Research Use Statement for Application for Genomic Data from NIAGADS

Objectives

The proposed work seeks to compare published large scale meta-analysis (Naj, AC et al. *Nat. Genet.* **43**, 436-441 10.1038/ng.801, 2011) to Linear Mixed and Bayesian-based models of GWAS on large-scale integrated data sets The harmonization / integration procedure (relative to 1000G) to be refined and tested is a novel approach adopted by us that we expect will decrease false positives arising from the merging itself. Specifically we seek to generate a genomic prediction model to assess risk and, more importantly, age-at-onset (AAO) for Alzheimer's disease. The preferred sample size to construct such a panel necessitates the integration of the multiple large data sets requested with this application.

This is not the same as meta-analyses, which is commonly performed by combining summary statistics resulting from independent analysis of each data set. Integrating GWAS datasets at the level of the raw genotype data while controlling for false positives requires bridging SNP identification across multiple analysis chips, as well as proper orientation of phase across multiple genomic regions. For many nucleotide substitutions, assigning proper phase orientation is not particularly difficult or complicated (eg: A/G and C/T). However, for other mutations, phase is difficult to infer (eg: A/T and G/C). We have generated software that uses a strategy for data harmonization that includes entropy and imputation techniques relative to reference genomic sequence data. The identification of SNPs across multiple assay platforms is complicated by incomplete, conflicting, and at times inaccurate annotation files that may accompany each GWAS assay chip.

Design and Analysis

We shall determine the predicted risk of AD and the predicted age-at-onset (AAO) and determine the power of analysis versus data set size and integration procedure. The results will be compared to meta-analysis. All subject IDs will be analyzed while controlling (as fixed effects) for sex, years of education, ancestry, and possibly cohort indicators. Population stratification will be handled as random variables with kinship coefficients computed as numerical (genetic) relatedness matrices. Genetic estimations will be performed with at least LMM and BSLMM using the GCTA and GEMMA codes, respectively. Imputation will be performed using MACH in conjunction with our software pipeline called MACHTools. MACHTools encapsulates capabilities for managing high throughput execution of the MACH process, applying quality controls on the data and performing necessary genotyping corrections required for merging data.

Replication will be constructed in a manner consistent with the meta-analysis published work of Naj et al.

Non-Technical Summary for Application for Genomic Data from NIAGADS

We propose to integrate the requested AD GWAS data sets to apply a form of genetic analysis that is not common in human genetic association testing. In most GWAS data analyses, marker by marker tests are performed to evaluate the marginal effects of individual SNPs. We are proposing to use methods that are grounded in the genetics of plant and animal breeders.

These methods traditionally use the entire suite of genetic information (variants) to predict the performance of the next generation of crops or livestock in an agricultural setting. These same techniques may be used with human data to estimate characteristics of complex disease traits. But because they require comprehensive genetic data, genomic prediction methods have not been widely used. We propose that with the completion of numerous public genotyping and sequencing projects (HGP, HapMap, 1000 Genomes), sufficient genetic data now exist to allow the use of genomic prediction methods in humans