Received: 7 February 2023 | Revised: 1 June 2023 | Accepted: 2 June 2023

DOI: 10.1002/alz.13370

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION

RESEARCH ARTICLE

The Early-Onset Alzheimer's Disease Whole-Genome Sequencing Project: Study design and methodology

Nicholas R. Ray[1,2] | Temitope Ayodele[1] | Melissa Jean-Francois[3,4] | Penelope Baez[1] | Victoria Fernandez[5,6] | Joseph Bradley[5,6] | Paul K. Crane[7] | Clifton L. Dalgard[8,9] | Amanda Kuzma[10] | Heather Nicaretta[10] | Rebecca Sims[11] | Julie Williams[11,12] | Michael L. Cuccaro[3,4] | Margaret A. Pericak-Vance[3,4] | Richard Mayeux[1,2,13,14] | Li-San Wang[10] | Gerard D. Schellenberg[10] | Carlos Cruchaga[5,6] | Gary W. Beecham[3,4] | Christiane Reitz[1,2,13,14]

[1]Gertrude H. Sergievsky Center, Columbia University, New York, New York, USA

[2]Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University, New York, New York, USA

[3]The John P. Hussman Institute for Human Genomics, University of Miami, Miami, Florida, USA

[4]Dr. John T. MacDonald Foundation Department of Human Genetics, University of Miami, Coral Gables, Florida, USA

[5]Department of Psychiatry, Neurology and Genetics, Washington University School of Medicine, St. Louis, Missouri, USA

[6]Neurogenomics and Informatic (NGI) Center, Washington University School of Medicine, St. Louis, Missouri, USA

[7]Division of General Internal Medicine, University of Washington, Seattle, Washington, USA

[8]Department of Anatomy, Physiology & Genetics, Uniformed Services University of the Health Sciences, Bethesda, Maryland, USA

[9]The American Genome Center, Uniformed Services University of the Health Sciences, Bethesda, Maryland, USA

[10]Penn Neurodegeneration Genomics Center, Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA

[11]Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK

[12]UK Dementia Research Institute, Cardiff University, Cardiff, UK

[13]Department of Neurology, Columbia University, New York, New York, USA

[14]Department of Epidemiology, Columbia University, New York, New York, USA

Correspondence
Christiane Reitz, Taub Institute for Research on the Aging Brain, Columbia University, New York, NY 10032, USA.
Email: cr2101@cumc.columbia.edu

Gary W. Beecham, John P. Hussman Institute for Human Genomics, Dr. John T. MacDonald Foundation Department of Human Genetics, University of Miami, Miller School of Medicine, Miami, FL, 33136, USA.
Email: gbeecham@med.miami.edu

Funding information
NIH, Grant/Award Numbers: R01AG078964, R01AG064614, RF1AG054080, U24AG056270, U24-AG041689, U54-HG003067, U19AG066567, U01AG006781, U01AG032984; Medical Research Council, Grant/Award Numbers:

Abstract

INTRODUCTION: Sequencing efforts to identify genetic variants and pathways underlying Alzheimer's disease (AD) have largely focused on late-onset AD although early-onset AD (EOAD), accounting for ~10% of cases, is largely unexplained by known mutations, resulting in a lack of understanding of its molecular etiology.

METHODS: Whole-genome sequencing and harmonization of clinical, neuropathological, and biomarker data of over 5000 EOAD cases of diverse ancestries.

RESULTS: A publicly available genomics resource for EOAD with extensive harmonized phenotypes. Primary analysis will (1) identify novel EOAD risk loci and druggable targets; (2) assess local-ancestry effects; (3) create EOAD prediction models; and (4) assess genetic overlap with cardiovascular and other traits.

DISCUSSION: This novel resource complements over 50,000 control and late-onset AD samples generated through the Alzheimer's Disease Sequencing Project (ADSP).

G0801418/1, MR/K013041/1, MR/L023784/1; Welsh Assembly Government, Grant/Award Number: SGR544:CADR; Moondance Charitable Foundation

The harmonized EOAD/ADSP joint call will be available through upcoming ADSP data releases and will allow for additional analyses across the full onset range.

KEYWORDS
Alzheimer's disease, early-onset Alzheimer's disease, genetics, study design, whole-genome sequencing

Highlights

- Sequencing efforts to identify genetic variants and pathways underlying Alzheimer's disease (AD) have largely focused on late-onset AD although early-onset AD (EOAD), accounting for ~10% of cases, is largely unexplained by known mutations. This results in a significant lack of understanding of the molecular etiology of this devastating form of the disease.
- The Early-Onset Alzheimer's Disease Whole-genome Sequencing Project is a collaborative initiative to generate a large-scale genomics resource for early-onset Alzheimer's disease with extensive harmonized phenotype data.
- Primary analyses are designed to (1) identify novel EOAD risk and protective loci and druggable targets; (2) assess local-ancestry effects; (3) create EOAD prediction models; and (4) assess genetic overlap with cardiovascular and other traits.
- The harmonized genomic and phenotypic data from this initiative will be available through NIAGADS.

1 | BACKGROUND

Although aging is the predominant biological risk factor for developing Alzheimer's disease (AD), about 5% to 10% of cases (eg, ~250,000 individuals) in the US alone[1] show symptom onset before 65 years and are classified as early-onset Alzheimer's disease (EOAD). A subset of EOAD cases are clinically similar to late-onset AD (LOAD), with predominant cognitive impairment in the memory domain.[2] However, EOAD tends to be more aggressive in its course[3,4] and shows a higher prevalence of atypical clinical features and impairment in other cognitive domains including impairment in executive dysfunction, apraxia, dyscalculia, visual dysfunction, and aphasia (fluent and non-fluent).[5–7] In line with these differences in clinical presentation, individuals with EOAD often show different profiles on brain imaging and neuropathological assessment, even at a similar stage of clinical impairment. EOAD tends to show less atrophy and neuropathological changes in medial temporal lobe structures (ie, hippocampus and entorhinal cortex) but more widespread and faster progressing cortical atrophy and hypometabolism, and a higher degree of tau pathology in neocortical regions.[8–13]

The early onset and clinical heterogeneity result in particularly detrimental medical, emotional, social, and financial consequences for patients and their families. Individuals with EOAD often receive a significantly delayed diagnosis,[14] are misdiagnosed with other psychiatric/neurodegenerative diseases such as frontotemporal dementia,[15–20] and are often excluded from clinical research trials[21] resulting in stigmatization and inadequate access to treatment, disease education, and patient and caregiver support resources. Disease onset during the prime earning years frequently results in significant deprivation of income, loss of employment, health insurance, and retirement benefits.

While EOAD has a considerable genetic basis with a heritability of over 90%,[22] variation in known EOAD genes (including *APP*, *PSEN1*, *PSEN2*) accounts for only 5% to 10% of cases.[23,24] Most cases are either sporadic or follow a non-Mendelian pattern of inheritance but are expected to be enriched for causative genetic factors.

Identifying this missing heritability is essential to understand the molecular mechanisms underlying this devastating form of the disease and identify more effective targets for screening, prevention, and treatment. However, individuals with EOAD have been significantly underrepresented in the major genomic efforts of AD. The leading national effort, the Alzheimer's Disease Sequencing Project (ADSP)[25] and its follow up study (Alzheimer's Disease Sequencing Project–Follow Up Study; ADSP-FUS),[26] focus mostly on the late-onset form of disease. The Dominantly Inherited Alzheimer Network is restricted to autosomal dominant EOAD accounting for only 2% to 5% of EOAD cases.[27] The Alzheimer's Disease Neuroimaging Initiative[28] and the Longitudinal Early-onset Alzheimer's Disease Study[29] are designed to track the progression of AD across disease stages with clinical, imaging, and biospecimen biomarkers. They are not, however, necessarily designed for gene discovery using large sample sizes.

To facilitate EOAD variant discovery we have implemented the Early-Onset Alzheimer's Disease Whole-genome Sequencing Project (R01AG064614), a collaborative initiative to generate and analyze a large-scale genomics resource for EOAD comprising several thousand EOAD cases of diverse ancestry. These case-control data will be complemented by whole-genome sequencing (WGS) data generated for over 200 multiplex families loaded for EOAD through complementary efforts (RF1AG054080, U24AG056270). Application of ADSP pipelines for processing and harmonization of genomic and phenotype data across all datasets ensures compatibility with ADSP and ADSP-FUS efforts. Inclusion of diverse ancestries will allow us to identify EOAD variants not detectable in individuals of European ancestry, providing critical information on mechanisms underlying EOAD subtypes and observed health disparities. Primary specific goals are to (1) create a publicly available large-scale genomics resource for EOAD with WGS data generated and processed using ADSP pipelines and extensive harmonized phenotype data; (2) identify novel genomic EOAD risk loci and loci modulating age at onset and decline in specific cognitive domains; (3) assess the role of polygenic and local-ancestry effects in EOAD etiology; (4) create EOAD-specific prediction models; (5) assess genetic overlap with cardiovascular and other potentially associated traits; and (6) identify druggable targets.

RESEARCH IN CONTEXT

1. **Systematic Review**: Relevant literature and related efforts were screened by reviewing PubMed, NIAGADS, and dbGaP for efforts on early-onset Alzheimer's disease (EOAD).
2. **Interpretation**: EOAD has been largely excluded from major AD genomics efforts, resulting in an extensive lack of understanding of its underlying molecular etiology. Generation of a large-scale EOAD whole-genome resource will allow for identification of genetic variants, genes, and molecular pathways underlying this form of AD.
3. **Future Directions**: Integration of the generated EOAD resource with the ADSP, ADSP-FUS, and related large-scale AD genomics, multi-omics and functional genomics efforts across a range of diverse ancestries will readily allow for examination of several additional critical hypotheses including mechanisms underlying changes in blood biomarkers, neuropathological measures, and structural and functional brain imaging phenotypes across the full spectrum of age at onset strata.

2 | METHODS

2.1 | Study design

The Early-Onset Alzheimer's Disease Whole-genome Sequencing Project is a collaborative large-scale WGS effort on EOAD led by the Taub Institute for Research on the Aging Brain at Columbia University, the Hussman Institute for Human Genomics at the University of Miami, and the NeuroGenomics and Informatics Center at Washington University School of Medicine in St. Louis in collaboration with the Alzheimer's Disease Genetics Consortium. The project leverages existing sample ascertainment, sample processing, and data generation and processing pipelines by major AD research centers. EOAD samples and extensive phenotype data were obtained from the NIH-funded Alzheimer's Disease Research centers (ADCs) via the National Centralized Repository for Alzheimer's Disease and Related Dementias and the National Alzheimer's Coordinating Center (NACC), with ADCs at Columbia University, University of Miami, Washington University School of Medicine in St. Louis, the Adult Changes of Thought Study,[30] and other sites providing additional samples. Descriptions of the individual cohorts can be found in the Supplemental Material. WGS data were generated at The American Genome Center (TAGC) at the Uniformed Services University of the Health Sciences (USUHS). Sequence data are being quality controlled, harmonized, and jointly called through the Genome Center for Alzheimer's Disease (GCAD) employing bioinformatics protocols implemented through the ADSP. Where missing, genome-wide association study data on the same samples are being generated, quality controlled, and imputed to the latest ancestry-specific reference panels.

2.2 | Inclusion/exclusion criteria

Included in the Early-Onset Alzheimer's Disease Whole-genome Sequencing Project are cognitively healthy individuals, and individuals with EOAD or early-onset mild cognitive impairment (MCI) of diverse ancestries with an age of onset <65 years meeting National Institute on Aging (NIA) criteria for AD or MCI.[31,32] While baseline age required for controls is 60 years, 96% of the control samples are 70 years or older, and mean age at last evaluation is 85 years. Both EOAD cases with predominant amnestic impairment as well as cases with predominant impairment in other cognitive domains (ie, non-amnestic presentation) and atypical presentation are included, allowing to identify the genetic variation underlying EOAD subtypes. A subset of individuals have a definite AD diagnosis through brain autopsy based on Braak and the Consortium to Establish a Registry for Alzheimer's Disease's criteria, or have cerebrospinal fluid (CSF), plasma, or imaging biomarkers.[33,34] Cases with competing diagnoses (Parkinson's disease, Huntington's disease, frontotemporal dementia, vascular dementia, depression, etc) or with a known mutation in *APP*, *PSEN1*, or *PSEN2* are excluded from the effort. For all samples selected and whole-genome sequenced for this project sequence, data in these genes are scrutinized ahead of any further downstream analyses to identify any additional samples potentially carrying pathogenic variants in these genes. All participants have provided informed consent according to the Declaration of Helsinki and the policies of the respective institutional review boards at the contributing centers.

2.3 | Ancestral diversity

The study sample specifically includes individuals of diverse ancestry—African American (AA), Hispanic (HI), non-Hispanic White, Asian. It is clear that genetic ancestry plays a critical role in complex diseases and observed health disparities in AD.[35–38] Persons of AA and HI ancestry have up to twice the incidence of AD as Non-Hispanic White (NHW)[39] individuals, and heritability of AD differs between ethnic groups.[37,40] Alleles in known AD genes (eg, *APOE* and *ABCA7*, among others) account for some disease risk variability. African ancestry-specific AD risk variants in *ABCA7*, *TREM2*, and other genes have been described by our group and others,[35,41,42] along with loci specific to HI individuals.[43–47] For *ABCA7* in particular, a 44 bp deletion is strongly associated with AD in those of AA[35] ancestry and is also present in HI individuals with a high proportion of African global ancestry (41.8%),[48] while other rare truncating and splice altering variants confer risk in the NHW population.[49–51] This suggests that some AD risk/protective variants will have European origins while others will have African or Native American origins, and still other variants may be rare, recent in origin, and unique to individual populations. These findings underscore the importance of investigating diverse populations for ancestry-specific AD risk variants, and the sample included in this project will allow to assess the ancestral background at identified genetic loci associated with EOAD.

3 | RESULTS

3.1 | Whole-genome sequencing (WGS) and downstream bioinformatics processing

In total, the Early-Onset Alzheimer's Disease Whole-genome Sequencing Project has sequenced 4097 EOAD and early-onset MCI samples meeting our minimum inclusion criteria (affection status, age at onset, sex, and adequate DNA). In addition, samples from 1109 elderly cognitive controls have been sequenced through this effort, selected to match case samples by sex and ancestry yielding the largest EOAD genomics resource to date. These samples complement over 50,000 control and LOAD samples, generated with similar protocols through the ADSP Discovery and Follow-up Studies and related efforts, all with harmonized phenotype and genomics data allowing for additional analyses across the full range of onset, including analyses of factors modulating age of onset and shared genetic heritability across age at onset groups. This harmonized EOAD/ADSP joint call will be available through the upcoming fifth ADSP data release.

3.1.1 | Sequencing library preparation and whole-genome sequencing

Sequencing library preparation and WGS of samples missing WGS data was performed through TAGC at the USUHS. The USUHS has extensive experience in large-scale WGS workflows, including several large consortia-based sequencing efforts (NIA Alzheimer's Disease Sequencing Program, National Institute of Mental Health Army Study to Assess Risk and Resilience in Servicemembers–Longitudinal Study, the National Institute of Neurological Disorders and Stroke Dementia Resolution Study, Applied Proteogenomics Organizational Learning and Outcomes, etc). Samples were assessed for quantity (Quant-iT PicoGreen dsDNA assay) by concentration. Sequencing libraries were prepared using the TruSeq PCR-Free Library Prep kit (Illumina) with unique dual index adapters and quantified using quantitative polymerase chain reaction (qPCR; KAPA Library Quant Kit). Libraries were normalized to 4 nM into a 24 to 26 sample pools. Pool concentration was quantified using qPCR and clustered onboard the NovaSeq 6000 platform (Illumina) with sequencing runs conducted on an S4 flow cell with paired-end 150 bp read length. After sequencing de-multiplexing was performed (bcl2fastq v2.20) and resequencing analysis on a quality assurance (QA) workflow (Illumina HAS2.2); data were reviewed for yield, read alignment percentage, bases greater than Q30, percent read duplicates, Picard mean coverage and contamination (FREEMIX < 0.05 by verifyBamID). QA-passing genomes were inventoried for data transfer of FASTQ sets to the Genome Center for Alzheimer's Disease (GCAD).

3.1.2 | Bioinformatics processing of WGS data

Sequence data generated by USUHS were processed and joint called by GCAD using the VCPA[52] pipeline developed by GCAD for ADSP-related projects. The approach uses Genome Analysis Toolkit (GATK)[53,54] for single nucleotide variant (SNV)/Indel calling. The workflow includes mapping reads to hg38, sorting in BAM format, duplicate marking, quality scores, and local read realignments around known indels. GATK HaplotypeCaller is then applied to generate individual genotype calls in genomic and project-level VCF formats.

3.1.3 | Quality control of WGS data

In line with ADSP efforts, variant-level quality metrics include VQSR quality tranches, call rates, average read depths, excessive read depths (>500 reads), and excess heterozygosity or departure from Hardy-Weinberg equilibrium.[25,55] Sample-level quality control includes within-sample genotype call rate, Ti/Tv ratio for SNVs, heterozygosity/homozygosity ratio, and excess burden of singleton/doubleton variants. The joint-genotype format called pVCF will be annotated using the pipeline which generates variant-level assessments of functional impact on genes and genetic regulation. Our pipeline is based upon the Ensembl Variant Effect Predictor, which overlays exon, transcript, and regulatory element information from the Ensembl database to generate all possible consequences (missense, frameshift, splicing, etc) a variant may have. Variant consequences relative to Ensembl/GENCODE transcripts are assigned an impact category (high, moderate, low, etc), and multiple variant scoring approaches are incorporated (CADD, REVEL, CATO, etc).

3.2 | Clinical and cognitive assessment and phenotype data harmonization

All individuals from all contributing sites have completed standard clinical assessments that include self-reporting, informant reporting, medical records, and direct assessment information. Additionally, past medical history, family history, and detailed neurological data have been obtained. A total of 3868 of the 5206 cases and controls were recruited from ADCs and have NACC Uniform Data Set assessments. Study personnel at the contributing sites conduct the clinical assessment, including an interview to assess subjective neuropsychiatric symptoms that pertain to activities of daily living, cognition, and mood. Disease history is collected, a score on the Clinical Dementia Rating Scale is calculated to assign degree of severity, a neurological examination is conducted, and extensive cognitive test batteries are employed.

To ensure compatibility across datasets and with the leading LOAD sequencing efforts such as the ADSP Discovery and ADSP-FUS, we will compile, harmonize, and generate phenotypes, subphenotypes (AD diagnoses), cognitive measures, demographics, stage, age-at-onset (AAO; case) or age-at-examination (AAE; control), sex, race/ethnicity, and genomic data across all datasets. All phenotype data were checked for quality, integrity, and consistency, and we have developed a common coding scheme to match covariates and value formats (eg, range and precision for continuous values, and codes for categorical data) from the different studies. We will recode the data using standardized measures whenever possible. We will compare summary statistics/distributions across studies and will conduct outlier studies to identify any potential coding errors or data collection bias.

3.2.1 | Diagnosis and AAO/AAE

We will utilize established criteria for the diagnosis of AD which are available in all cohorts. The diagnoses of mild cognitive impairment and probable/possible AD will be made using the NIA Alzheimer's Association workgroup diagnostic guidelines[31,32] based on the in-person assessment by the study staff and norms based on age, education, and ethnic group. To assess AAO, we require information from a knowledgeable caregiver or family member concerning when the person manifested constant forgetfulness resulting in an inability to manage his schedule or daily activities. For normal controls without cognitive impairment, AAE will be the age when the individual was last examined.

3.2.2 | Harmonization of neuropsychological data

Harmonization of neuropsychological data will be done in collaboration with the ADSP Phenotype Harmonization Consortium (U24AG074855). Each individual dataset has an extensive cognitive battery examining a variety of cognitive functions including memory, visuospatial awareness, language, and executive function. We will evaluate the internal consistency of each study's battery using Cronbach's α.[56] To derive harmonized composite scores for cognitive domains across cohorts (memory, visuospatial awareness, language, executive function), we will employ modern psychometric methods to the pooled sample, which tend to have better validity than scores derived from standard approaches, and are specifically recommended for genetic analyses.[57–62] Using information from time of first diagnosis for cases and last visit for cognitively healthy controls, we will recode observed item responses to avoid sparse response categories, preserving variability at the extremes of the distribution. Separately for cases only and the total sample, we will then fit, for each domain, factor analyses (single factor models assuming no residual relationships, and bifactor models assuming covariance by cognitive subdomains or methods effects).[63] To determine which model is superior, we will compare single factor and bi-factor models for both sets of samples assessing the correlation between factor scores, compare the loadings for each indicator on the overall domain factor with and without the secondary domain structure, and use fit statistics.[64] Missing data will be handled using full information maximum likelihood estimation.[65]

3.2.3 | Functional impairment categories for cognitive domains

Thresholds to define "substantial" relative impairments will be calculated as previously described.[66] This will create, for each subject, labels reflecting the predominant EOAD subtype (ie, AD-Memory, AD-Executive, AD-Language, AD-Visuospatial, AD-No Domains, and AD-Multiple Domains). We will also analyze groups with a prominent or neutral memory impairment versus those with relatively intact memory (ie, AD-Memory, AD-No Domains, and AD-Multiple Domains, vs the other three subtypes). These constructed categorical variables will provide a set of harmonized measures differently capturing cognitive impairment that can be readily used in genomic[67] and clinical[68] analyses.

3.2.4 | AD biomarkers normalization

A subset of participants have been recruited on this study through CSF, plasma (Aβ, tau, ptau, TREM2, NFL, SNAP25), or amyloid imaging (see Table 1), and we expect this number to further increase through complementary efforts. As these biomarkers have been generated in different centers using different platforms it is not possible to simply combine the data across studies. We have developed robust approaches to harmonize AD biomarkers across datasets.[69,70] Briefly, normalized z-scores are calculated by using the mean and standard deviation units across each cohort and applied to the entire endophenotype in order to account for within cohort variation. Then, we used a mixture modeling, which is a statistical method for estimating subpopulations within an overall group, to determine the biomarker-positive and negative individuals. We assume that there are two normally distributed subgroups within each dataset. Using an expectation-maximization algorithm in the R package mixtools v1.0.4,[71] we can

TABLE 1 Demographic and clinical characteristics of Early-Onset Alzheimer's Disease Whole-genome Sequencing Project samples sequenced to date.

	Affected	Unaffected
Individuals, n	4097	1109
Early-onset MCI	541	–
Early-onset AD	3490	–
Early-onset other dementia	66	–
Female, n (%)	2178 (53.16)	695 (62.67)
Age at last evaluation (years), mean	69.46	84.90
Age at onset (years), mean	61.21	–
Early-onset MCI	64.57	–
Early-onset AD	60.73	–
Early-onset other dementia	58.98	–
Ethnicity		
NHW	3506	962
HI	310	77
AA	171	69
Other	99	1
Unknown	11	0
CDR		
0	37[a]	899
0.5	917	48
1	902	4
> = 2	1803	2
% CSF biomarkers	6.66	8.39
% plasma biomarkers	6.47	12.17

Abbreviations: AA, African American; AD, Alzheimer's disease; CDR, Clinical Dementia Rating Scale; CSF, cerebrospinal fluid; HI, Hispanic; MCI, mild cognitive impairment; NHW, Non-Hispanic White.
[a]These individuals are affected with MCI and have a clinical judgment of impaired cognition.

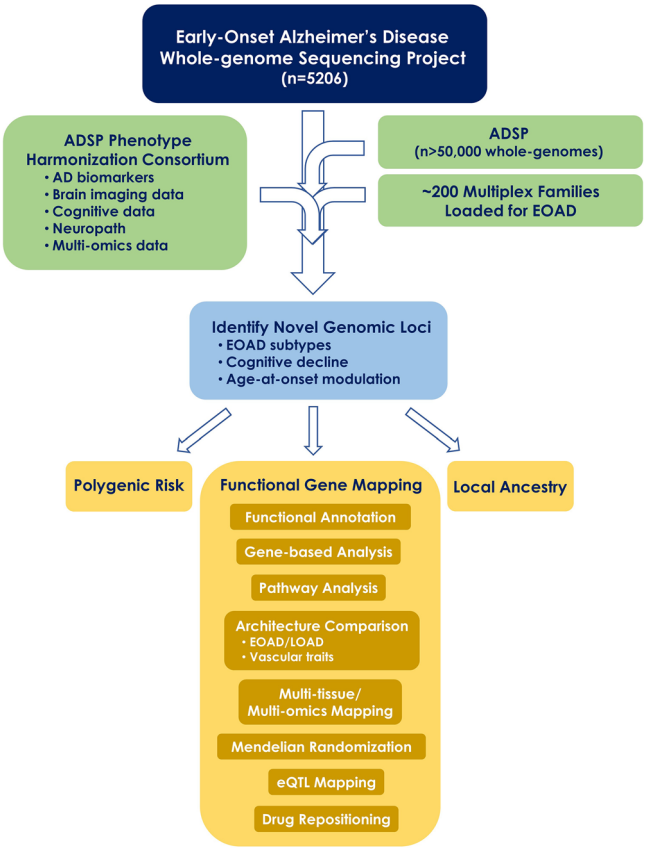

FIGURE 1 Project flow and primary aims of the Early-Onset Alzheimer's Disease Whole-genome Sequencing Project. The 5206 samples (4097 cases and 1109 controls) sequenced by the Early-Onset Alzheimer's Disease Whole-genome Sequencing Project will be integrated with over 50,000 whole-genomes collected by the Alzheimer's Disease Sequencing Project (ADSP) and 200 multiplex families loaded for early-onset Alzheimer's Disease (EOAD). Phenotype information will be harmonized using AD biomarkers, brain imaging, cognitive, neuropath, and multi-omics data. First-pass analyses will be conducted to identify novel loci associated with EOAD, cognitive decline, and age-at-onset modulation. Follow-up analyses will be conducted to assess polygenic risk, mechanistic pathways, comparison of the genetic architecture between EOAD and late-onset Alzheimer's Disease (LOAD) and vascular traits, and to examine local ancestry. eQTL, expression quantitative trait loci.

calculate estimated means, standard deviations, and subgroup proportions for each study. We can calculate the intersection of the estimated Gaussian curves. Based on the assumption of two univariate normal distributions within each study we will obtain two estimated means ($\mu 1$ and $\mu 2$), two estimated standard deviations ($\sigma 1$ and $\sigma 2$), and two estimated mixing proportions. From these models we can determine biomarker status for each of the specific analytes, and perform further analyses.[69,70]

4 | DISCUSSION

Besides creating a publicly available large-scale genomics resource for genetic research on EOAD and AAO with extensive harmonized phenotype and biomarker data, the Early-Onset Alzheimer's Disease Whole-genome Sequencing Project has several immediate analysis goals. Harmonized WGS data will be scrutinized with a wide array of computational tools and statistical approaches to identify novel genomic risk and protective loci for EOAD subtypes, decline in specific cognitive domains, and loci modulating AAO. These analyses include single variant, gene-based, and sliding window analyses and are expected to identify novel genes, pathways, and etiologic mechanisms that are shared or specific to a particular ethnic group (see Figure 1). Comparison with WGS data on multiplex EOAD families generated by our groups will allow us to determine which variants are associated with familial versus sporadic disease. Incorporation of LOAD genomic data will allow us to identify loci, genes, and mechanistic pathways modulating AAO, map genomic loci shared between the early- and late-onset forms, and calculate the extent of shared EOAD/LOAD heritability. These findings will be pivotal steps in dis-

entangling the genetic and mechanistic overlap with the late-onset form and clarifying whether both forms are in fact distinct disease entities. A wide array of computational approaches such as linkage disequilibrium score regression, genome-wide complex trait analysis, colocalization, and Mendelian randomization approaches coupled with extensive multi-tissue multi-omics data available to us will be employed to infer causality of identified variants and genes, and identify potential druggable targets. Second, we will comprehensively assess the role of polygenic effects by aiming to understand EOAD etiology and EOAD genetic risk, develop effective screening tools, and identify druggable targets. It is critical to determine whether EOAD and its subtypes are the result of rare genetic variation of strong effect, or if it is highly polygenic with weak effects of individual variants. Analyses of networks and polygenic effects in the late-onset form of AD have identified several specific gene-sets that seem to influence disease including immune response, inflammation, and endocytosis.[72,73] To test these polygenic hypotheses in EOAD, we will perform in-depth network and pathway-based tests of association, and construct and assess the utility of genetic risk scores (calculated by summing an individual's genome-wide genotypes weighted by their corresponding z-scores) employing state-of-the-art methods specifically developed for these analyses. Risk score sub–analyses will include non-genetic factors as this can improve predictive power of polygenic scores significantly.[74] Third, we will comprehensively assess the role of ancestry in EOAD and its subtypes, capitalizing on the rich diversity of this dataset. All analyses will be conducted within and across ancestry groups, and we will utilize a wide array of tools to assess global ancestry, local ancestry, admixture, and the evolutionary history of identified risk and protective alleles. These analyses will determine variants, loci, and pathways that are shared across ethnic groups, as well as variants that are specific to a particular ethnic group. The results will provide pivotal information for development of personalized preventive and therapeutic measures, and disentangling observed health disparities. Fourth, we will assess genetic overlap with traits potentially sharing or impacting etiologic mechanisms such as cardiovascular disease by employing computational approaches developed to determine shared heritability. Finally, bioinformatics and phenotype harmonization protocols in line with the ADSP and ADSP-FUS studies will allow for joint examination across these efforts allowing an extensive array of additional critical hypotheses to disentangle EOAD etiology, including assessment of blood biomarkers, neuropathological changes, and structural and functional brain imaging phenotypes across the full spectrum of AAO strata. The harmonized EOAD/ADSP joint call will be available through upcoming ADSP data releases via the NIA Genetics of Alzheimer's Disease Data Storage Site (NIAGADS; https://dss.niagads.org/).

ACKNOWLEDGMENTS

All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the National Institute on Aging or the National Institutes of Health. We thank the participants of the various Alzheimer's Disease Research Centers, study participants from Columbia University, University of Miami, the Knight ADRC, the Adult Changes in Thought (ACT) study and other ADRCs for the data they have provided and the many investigators and staff who steward that data. Details on the ACT research program can be found at https://actagingresearch.org/about/research-program. This study was supported by NIH grants R01AG078964, R01AG064614, RF1AG054080, U24AG056270, U24-AG041689, U54-HG003067, U19AG066567, U01AG006781, and U01AG032984. Cardiff University was supported by the Medical Research Council (MRC; grants G0801418/1, MR/K013041/1, MR/L023784/1), the Welsh Assembly Government (grant SGR544:CADR) and a donation from the Moondance Charitable Foundation.

CONFLICT OF INTEREST STATEMENT

Carlos Cruchaga has received research support from GSK and EISAI. The funders of the study had no role in the collection, analysis, or interpretation of data; in the writing of the report; nor in the decision to submit the paper for publication. Carlos Cruchaga is a member of the advisory board of Vivid Genomics and Circular Genomics. There were no other potential conflicts. Author disclosures are available in the supporting information.

CONSENT STATEMENT

All human subjects provided informed consent.

DATA AVAILABILITY STATEMENT

The data have been deposited at NIAGADS (https://dss.niagads.org/) and are available as qualified access. To request access to the dataset, researchers can submit a data access request to NIAGADS for the ADSP dataset (ng00067; https://dss.niagads.org/datasets/ng00067/).

REFERENCES

1. Alzheimer's Association. 2019 Alzheimer's disease facts and figures. *Alzheimers Dement*. 2019;15(3):321-387.
2. Reitz C, Brayne C, Mayeux R. Epidemiology of Alzheimer disease. *Nat Rev Neurol*. 2011;7(3):137-152.
3. Jacobs D, Sano M, Marder K, et al. Age at onset of Alzheimer's disease: relation to pattern of cognitive dysfunction and rate of decline. *Neurology*. 1994;44(7):1215-1220.
4. Seltzer B, Sherwin I. A comparison of clinical features in early- and late-onset primary degenerative dementia. One entity or two? *Arch Neurol*. 1983;40(3):143-146.
5. van der Flier WM, Pijnenburg YA, Fox NC, Scheltens P. Early-onset versus late-onset Alzheimer's disease: the case of the missing APOE ε4 allele. *Lancet Neurol*. 2011;10(3):280-288.
6. Ryan NS, Rossor MN. Correlating familial Alzheimer's disease gene mutations with clinical phenotype. *Biomark Med*. 2010;4(1):99-112.
7. Bateman RJ, Aisen PS, De Strooper B, et al. Autosomal-dominant Alzheimer's disease: a review and proposal for the prevention of Alzheimer's disease. *Alzheimers Res Ther*. 2011;3(1):1.
8. Cho H, Jeon S, Kang SJ, et al. Longitudinal changes of cortical thickness in early- versus late-onset Alzheimer's disease. *Neurobiol Aging*. 2013;34(7):1921 e9-1921 e15.
9. La Joie R, Visani AV, Baker SL, et al. Prospective longitudinal atrophy in Alzheimer's disease correlates with the intensity and topography of baseline tau-PET. *Sci Transl Med*. 2020;12(524):eaau5732.
10. Migliaccio R, Agosta F, Possin KL, et al. Mapping the progression of atrophy in early- and late-onset Alzheimer's disease. *J Alzheimers Dis*. 2015;46(2):351-364.

11. Moller C, Vrenken H, Jiskoot L, et al. Different patterns of gray matter atrophy in early- and late-onset Alzheimer's disease. *Neurobiol Aging*. 2013;34(8):2014-2022.
12. Rabinovici GD, Furst AJ, Alkalay A, et al. Increased metabolic vulnerability in early-onset Alzheimer's disease is not related to amyloid burden. *Brain*. 2010;133(2):512-528. Pt.
13. Yasuno F, Imamura T, Hirono N, et al. Age at onset and regional cerebral glucose metabolism in Alzheimer's disease. *Dement Geriatr Cogn Disord*. 1998;9(2):63-67.
14. van Vliet D, de Vugt ME, Bakker C, et al. Time to diagnosis in young-onset dementia as compared with late-onset dementia. *Psychol Med*. 2013;43(2):423-432.
15. Beach TG, Monsell SE, Phillips LE, Kukull W. Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005-2010. *J Neuropathol Exp Neurol*. 2012;71(4):266-273.
16. Crutch SJ, Lehmann M, Schott JM, Rabinovici GD, Rossor MN, Fox NC. Posterior cortical atrophy. *Lancet Neurol*. 2012;11(2):170-178.
17. Alladi S, Xuereb J, Bak T, et al. Focal cortical presentations of Alzheimer's disease. *Brain*. 2007;130(10):2636-2645. Pt.
18. Forman MS, Farmer J, Johnson JK, et al. Frontotemporal dementia: clinicopathological correlations. *Ann Neurol*. 2006;59(6):952-962.
19. Kertesz A, McMonagle P, Blair M, Davidson W, Munoz DG. The evolution and pathology of frontotemporal dementia. *Brain*. 2005;128(9):1996-2005. Pt.
20. Varma AR, Snowden JS, Lloyd JJ, Talbot PR, Mann DM, Neary D. Evaluation of the NINCDS-ADRDA criteria in the differentiation of Alzheimer's disease and frontotemporal dementia. *J Neurol Neurosurg Psychiatry*. 1999;66(2):184-188.
21. Szigeti K, Doody RS. Should EOAD patients be included in clinical trials? *Alzheimers Res Ther*. 2011;3(1):1-5.
22. Wingo TS, Lah JJ, Levey AI, Cutler DJ. Autosomal recessive causes likely in early-onset Alzheimer disease. *Arch Neurol*. 2012;69(1):59-64.
23. Campion D, Dumanchin C, Hannequin D, et al. Early-onset autosomal dominant Alzheimer disease: prevalence, genetic heterogeneity, and mutation spectrum. *Am J Hum Genet*. 1999;65(3):664-670.
24. Janssen JC, Beck JA, Campbell TA, et al. Early onset familial Alzheimer's disease: mutation frequency in 31 families. *Neurology*. 2003;60(2):235-239.
25. Sims R, van der Lee SJ, Naj AC, et al. Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. *Nat Genet*. 2017;49(9):1373-1384.
26. Mena PR, Kunkle BW, Faber KM, et al. The Alzheimer's Disease Sequencing Project – Follow Up Study (ADSP-FUS): increasing ethnic diversity in Alzheimer's genetics research with the addition of potential new cohorts. *Alzheimers Dement*. 2020;16:e046400.
27. Morris JC, Aisen PS, Bateman RJ, et al. Developing an international network for Alzheimer research: the Dominantly Inherited Alzheimer Network. *Clin Investig (Lond)*. 2012;2(10):975-984.
28. Weiner MW, Aisen PS, Jack Jr CR, et al. The Alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimers Dement*. 2010;6(3):202-211.e7.
29. Apostolova LG, Aisen P, Eloyan A, et al. The longitudinal early-onset Alzheimer's disease study (LEADS): framework and methodology. *Alzheimers Dement*. 2021;17(12):2043-2055.
30. Kukull WA, Higdon R, Bowen JD, et al. Dementia and Alzheimer disease incidence: a prospective cohort study. *Arch Neurol*. 2002;59(11):1737-1746.
31. McKhann GM, Knopman DS, Chertkow H, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*. 2011;7(3):263-269.
32. Albert MS, DeKosky ST, Dickson D, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*. 2011;7(3):270-279.
33. Grober E, Qi Q, Kuo L, Hassenstab J, Perrin RJ, Lipton RB. The free and cued selective reminding test predicts Braak stage. *J Alzheimers Dis*. 2021;80(1):175-183.
34. Morris JC, Heyman A, Mohs RC, et al. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology*. 1989;39(9):1159-1165.
35. Cukier HN, Kunkle BW, Vardarajan BN, et al. ABCA7 frameshift deletion associated with Alzheimer disease in African Americans. *Neurol Genet*. 2016;2(3):e79.
36. Hohman TJ, Cooke-Bailey JN, Reitz C, et al. Global and local ancestry in African-Americans: implications for Alzheimer's disease risk. *Alzheimers Dement*. 2016;12(3):233-243.
37. Rajabli F, Feliciano BE, Celis K, et al. Ancestral origin of ApoE epsilon4 Alzheimer disease risk in Puerto Rican and African American populations. *PLoS Genet*. 2018;14(12):e1007791.
38. Vardarajan BN, Faber KM, Bird TD, et al. Age-specific incidence rates for dementia and Alzheimer disease in NIA-LOAD/NCRAD and EFIGA families: national Institute on Aging Genetics Initiative for Late-Onset Alzheimer Disease/National Cell Repository for Alzheimer Disease (NIA-LOAD/NCRAD) and Estudio Familiar de Influencia Genetica en Alzheimer (EFIGA). *JAMA Neurol*. 2014;71(3):315-323.
39. Kohli MA, Cukier HN, Hamilton-Nelson KL, et al. Segregation of a rare TTC3 variant in an extended family with late-onset Alzheimer disease. *Neurol Genet*. 2016;2(1):e41.
40. Bussies PL, Rajabli F, Griswold A, et al. Use of local genetic ancestry to assess TOMM40-523' and risk for Alzheimer disease. *Neurol Genet*. 2020;6(2):e404.
41. Reitz C, Jun G, Naj A, et al. Variants in the ATP-binding cassette transporter (ABCA7), apolipoprotein E 4,and the risk of late-onset Alzheimer disease in African Americans. *JAMA*. 2013;309(14):1483-1492.
42. Schindler SE, Cruchaga C, Joseph A, et al. African Americans have differences in CSF soluble TREM2 and associated genetic variants. *Neurol Genet*. 2021;7(2):e571.
43. Cheng R, Tang M, Martinez I, et al. Linkage analysis of multiplex Caribbean Hispanic families loaded for unexplained early-onset cases identifies novel Alzheimer's disease loci. *Alzheimers Dement*. 2018;10:554-562.
44. Barral S, Cheng R, Reitz C, et al. Linkage analyses in Caribbean Hispanic families identify novel loci associated with familial late-onset Alzheimer's disease. *Alzheimers Dement*. 2015;11(12):1397-1406.
45. Tang M, Alaniz ME, Felsky D, et al. Synonymous variants associated with Alzheimer disease in multiplex families. *Neurol Genet*. 2020;6(4):e450.
46. Vardarajan BN, Barral S, Jaworski J, et al. Whole genome sequencing of Caribbean Hispanic families with late-onset Alzheimer's disease. *Ann Clin Transl Neurol*. 2018;5(4):406-417.
47. Tosto G, Fu H, Vardarajan BN, et al. F-box/LRR-repeat protein 7 is genetically associated with Alzheimer's disease. *Ann Clin Transl Neurol*. 2015;2(8):810-820.
48. Bryc K, Auton A, Nelson MR, et al. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci U S A*. 2010;107(2):786-791.
49. Kunkle BW, Carney RM, Kohli MA, et al. Targeted sequencing of ABCA7 identifies splicing, stop-gain and intronic risk variants for Alzheimer disease. *Neurosci Lett*. 2017;649:124-129.
50. De Roeck A, Van den Bossche T, van der Zee J, et al. Deleterious ABCA7 mutations and transcript rescue mechanisms in early onset Alzheimer's disease. *Acta Neuropathol*. 2017;134(3):475-487.

51. Cuyvers E, De Roeck A, Van den Bossche T, et al. Mutations in ABCA7 in a Belgian cohort of Alzheimer's disease patients: a targeted resequencing study. *Lancet Neurol*. 2015;14(8):814-822.
52. Leung YY, Valladares O, Chou YF, et al. VCPA: genomic variant calling pipeline and data management tool for Alzheimer's Disease Sequencing Project. *Bioinformatics*. 2019;35(10):1768-1770.
53. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a Mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303.
54. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43(5):491-498.
55. Bis JC, Jian X, Kunkle BW, et al. Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. *Mol Psychiatry*. 2020;25(8):1859-1875.
56. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16(3):297-334.
57. Lee CS, Lee ML, Gibbons LE, et al. Associations between retinal artery/vein occlusions and risk of vascular dementia. *J Alzheimers Dis*. 2021;81(1):245-253.
58. Gross AL, Sherva R, Mukherjee S, et al. Calibrating longitudinal cognition in Alzheimer's disease across diverse test batteries and datasets. *Neuroepidemiology*. 2014;43(3-4):194-205.
59. Crane PK, Carle A, Gibbons LE, et al. Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Brain Imaging Behav*. 2012;6(4):502-516.
60. Gibbons LE, Carle AC, Mackin RS, et al. A composite score for executive functioning, validated in Alzheimer's Disease Neuroimaging Initiative (ADNI) participants with baseline mild cognitive impairment. *Brain Imaging Behav*. 2012;6(4):517-527.
61. Mukherjee S, Trittschuh E, Gibbons LE, Mackin RS, Saykin A, Crane PK. Dysexecutive and amnesic AD subtypes defined by single indicator and modern psychometric approaches: relationships with SNPs in ADNI. *Brain Imaging Behav*. 2012;6(4):649-660.
62. van der Sluis S, Verhage M, Posthuma D, Dolan CV. Phenotypic complexity, measurement bias, and poor phenotypic resolution contribute to the missing heritability problem in genetic association studies. *PLoS One*. 2010;5(11):e13929.
63. Reise SP, Morizot J, Hays RD. The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Qual Life Res*. 2007(16):19-31. Suppl 1.
64. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care*. 2007;45(5):S22-S31. Suppl 1.
65. Tucker-Drob EM, Salthouse TA. Confirmatory factor analysis and multidimensional scaling for construct validation of cognitive abilities. *Int J Behav Dev*. 2009;33(3):277-285.
66. Crane PK, Trittschuh E, Mukherjee S, et al. Incidence of cognitively defined late-onset Alzheimer's dementia subgroups from a prospective cohort study. *Alzheimers Dement*. 2017;13(12):1307-1316.
67. Mukherjee S, Mez J, Trittschuh EH, et al. Genetic data and cognitively defined late-onset Alzheimer's disease subgroups. *Mol Psychiatry*. 2020;25(11):2942-2951.
68. Bauman J, Gibbons LE, Moore M, et al. Associations between depression, traumatic brain injury, and cognitively-defined late-onset Alzheimer's disease subgroups. *Journal of Alzheimer's Disease*. 2019;70(2):611-619.
69. Ali M, Sung YJ, Wang F, et al. Leveraging large multi-center cohorts of Alzheimer disease endophenotypes to understand the role of Klotho heterozygosity on disease risk. *PLoS One*. 2022;17(5):e0267298.
70. Deming Y, Li Z, Kapoor M, et al. Genome-wide association study identifies four novel loci associated with Alzheimer's endophenotypes and disease modifiers. *Acta Neuropathol*. 2017;133(5):839-856.
71. Tatiana B, Didier C, David R, Derek Y. mixtools: an R package for analyzing finite mixture models. *Journal of Statistical Software*. 2009;32(6):1-29.
72. Kunkle BW, Schmidt M, Klein HU, et al. Novel Alzheimer disease risk loci and pathways in African American individuals using the African genome resources panel: a meta-analysis. *JAMA Neurol*. 2021;78(1):102-113.
73. Kunkle BW, Grenier-Boley B, Sims R, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. *Nat Genet*. 2019;51(3):414-430.
74. van Dam S, Folkertsma P, Castela Forte J, et al. The necessity of incorporating non-genetic risk factors into polygenic risk score models. *Sci Rep*. 2023;13(1):1351.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ray NR, Ayodele T, Jean-Francois M, et al. The Early-Onset Alzheimer's Disease Whole-Genome Sequencing Project: Study design and methodology. *Alzheimer's Dement*. 2023;1-9.
https://doi.org/10.1002/alz.13370