Powerful and novel statistical methods to detect genetic variants associated with or putative causal to Alzheimer’s disease

Research Use Statement (RUS). The statement should include the following components (2,200 characters max):

- Objectives of the proposed research;
- Study design;
- Analysis plan, including the phenotypic characteristics that will be evaluated in association with genetic variants;
- If applicable, a brief description of any planned collaboration with researchers at other institutions, including the name of the collaborator(s) and their institutions(s).

We have been developing more powerful statistical methods to detect common variant (CV)- or rare variant (RV)-complex trait associations and/or putative causal relationships for GWAS and DNA sequencing data. Here we propose applying our new methods, along with other suitable existing methods, to the existing ADSP sequencing data and other AD GWAS data provided by NIA, hence requesting approval for accessing the ADSP sequencing and other related GWAS/genetic data. We have the following two specific Aims:

Aim1. Association testing under genetic heterogeneity: For complex traits, genetic heterogeneity, especially of RVs, is ubiquitous as well acknowledged in the literature, however there is barely any existing methodology to explicitly account for genetic heterogeneity in association analysis of RVs based on a single sample/cohort. We propose using secondary and other omic data, such as transcriptomic or metabolomic data, to stratify the given sample, then apply a weighted test to the resulting strata, explicitly accounting for genetic heterogeneity that causal RVs may be different (with varying effect sizes) across unknown and hidden subpopulations. Some preliminary analyses have confirmed power gains of the proposed approach over the standard analysis.

Aim 2. Meta analysis of RV tests: Although it has been well appreciated that it is necessary to account for varying association effect sizes and directions in meta analysis of RVs for multi-ethnic cohorts, existing tests are not highly adaptive to varying association patterns across the cohorts and across the RVs, leading to power loss. We propose a highly adaptive test based on a family of SPU tests, which cover many existing meta-analysis tests as special cases. Our preliminary results demonstrated possibly substantial power gains.

Specifically, for Aim 1, we will first use the AD GWAS dataset NG00073, available from the NIA Genetics of Alzheimer’s Disease Data Storage Site website (https: //www.niagads.org/datasets/ng00073), as the secondary data. The data contain five cognitively defined AD subgroups, each with 3444 controls and between 141 and 974 AD cases: 1. Controls vs. memory predominant AD; 2. Controls vs. visuospatial predominant AD; 3. Controls vs. language predominant AD; 4. Controls vs. no domain with substantial relative impairment AD group; 5. Controls vs. multiple substantial relative impairment AD group. We will use the SNPs in the exome to build a model (e.g. by the nearest shrunken-centroid method, or supervised principal component analysis method) to cluster and thus predict the AD subgroups. We will then apply the constructed predictive model to the ADSP WES and WGE (but using only SNVs in the exome) data to partition the AD subjects into the corresponding AD subgroups, before applying our powerful gene-based adaptive and weighted aSPU test (called CaSPU) to identify AD-associated RVs after accounting for the heterogeneity of AD.

For Aim 2, we will first conduct the gene-based aSPU test on each stratum of the ADSP WES and WGS data, stratified by the combination of ethnic groups (e.g. EA versus CH), sequencing platforms (i.e. WES versus WGS) and study types (i.e. case-control versus family-based studies), then apply our powerful and adaptive aSPU test for meta-analysis of these multi-ethnic groups. Alternatively, we can further partition each of the above strata into AD subgroups as proposed in Aim 1, analyze each subgroup (within each stratum) (with our weighted aSPU testing proposed in Aim 1), then combine them by meta analysis.

Our methods can be also applied as weighted analysis to integrate GWAS or DNA-sequencing data with functional annotations and other omic data.

6. **Non-Technical Summary.** This non-technical summary of your proposed research plan will be made publicly available for lay audiences to read (1,100 characters max).

We propose applying our newly developed statistical analysis methods, along with other suitable existing methods, to the existing ADSP sequencing data and other AD GWAS data to detect common or rare genetic variants associated with Alzheimer's disease (AD). The novelty and power of our new methods are in two aspects: first, we consider and account for possible genetic heterogeneity with several subcategories of AD; second, we apply powerful meta-analysis methods to combine the association analyses across multiple subcategories of AD. The proposed research is feasible, promising and potentially significant to AD research. In addition, our proposed analyses of the existing large amount of ADSP sequencing data and other AD GWAS data with our developed new methods are novel, powerful and cost-effective.

7. **Derived/Secondary Data Return Plan**. The Investigator must describe the derived/secondary data that will be returned to the NIA Genetics of Alzheimer's Disease Data Storage Site (NIAGADS): see Sample-Data-Return-NIAGADS.

We will contact NIAGADS to start the process of submitting derived data when a publication has been accepted, so the data will become available to the public in compliance with the timeline indicated in the GDS.

Some main data types are:

1) Case/control or endophenotype association studies: (meta analysis of) genome-wide association significance, allele frequency, and summary statistics;
2) Bioinformatic annotations of variants that are used in published analysis.