December 21, 2012

Note:

This is a summary of a plan that is still being developed for the Alzheimer's Disease Genome Sequencing Project (ADSP), an initiative jointly funded by the National Human Genome research Institute (NHGRI) and the National Institute on Aging (NIA). Readers should understand that, because this plan is still under development and has not been formally approved or announced, some important details may be subject to change.

This summary was extracted from the unfinished ADSP plan by NIH staff. The non-NIH ADSP participants that are working on the ADSP plan are not responsible for any omissions that may have resulted.

NIA is posting this summary to aid applicants to program announcement http://grants.nih.gov/grants/guide/pa-files/PAR-12-183.html "National Institute on Aging Analysis of Alzheimer's Disease Genome Sequencing Project Data [U19]" in developing their research plan. Applicants are encouraged to contact the NIA program director for additional details:

Marilyn M. Miller, Ph.D.
Program Director
Genetics of Alzheimer's Disease
Division of Neuroscience
NIA / NIH / DHHS
7201 Wisconsin Suite 350
Bethesda, Maryland 20892
voice: 1 301-496-9350
Email: millerm@nia.nih.gov

The finalized ADSP plan will be posted by February 18th, 2013.

Over View and Summary of the Alzheimer's Disease Whole Genome Sequencing Project Proposal

Overall objectives:

• Objective 1: Identify novel risk raising genes and alleles for late-onset AD.

• Objective 2: Identify novel protective genes and alleles for late-onset AD.

Four components:

• **Family-based** sequencing (whole genome sequencing of 100 informative multiplex families) to identify genomic regions associated with increased risk of AD

Timeline for data production: March 2013-December 2013

• **Case-control** sequencing (whole exome capture sequencing of 5,000 cases / 5,000 controls) for both risk raising and protective loci; plus whole exome capture sequencing from an additional case group made-up of one individual from 1,000 additional AD-enriched families) to identify regions associates with increased risk or protection from AD

Timeline for data production: June 2013-December 2014

• **Replication and validation** of regions identified from case-control and family sequencing in >50,000 samples by targeted sequencing and/or genotyping

Timeline for data production: June 2014-June 2015

• Deep **targeted** sequencing of candidate AD regions identified by previous linkage and chip-based association GWAS and ADSP sequenced-based, analyses (targeted sequencing in 5,000 cases /5,000 controls and 100 multiplex families) to identify potential functional variants beyond the exomes in regions implicated in and validated for AD risk and protection

Timeline for data production: June 2014-December 2015

Samples

• Samples for the family studies will be drawn from NIA-LOAD family study, NCRAD families and Miami families

• Samples for the case/control study will be drawn from the Alzheimer's Disase Genetics Consortium (ADGC) and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortia

Data All project sequence and other data will be made available through NIAGADS website https://www.niagads.org/) including pre-existing phenotype, GWAS and gene chip data for the samples to be used.
Additional details:

Family study: Whole genome sequencing (WGS) for select subjects from large multigenerational extended families with late onset Alzheimer disease to identify novel genes and alleles associated with the occurrence of late-onset AD. Existing qualifying multiplex families from diverse race/ethnic backgrounds (N=102) have been identified for inclusion.

These families represent several different cohorts including the NIA-Late Onset of Alzheimer's Disease (LOAD) family study, the National Cell Repository for Alzheimer's Disease (NCRAD), Multi-Institutional Research in Alzheimer's Genetic Epidemiology (MIRAGE), Caribbean Hispanic Families collected by Dr. Richard Mayeux, and families from cohorts collected by investigators at the University of Miami, and Vanderbilt University families. The goal is to create an unrelated case series derived from the family dataset. Many of these families have a strong family history of disease but were excluded because there were insufficient numbers of affected individuals with DNA available.

Families were given priority status based on the number of affected individuals and generations affected, absence of known mutations, age, and common risk variants (APOE*4). Five tiers were created to rank the largest families. Initial sequencing will focus on tier 1 and 2 families. Briefly, tier 1 families were identified as having multiple affected individuals with DNA samples (>5 individuals) in two branches (e.g. cousins) who are not "clustered" for the APOE*4 allele. In addition, tier 2 was defined as families that still have a high density of affected individuals and have at least one affected individual without an APOE*4 allele but also may have one or more with the APOE*4 allele so long as the age of onset is less than 72 years.

Over 1500 families were reviewed yielding 256 pedigrees containing four or more affected individuals without known mutations or risk genes. 102 multiplex families (64 tier 1 and 41 tier 2) have been vetted for initial sequencing, but another 76 families (tiers 3 to 5) are available for further consideration based on the results of the tier 1 and tier 2 data analyses.

There are nearly 1,000 additional families which are available but smaller with three or fewer affected family members but with strong family histories. The family selection committee has also been asked to select one individual from each of the remaining families for inclusion in the Case/Control project.

Case / Control: The objective is to use a case-control design and whole exome capture sequencing (WECS) in a sample of European-Americans to identify novel genes and alleles associated with the increased risk of or protection from late-onset AD. Using sex, age and APOE genotype to calculate risk 5,000 case samples have been identified having definite or probably AD who were at lowest risk for AD. Control samples (N=5,000) have been identified as those free of disease having the least

amount of expected misclassification. Based on previous simulations and power calculations under a variety of scenarios, the cases and controls for the risk raising and protective analyses are the same.

Samples for the case-control design will be drawn from two consortia:

> **ADGC.** This consortium currently has available for sequencing10,273 cases and 10,575 controls. The ADGC is continually expanding this sample by collecting new cases and controls form the 29 NIA-funded Alzheimer Disease Centers (ADCs) and by establishing new collaborations with US and foreign researchers.
>
> **CHARGE.** The CHARGE consortium and collaborating cohorts currently have 6941 cases and 43,207 controls of European American ancestry.

For both the power calculations and the actual selection of samples, investigators calculated a "risk score" that included *APOE* genotype, sex and either onset age for cases or age at last exam for controls (age at death for neuropathology controls). For AD cases, the risk score is the incidence of AD at the sample's age at onset for their sex and APOE genotype. For controls, the risk score is the probability of developing AD between the age at last exam and age 85, accounting for their known *APOE* genotypes, sex, age at last exam, and availability of neuropathology information. The risk score was computed in both controls and cases that have readily available genomic DNA (no cell line samples). The risk profile is based on direct genotyping of *APOE.* Investigators did not consider the risk alleles at other AD loci as these have small effects and incorporating these into our sample selection process would limit our ability to detect rare causal variants at these recently identified loci. Details of the risk score calculation are provided below.

Genome-wide searches for rare variants protecting against AD are complicated by misclassification of controls who may have a genetic liability that has not yet manifested clinically. To mitigate this uncertainty, one can select controls based on a risk score that accounts for the probability of developing AD in the future based on population incidence data for sex and age, and the known genetic risk factors for AD. The major determinants of this risk score are age and APOE genotype with a comparatively smaller effect from allelic data for the other 10 genes. Thus, controls who have the lowest probability of converting would be a subgroup of persons who are ages 85 and above and who possess the least genetic liability (i.e., combination of the *APOE* ε2/ε2 genotype and no risk alleles at other AD loci). Among these older controls are also persons who have survived a high genetic risk (based on genotypes at *APOE* and other risk loci) but have escaped the disease, and these persons would be enriched for protective loci.

Power analyses suggest that the gain in power attenuates beyond a sample size of 5,000 AD cases and 5,000 controls for detecting association with rare variants assessed individually or collectively by a gene-based test. This number of AD cases is also the minimum required to confidently detect rare

variants with frequencies as small as 0.1%. All controls will be at least 60 years old and have either clinical assessment for dementia or absence of Alzheimer features upon neuropathology examination. All cases meet criteria for probable or definite AD based on clinical assessment, or had presence of AD features upon neuropathology examination. The control samples are selected as those having the least amount of expected misclassification (misclassification was defined using age at last exam, sex, and APOE genotype specific incidence measures). The case samples are selected to be of lowest risk for AD, based on their age, sex, and APOE status (also calculated using measures of AD incidence).

Using the same control group and a new case group consisting of one individual selected from families aggregating for late onset AD, it is possible to identify novel genes and alleles associated with the increased risk of AD. This "enriched case set" consists of one member of each European-American family. It is anticipated that approximately 1,000 individuals will be in this "enriched" case set, of which 100 will have already been sequenced. A previous study selected the earliest onset case with the most definitive diagnosis of AD from each of 867 NIA-LOAD families. These individuals were then sequenced for pathogenic mutations in the familial AD and frontotemporal dementia genes: *APP, PSEN1, PSEN2, MAPT* and *GRN* (Cruchaga et al., 2012). All individuals were also screened for expansion of the intronic repeat in *C9orf72* that is associated with FTD/ALS (Harms et al., 2012). Sixty-five families were removed based on the presence of a pathogenic mutation in one of these genes. It is anticipated that the remaining mutation negative families will be enriched for novel AD risk alleles.

Replication: Replicate and validate the discoveries made in the above studies in independent samples of European-American AD cases and controls (n>50,000 samples). In addition, evaluate the generalizability of these discoveries in European-Americans in other race/ethnic groups: African-American, Hispanic and Afro-Caribbean, East Asians and genetically isolated populations including the Amish, Icelanders and Israeli-Arabs.

Samples will be drawn from the above CHARGE and ADGC consortia, as well as other consortia both within and separate from the International Genomics Alzheimer Project (IGAP) consortium