) in PD vs. health controls were initially discovered from the analysis of the PPMI cohort. Then we replicated those significant DEGs in a cross-sectional analysis of the PDBP and BioFIND cohorts.

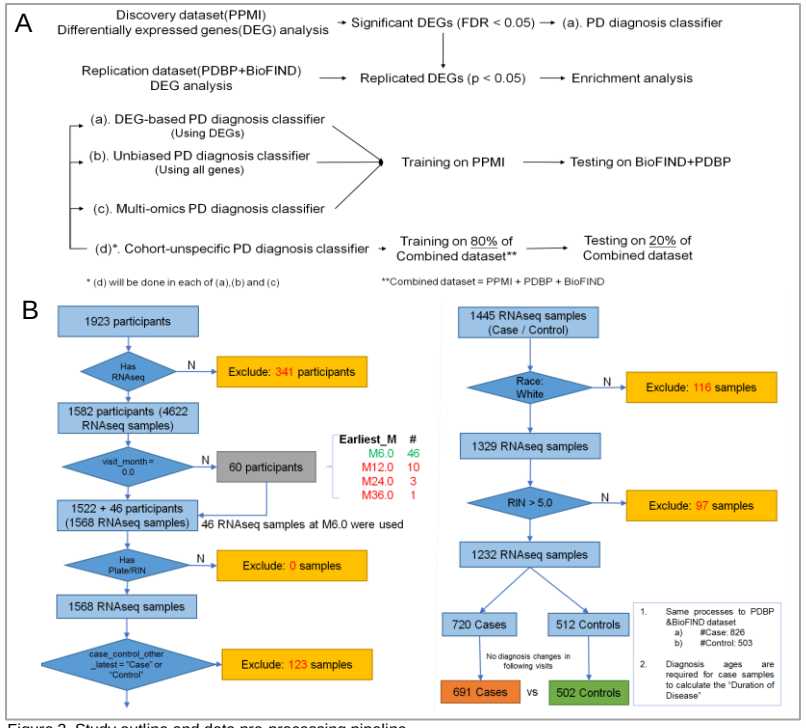

Figure 2. Study outline and data pre-processing pipeline.

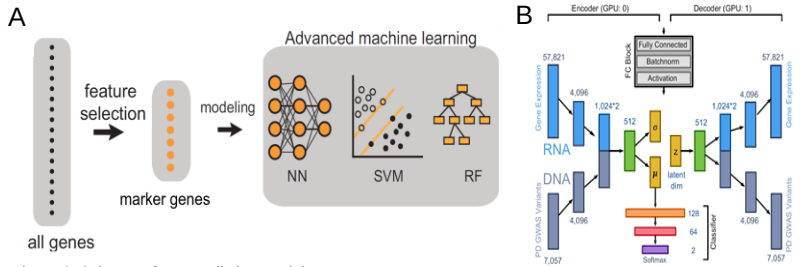

Figure 3. Schema of PD prediction models

Results

1,969 DEGs were obtained from the analysis of PPMI dataset. Among the 1969 DEGs, 965 were replicated in PDBP&BioFIND data analysis. 84 genes were further confirmed in the analysis of brain samples[1].

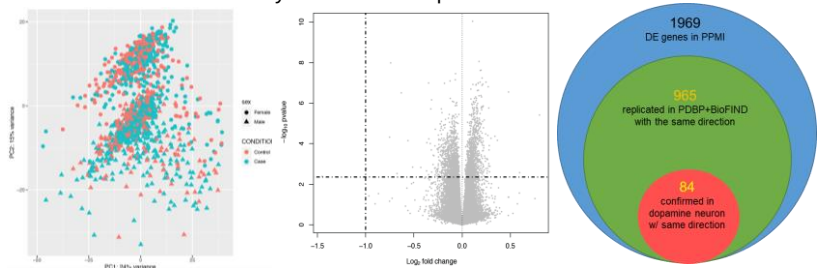

Figure 4. Results of differentially expressed genes. There is no clear cluster between case and control groups and males and females were separated in PCA plot.

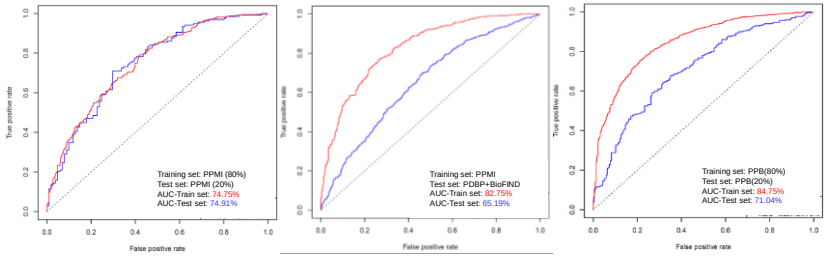

Figure 5 .Performances of DEG-based PD diagnisos classifiers. Left: The classifier was trained on 80% samples of PPMI dataset and tested on the rest of 20% samples, the testing AUC is 74.75%. Middle: The classifier was trained on the whole PPMI dataset and tested on the samples from PDBP + BioFIND datasets. The testing AUC is 65.19%. Right: To remove the cohort bias, PPMI,PDBP, and BioFIND samples were combined and mixed together as the PPB dataset. The classifier was trained on 80% samples of PPB dataset and tested on the rest of 20% samples, the testing AUC is 71.04%.

Discussion

This is our preliminary results of an ongoing investigation. In our first step, we built the multi-genes classifiers. We next will build multi-omics classifiers by combining both PD-associated RNAs and DNAs signals with state-of-the-art deep learning techniques (e.g. variational autoencoder). We will also include novel RNAs like circRNAs and eRNAs into the analysis. This analysis will powerfully delineate - for the first time - the full spectrum of known and novel, coding and noncoding RNAs linked to PD and detectable in circulating blood cells in a harmonized, large-scale data set. It will develop and test highly innovative multi-omics classifiers and provide a generally useful computational framework for large-scale, unbiased PD biomarker discovery.

Reference: [1] Dong X, Liao Z, Gritsch D, *et.al.* Nature neuroscience. 2018 Oct;21(10):1482-92.