Title: Alzheimer's Disease Sequencing Project Release 4 Whole Genome Sequencing Dataset

Yuk Yee Leung^{1,2,*#}, Wan-Ping Lee^{1,2#}, Amanda B Kuzma^{1,2}, Heather Nicaretta^{1,2}, Otto Yuk Yee Leung^{+,2,} ", Wan-Ping Lee^{+,2}", Amanda B Kuzma^{+,2}, Heather Nicaretta^{+,2}, Otto
Valladares^{1,2}, Prabhakaran Gangadharan^{1,2}, Liming Qu^{1,2}, Yi Zhao^{1,2}, Youli Ren^{1,2}, Po-L
Cheng^{1,2}, Pavel P Kuksa^{1,2} Valladares^{4,2}, Prabhakaran Gangadharan^{4,2}, Liming Qu^{4,2}, Yi Zhao^{4,2}, Youli Ren^{4,2}, Po-Liang
Cheng^{1,2}, Pavel P Kuksa^{1,2}, Hui Wang^{1,2}, Heather White^{1,2}, Zivadin Katanic^{1,2}, Lauren Bass^{1,}
Naveen Saravana Cheng^{2,2}, Pavel P Kuksa^{2,2}, Hui Wang^{2,2}, Heather White^{2,2}, Zivadin Katanic^{2,2}, Lauren Bass^{2,2},
Naveen Saravanan^{1,2}, Emily Greenfest-Allen^{1,2}, Maureen Kirsch^{1,2}, Laura Cantwell^{1,2}, Taha
Iqbal^{1,3}, Nich Naveen Saravanan^{2,2}, Emily Greenfest-Allen^{2,2}, Maureen Kirsch^{2,2}, Laura Cantwell^{2,2}, Taha
Iqbal^{1,3}, Nicholas R Wheeler^{4,5}, John J. Farrell^{6,7}, Congcong Zhu^{6,7}, Shannon L Turner^{8,9}, Tan
Gunasekaran^{10,11} Iqbal^{4,5}, Nicholas R Wheeler^{4,5}, John J. Farrell^{9,7}, Congcong Zhu⁹⁷, Shannon L Turner^{9,5}, Tamil I
Gunasekaran^{10,11}, Pedro R Mena¹², Jimmy Jin^{1,2}, Luke Carter^{1,2}, Alzheimer's Disease Sequencir
Project, Xia Gunasekaran^{19,21}, Pedro R Mena²², Jimmy Jin^{2,2}, Luke Carter^{2,2}, Alzheimer's Disease Sequencing
Project, Xiaoling Zhang^{6,7}, Badri N Vardarajan^{10,11}, Arthur Toga¹³, Michael Cuccaro¹², Timothy J
Hohman^{8,9}, Project, Xiaoling Zhang^{o,}", Badri N Vardarajan^{10,11}, Arthur Toga¹⁵, Michael Cuccaro¹², Timothy J
Hohman^{8,9}, William S Bush^{4,5}, Adam C Naj^{1,3}, Eden Martin¹², Clifton Dalgard¹⁴, Brian W Kunkle¹²
Lindsay A Hohman^{9,9}, William S Bush^{3,9}, Adam C Naj^{3,9}, Eden Martin⁴⁴, Clitton Dalgard⁴⁴, Brian W Kunkle⁴⁴,
Lindsay A Farrer^{6,7}, Richard P Mayeux^{10,11}, Jonathan L Haines^{4,5}, Margaret A Pericak-Vance¹²,
Gerard D Sc

Lindsay A Farrer^{9,7}, Richard P Mayeux^{20,21}, Jonathan L Haines^{2,5}, Margaret A Pericak-Vance²²,
Gerard D Schellenberg¹, Li-San Wang^{1,2,*}
¹Department of Pathology and Laboratory Medicine, Perelman School of Medi Gerard D Schellenberg⁺, Li-San Wang^{+,4}
¹Department of Pathology and Laborator
of Pennsylvania
²Penn Neurodegeneration Genomics Cer
Medicine, Perelman School of Medicine, 1
2
3 1 Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University

Of Pennsylvania

Penn Neurodegeneration Genomics Center, Department of Pathology and Laboratory

Medicine, Perelman School of Medicine, University of Pennsylvania

Department of Biostatistics, Epidemiology, and Informatics 2

Penn Neurodeg
Medicine, Perelr
³Department of
Perelman Schoo Penn Medicine, Perelman School of Medicine, University of Pennsylvania
Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania
Perelman School of Medicine, Philadelphia, PA, USA
Department of Ĭ

³Department of Biostatistics, Epidemiology, and Informatics, Univers
Perelman School of Medicine, Philadelphia, PA, USA
⁴Department of Population and Quantitative Health Sciences, Case \
Cleveland, OH, USA E
 $\frac{1}{4}$
 $\frac{1}{4}$ Perelman School of Medicine, Philadelphia, PA, USA
Department of Population and Quantitative Health Sciences, Case Western Reserve Un
Cleveland, OH, USA
Department of Genetics and Genome Sciences, School of Medicine, Case ⁴Department of Population and Quantitative Health
Cleveland, OH, USA
⁵Department of Genetics and Genome Sciences, Scho
University, Cleveland, OH, USA 4
(
5
(1

Speeland, OH, USA

Speeland, OH, USA

Speeland of Genetics and Genome Sciences, School of Medicine, Case Western Reserve

University, Cleveland, OH, USA

Speeland of Medicine, Biostatistics & Bioinformatics, Boston Univers Cleveland, OH, USA

⁵ Department of Genetics and Genome Sciences,

University, Cleveland, OH, USA

⁶ Department of Medicine, Biostatistics & Bioinfo

Avedisian School of Medicine, Boston, MA, USA 5 ll e / 5 Department of Genetics and Genome Sciences, School of Medicine, Case Western Reserve

Department of Medicine, Biostatistics & Bioinformatics, Boston University Chobanian &
Department of Medicine, Biostatistics & Bioinformatics, Boston University Chobanian &
Nedisian School of Medicine, Boston, MA, USA
Depar ⁶ Department of Medicine, Biost
Avedisian School of Medicine, E
⁷ Department of Biostatistics, Bo $rac{1}{2}$ Nedisian School of Medicine, Boston, MA, USA
Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA
Department of Neurology, Vanderbilt University Medical Center, Nashville, TN, USA

 7 Department of Biostatistics, Boston University S
 8 Department of Neurology, Vanderbilt Universit
 9 Vanderbilt Genetics Institute, Vanderbilt Univer $\begin{array}{c} 7 \\ 5 \\ 1 \end{array}$ \overline{a}

 $\begin{array}{c} 2 \\ 3 \end{array}$ ⁸Department of Neurology, Vanderbilt University Medical Center, Nashville, TN, USA

Department of Neurology, Vanderbilt University Medical Center, Nashville, TN, USA
Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA
⁰Columbia University Irving Medical Center, New Yo $\overline{ }$

Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, U
⁰Columbia University Irving Medical Center, New York, NY, USA
¹Gertrude H. Sergievsky Center, Taub Institute for Research on the Agi $\begin{array}{cc} 1 & 1 \\ 1 & 1 \end{array}$ Columbia University Irving Medical Center, New York, NY, USA
¹Gertrude H. Sergievsky Center, Taub Institute for Research on the Aging Brain, Departr
of Neurology, Psychiatry, and Epidemiology, College of Physicians and S $\begin{array}{c} 1 \\ 1 \end{array}$ ¹⁰Columbia University Irving Medical Center, New York, NY, USA
¹¹Gertrude H. Sergievsky Center, Taub Institute for Research on t
of Neurology, Psychiatry, and Epidemiology, College of Physician
University, New York, NY $\begin{array}{c}\n1 \\
1 \\
1\n\end{array}$ ⁺⁺Gertrude H. Sergievsky Center, Taub Institute for Research on the Aging Brain, Departments
of Neurology, Psychiatry, and Epidemiology, College of Physicians and Surgeons, Columbia
University, New York, NY, USA
NOTE: Th University, New York, NY, USA
NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

It is made available under a CC-BY-ND 4.0 International license. medRxiv preprint doi: [https://doi.org/10.1101/2024.12.03.24317000;](https://doi.org/10.1101/2024.12.03.24317000) this version posted December 6, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted

1
|
1
| ¹¹ Department of Human Genetics and John P. Hussman Institute for Human Genomics,
University of Miami Miller School of Medicine, Miami, FL, USA
¹³ Laboratory of Neuro Imaging, USC Stevens Neuroimaging and Informatics I ¹³Laboratory of Neuro Imaging, USC Stevens Neuroimaging and
School of Medicine of USC, University of Southern California
¹⁴Department of Anatomy, Physiology and Genetics, School of
University of the Health Sciences, Be

¹⁻'Laboratory of Neuro Imaging, USC Stevens Neuroimaging and Informatics Institute, Keck
School of Medicine of USC, University of Southern California
¹⁴Department of Anatomy, Physiology and Genetics, School of Medicine Medicine of Anatomy, Physiology and Genetics, School of University of the Health Sciences, Bethesda, MD, USA
* Corresponding authors
* Authors contribute equally **Department of Anatomy, Physiology and Genetics, School of Medicine, Uniformed Services
University of the Health Sciences, Bethesda, MD, USA
* Corresponding authors
Authors contribute equally
Corresponding authors:

University of the Health Sciences, Bethesda, MD, 2014
* Corresponding authors
Corresponding authors:
M. W. USA *
#

C

Authors contribute equeries
Corresponding authors:
Yuk Yee Leung, Ph.D.
Department of Pathology Authors contribute equally
Corresponding authors:
Yuk Yee Leung, Ph.D.
Department of Pathology and
University of Pennsylvania Pe くししにここ Yuk Yee Leung, Ph.D.
Department of Patholog
University of Pennsylvan
3700 Hamilton Walk, Ric
Philadelphia PA 19104 ていこ トコ Department of Pathol
University of Pennsyl
3700 Hamilton Walk,
Philadelphia PA 1910
Tel: (215) 573-372
Email: <u>yyee@pennme</u> University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
3700 Hamilton Walk, Richards Medical Research Building
Philadelphia PA 19104
Tel: (215) 573-3729
Email: <u>yyee@pennmedicine.upenn.edu</u>
W.S. W. S. 1700 Hamilton Walk, Richards Medical Research Building
Philadelphia PA 19104
Tel: (215) 573-3729
Email: <u>yyee@pennmedicine.upenn.edu</u>
Li-San Wang, Ph.D. Philadelphia PA 19104

Tel: (215) 573-3729

Email: <u>yyee@pennmedicine.upenn.edu</u>

Li-San Wang, Ph.D.

Department of Pathology and Laboratory Medicine, Penn

Tel: (215) 573-3729
Email: <u>yyee@pennmed</u>
Li-San Wang, Ph.D.
Department of Patholo
University of Pennsylva Email: yyee@pennmed
Li-San Wang, Ph.D.
Department of Patholo,
University of Pennsylva
3700 Hamilton Walk, Ri Li-San Wang, Ph.D.
Li-San Wang, Ph.D.
Department of Pathology and Laborator
University of Pennsylvania Perelman Sch
3700 Hamilton Walk, Richards Medical F
Philadelphia PA 19104 せいしこ トコー Department of Path
Department of Path
Jniversity of Penns
3700 Hamilton Wal
Philadelphia PA 191
Tel: (215) 746-70
Email: <u>Iswang@pen</u> Department of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
3700 Hamilton Walk, Richards Medical Research Building
Philadelphia PA 19104
Tel: (215) 746-7015
Email: <u>Iswang@pennmedicine.upenn.edu</u> University of Pennsylvania Perennia Perennia Perelman School of Pennsylvania Perelman School of Philadelphia, P
Philadelphia, PA, 19104
Tel: (215) 746-7015
Email: <u>Iswang@pennmedicine.upenn.edu</u> 2012 Hamilton Walk, Antonio Medical Research
2700 Philadelphia PA 19104
2700 Email: <u>Iswang@pennmedicine.upenn.edu</u> Tel: (215) 746-7015
Email: <u>Iswang@pennm</u>
Email: <u>Iswang@pennm</u> Tel: (215) 746-7015 Email: lswang@pennmedicine.upenn.edu

Abstract

The Alzheimer's Disease Sequencing Project (ADSP) is a national initiative to understand the genetic architecture of Alzheimer's Disease and Related Dementias (AD/ADRD) by sequencing
whole genomes of affected participants and age-matched cognitive controls from diverse
populations. The Genome Center for Alzheimer' whole genomes of affected participants and age-matched cognitive controls from diverse
populations. The Genome Center for Alzheimer's Disease (GCAD) processed whole-genome
sequencing data from 36,361 ADSP participants, inc more genomes of a fareomerator and age-matched cognitions. The Genome Center for Alzheimer's Disease (GCAD) processed whole-genom
sequencing data from 36,361 ADSP participants, including 35,014 genetically unique
participa populations. The Genome Center form 36,361 ADSP participants, including 35,014 genetically unique
participants of which 45% are from non-European ancestry, across 17 cohorts in 14 countrie
this fourth release (R4). This se participants of which 45% are from non-European ancestry, across 17 cohorts in 14 c
this fourth release (R4). This sequencing effort identified 387 million bi-allelic variant
million short insertions/deletions, and 2.2 mil participants release (R4). This sequencing effort identified 387 million bi-allelic variants, 42
million short insertions/deletions, and 2.2 million structural variants. Annotations and quality
control data are available f million short insertions/deletions, and 2.2 million structural variants. Annotations and qua
control data are available for all variants and samples. Additionally, detailed phenotypes fi
15,927 participants across 10 domai control data are available for all variants and samples. Additionally, detailed phenotypes from
15.927 participants across 10 domains are also provided. A linkage disequilibrium panel was 15,927 participants across 10 domains are also provided. A linkage disequilibrium panel was
created using unrelated AD cases and controls. Researchers can access and analyze the genetic
data via NIAGADS Data Sharing Servic 1994) participants across the acrossive of provided in the quart plant can be created using unrelated AD cases and controls. Researchers can access and analyze the gene
data via NIAGADS Data Sharing Service, the VariXam to

l, Introduction

data via NIAGADS Data Sharing Service, the VariXam tool, or NIAGADS GenomicsDB.
 Introduction

Alzheimer's disease (AD) is a neurodegenerative condition characterized by the abnorm

buildup of amyloid-ß peptides in extra Alzheimer's disease (AD) is a neurodegenerative condition characterized by the abnormal
buildup of amyloid-β peptides in extracellular plaques and hyperphosphorylated tau in
intracellular neurofibrillary tangles. This pro / k i c c t . decline in cognitive and functional abilities. Genetic variants play a significant role in the
development of late-onset AD (LOAD). The first notable finding in 1993 reported the ε 4 allele of
the apolipoprotein E (APO intracellular neurofibrillary tangles. This progressive neurodegeneration results in a gra
decline in cognitive and functional abilities. Genetic variants play a significant role in the
development of late-onset AD (LOAD) decline in cognitive and functional abilities. Genetic variants play a significant role in the
development of late-onset AD (LOAD). The first notable finding in 1993 reported the ε 4 allel
the apolipoprotein E (APOE) ge development of late-onset AD (LOAD). The first notable finding in 1993 reported the ϵA a
the apolipoprotein E (APOE) gene was associated with the risk of developing AD [1]. The
identification of additional genetic fact development of late onset AD (LOAD). The first notable finding in 1993 reported the ε4 allele of
the apolipoprotein E (APOE) gene was associated with the risk of developing AD [1]. The
identification of additional genetic the aponpoprotein E (APOE) gene was associated with the risk of developing AD [1]. The
identification of additional genetic factors for LOAD accelerated with the advent of high-
throughput genomic technologies, such as gen effects on disease susceptibility as the studies focused on genotype array data. These variants
contribute minimally to the overall genetic liability for the disease, as a study showed that the the list of genetic factors continues to expand with the inclusion of bigger samples sizes through
international consortium efforts, notably between the Alzheimer's Disease Genetics
Consortium (ADGC) and International Geno international consortium efforts, notably between the Alzheimer's Disease Genetics
Consortium (ADGC) and International Genomics of Alzheimer's Project (IGAP) [4-6]. However,
most of these identified genetic variants are co most of these identified genetic variants are common alleles with individually small causal
effects on disease susceptibility as the studies focused on genotype array data. These variants
contribute minimally to the overal effects on disease susceptibility as the studies focused on genotype array data. These varia
contribute minimally to the overall genetic liability for the disease, as a study showed that
single nucleotide polymorphism (SNP contribute minimally to the overall genetic liability for the disease, as a study showed that the
single nucleotide polymorphism (SNP)-heritability estimate from the largest AD genome-wide
association studies (GWAS) to dat single nucleotide polymorphism (SNP)-heritability estimate from the largest AD genome-wide
association studies (GWAS) to date is 3.1% [7], which is significantly smaller than the heritabilit
estimates for AD obtained from

association studies (GWAS) to date is 3.1% [7], which is significantly smaller than the heritabili
estimates for AD obtained from twin studies, which ranges from 60 to 80% [8, 9].
Whole-genome sequencing (WGS) can address estimates for AD obtained from twin studies, which ranges from 60 to 80% [8, 9].
Whole-genome sequencing (WGS) can address this missing heritability challenge by 1) providing
a more comprehensive view of the genetic archit Whole-genome sequencing (WGS) can address this missing heritability challenge k
a more comprehensive view of the genetic architecture via a full-spectrum of varia
identifying rare variants with potentially larger phenotypi \
1 c
F More genome sequencing (WSS) can address intermediate intermediately manifology and a process and a process the above gaps. WGS studies are more costly compared to genotyping array, and as a result fewer participants of no identifying rare variants with potentially larger phenotypic effects. Analyzing many samples is
necessary to address the above gaps. WGS studies are more costly compared to genotyping
array, and as a result fewer participa Francessary to address the above gaps. WGS studies are more costly compared to genotyping
array, and as a result fewer participants of non-European descent have been sequenced [10].
Expanding AD research to ancestrally div necessary of addets the above gaps. The contribution in the cost, comparison gains γ_{p} are alread in a paraly diverse populations is crucial for several reasons.
Expanding AD research to ancestrally diverse populati Expanding AD research to ancestrally diverse populations is crucial for several reasons.
Most genetic studies of Alzheimer's disease (AD) have focused on non-Hispanic White (NHW)
populations. However, genetic risk factors Host genetic studies of Alzheimer's disease (AD) have focused on non-Hispanic White (
populations. However, genetic risk factors identified in NHW populations may not fully
populations. populations. However, genetic risk factors identified in NHW populations may not fully explain populations. However, genetic risk factors identified in NHW populations may not fully explain

late-onset AD in NHW individuals, its predictive power is weaker and more inconsistent in
African American (AA) and Hispanic or Latino (HL) populations [11-13]. Investigating how
genetic risk factors for AD vary among ethn African American (AA) and Hispanic or Latino (HL) populations [11-13]. Investigating how
genetic risk factors for AD vary among ethnic groups could pave the way for more effective
tailored treatments and interventions. Not African American (AA) and American Coupty propenduate (LEV) produces and genetic risk factors for AD vary among ethnic groups could pave the way for more effective tailored treatments and interventions. Notably, certain ge tailored treatments and interventions. Notably, certain genetic variants—such as those in *SORL1*, *ABCA7*, and *ACE*—exhibit stronger associations with AD risk in specific groups, included Asians [14], AA [15], and Israel SORL1, ABCA7, and ACE—exhibit stronger associations with AD risk in specific groups, including
Asians [14], AA [15], and Israeli-Arabs [16]. These findings suggest that gene therapies targeting
these genes may have varying

Asians [14], AA [15], and Israeli-Arabs [16]. These findings suggest that gene therapies targeting
these genes may have varying levels of effectiveness across different ethnicities.
Funded through cooperative agreements an These genes may have varying levels of effectiveness across different ethnicities.

Funded through cooperative agreements and research grants, the Alzheimer's Disease

Sequencing Project (ADSP) brings together 497 investig Funded through cooperative agreements and research grants, the Alzheimer's Di
Sequencing Project (ADSP) brings together 497 investigators from institutions wo
2023, it successfully completed the "Follow-Up Study (FUS) Phas Funder through Project (ADSP) brings together 497 investigators from institutions worldwider and the Uniformed Services of AA and pan-Hispanic ancestry at The American Genomerat the Uniformed Services University of the Hea 2023, it successfully completed the "Follow-Up Study (FUS) Phase" (the third phase of ADSP)
sequencing existing cohorts of AA and pan-Hispanic ancestry at The American Genome Cente
at the Uniformed Services University of t 2023, it successfully completed the Main Presence of AA and pan-Hispanic ancestry at The American Genome Center
at the Uniformed Services University of the Health Sciences (USUHS) and John P. Hussman
Institute for Human Ge sequencing existing critics University of the Health Sciences (USUHS) and John P. Hussman
Institute for Human Genomics (HIHG). This effort was conducted in collaboration with
established NIH-funded AD infrastructure like t at the Unity of Human Genomics (HIHG). This effort was conducted in collaboration with
established NIH-funded AD infrastructure like the National Cell Repository for Alzheimer's
Disease (NCRAD), National Institute on Aging established NIH-funded AD infrastructure like the National Cell Repository for Alzheime
Disease (NCRAD), National Institute on Aging Genetics of Alzheimer's Disease Data Sto
(NIAGADS), and the Genome Center for Alzheimer's Disease (NCRAD), National Institute on Aging Genetics of Alzheimer's Disease Data Storage
(NIAGADS), and the Genome Center for Alzheimer's Disease (GCAD). Additionally, participa
from NHW and Asian ancestries have been seq (NIAGADS), and the Genome Center for Alzheimer's Disease (GCAD). Additionally, participants
from NHW and Asian ancestries have been sequenced. All genomes, including those from
previous ADSP phases and other collaborative (NIAG) we are sequenced. All genomes, including those from
previous ADSP phases and other collaborative projects, have been processed using a unified
pipeline, subjected to comprehensive quality control, and annotated usin from NHW and Asian ancestries have been sequenced. All genomes, including those from
previous ADSP phases and other collaborative projects, have been processed using a unified
pipeline, subjected to comprehensive quality c pipeline, subjected to comprehensive quality control, and annotated using various resources. phis collection, the Release 4 (R4) of ADSP data, forms the world's largest publicly available AD
genome resource.
Results
Sequence existing ancestrally diverse cohorts via the ADSP Follow-Up Study (ADSP-FUS) This collection, the Release 4 (R4) of ADSP (RM4) collection, forms the young publicly available Le
genome resource.
Sequence existing ancestrally diverse cohorts via the ADSP Follow-Up Study (ADSP-FUS)

Results

֦ Sequence existing ancestrally diverse cohorts via the ADSP Follow-Up Study (ADSP-FUS)

o
Results
<u>Sequence existing</u>
The ADSP-FUS is a $\begin{array}{c} \underline{\mathsf{S}} \\ \underline{\mathsf{S}} \end{array}$ The ADSP-FUS is a National Institute on Aging (NIA) initiative focused on identifying genetic risk The ADSP cohorts beyond primalism progression and protective variants for AD by expanding the ADSP cohorts beyond primarily participants
with NHW. Given limitations in population diversity in the ADSP, the ADSP-FUS was des and preserve entity terms in perpending and the ADSP, the ADSP-FUS was designed
sequence existing ancestrally diverse and unique cohorts. ADSP-FUS 2.0 (The Diverse
Population Initiative) focuses on HL, non-Hispanic Black w sequence existing ancestrally diverse and unique cohorts. ADSP-FUS 2.0 (The Diverse
Population Initiative) focuses on HL, non-Hispanic Black with African Ancestry, and Asian
populations (e.g., the Asian cohort for Alzheime Population Initiative) focuses on HL, non-Hispanic Black with African Ancestry, and As
populations (e.g., the Asian cohort for Alzheimer's disease). ADSP-FUS intends to sequence
100,000 participants from diverse ancestries Populations (e.g., the Asian cohort for Alzheimer's disease). ADSP-FUS intends to sequen
100,000 participants from diverse ancestries (Methods - Sequence existing ancestrally d
cohorts via the ADSP Follow-Up Study (FUS)). populations (e.g., 100,000 participants from diverse ancestries (Methods - Sequence existing ancestrally diverse
cohorts via the ADSP Follow-Up Study (FUS)). ADSP has developed a workflow (Supplementary
Figure 1) to suppor 100,000 participants from diverse ancestries (Methods - Sequence existing ancestrally diverse

cohorts via the ADSP Follow-Up Study (FUS)). ADSP has developed a workflow (Supplementary

Figure 1) to support biospecimens pr Figure 1) to support biospecimens processing, DNA preparation, and sequencing at USUHS and
HIHG. Figure 1) to support biospecimens processing, DNA preparation, and sequencing at OSONS and
HIHG.

Sample characteristics

20,771 WGS data (8,159 new) from the ADSP-FUS phases are included in the ADSP Release 4 (R4) dataset, bringing the total number to 36,361 across 17 cohorts/studies in 14 countries
(Figure 1A). Sequencing was carried out at ten sequencing centers using Illumina technology.
Most were generated using the PCR-fr (Figure 1A). Sequencing was carried out at ten sequencing centers using Illumina technolog
Most were generated using the PCR-free protocol (91%) and 150bp in read length (94%). The
sequencing platforms were used: Illumina (Figure 1A). Sequencing was carried out at ten sequencing centers using murning technology.

Most were generated using the PCR-free protocol (91%) and 150bp in read length (94%). Thre

sequencing platforms were used: Illum Most were generated using the Protocol (720) and 150 kpc were useded using the provides more details on sequencing configurations. GCAD processed all 36,361 WGS samples from read mapping to variant calling using a standard sequencing procession were used in an interesting and the processed all 36,361 WGS samples from read mapping to variant calling using a standard pipeline (VCPA1.1) to harmonize all data and minimize batch effects [17] (Met machines. Supplementary Table S1 provides more details on sequencing comigurations. GCAD
processed all 36,361 WGS samples from read mapping to variant calling using a standard
pipeline (VCPA1.1) to harmonize all data and m

pipeline (VCPA1.1) to harmonize all data and minimize batch effects [17] (Methods - Data
and Sample processing protocol on SNVs and indels).
Of these 36,361 samples, 35,014 participants are genetically unique. Based on eth pipeline (VCFA1.1) to harmonize all data and minimize batch effects [17] (Methods - Dataflow
and Sample processing protocol on SNVs and indels).
Of these 36,361 samples, 35,014 participants are genetically unique. Based on and Sample processing protocol on SNVS and muers).
Of these 36,361 samples, 35,014 participants are genereported by cohorts, there are 5,260 AA, 78 American
Asian, 10,972 Hispanics, 9 Native Hawaiian or Pacific Is
particip Cr/Fi reported by cohorts, there are 5,260 AA, 78 American Indian or Alaskan Native (AlorAN)
Asian, 10,972 Hispanics, 9 Native Hawaiian or Pacific Islander (NHoPI), 13,969 NHW
participants, and 1,896 participants were of unknown reported by controlly interestingly and by controllation manufal controlly, 13,969 NHW
Asian, 10,972 Hispanics, 9 Native Hawaiian or Pacific Islander (NHoPI), 13,969 NHW
participants, and 1,896 participants were of unknown

participants, and 1,896 participants were of unknown ethnicity. A breakdown of der
information is summarized in Figure 1A and Supplementary Table S2.
We inferred the genetic ancestry of each individual using called genetic particles information is summarized in Figure 1A and Supplementary Table S2.
We inferred the genetic ancestry of each individual using called genetic data to investigate the
discordance between the reported and genetically Information is summarized in Figure 1A and Supplementary Table 32.
We inferred the genetic ancestry of each individual using called geneti
discordance between the reported and genetically inferred ancestries.
reported were \
cr ∈ ∈ (Mexicance between the reported and genetically inferred ancestries. Most discordance
reported were in admixed participants [18], presenting additional challenges in identifying
ancestry-specific variants. We used GRAF-pop reported were in admixed participants [18], presenting additional challenges in identifying
ancestry-specific variants. We used GRAF-pop [19] (Methods - GRAF-POP), which assume
each individual is an admixture of three ance ancestry-specific variants. We used GRAF-pop [19] (Methods - GRAF-POP), which assumes
each individual is an admixture of three ancestral groups: European (e), African (f), and Asia
(a). GRAF-pop estimates ancestry componen each individual is an admixture of three ancestral groups: European (e), African (f), and Asian (a). GRAF-pop estimates ancestry components Pe , Pf , and Pa , which are then used to assign participants to population grou (a). GRAF-pop estimates ancestry components *Pe, Pf,* and *Pa*, which are then used to assign participants to population groups, including European, African/African American, Latin American, Asian-Pacific Islander, and Ea American, Asian-Pacific Islander, and East/South Asian. Using the software's default settings, participants to population groups, including European, American, Asian-Pacific Islander, and East/South Asian. Using the software's default set
the match rates between reported and genetically inferred ancestries were 99.4 American, Asian-Pacific Islander, and 200, 2000 and 200, 2000 Performance incrementation,
The match rates between reported and genetically inferred ancestries were 99.4% for
European, 98.3% for African American, 96.8% for

European, 98.3% for African American, 96.8% for Asian, and 80.0% for Hispanic partici
(Figures 1B and 1C).
We also performed principal component analysis (PCA) based on genotypes derived fr
(Methods – Population substructu European, 98.3% for African American, 98.3% for African, and 80.00% for American American American American American
We also performed principal component analysis (PCA) based on genotypes derived from W
(Methods – Populat (Figures 1B and 1C).
We also performed p
(Methods – Population)
quality and performed
calculated principal control the ances (Methods – Population substructure). We selected common variants (MAF>0.02) of high
quality and performed linkage disequilibrium (LD) pruning to yield 146,964 variants, then
calculated principal components (PCs) and geneti quality and performed linkage disequilibrium (LD) pruning to yield 146,964 variants, then calculated principal components (PCs) and genetic relationship matrix (GRM). We then performed the ancestry inference analysis using performed the ancestry inference analysis using 145,278 variants common in both the ADSP R4
and gnomAD data [20], a publicly available population genetics resource generated on 76,215
diverse samples. Subjects were assigne performed the ancestry inference analysis using 145,278 variants common in both the *i*
and gnomAD data [20], a publicly available population genetics resource generated on
diverse samples. Subjects were assigned to an anc performand and gnomable and gnomable and genetics and generated on 76,215
diverse samples. Subjects were assigned to an ancestry group in which it has the highest
ancestry proportion value. The match rates between reported and generate and particle population generation generation generation is generated on diverse samples. Subjects were assigned to an ancestry group in which it has the highest ancestry proportion value. The match rates betw ancestry proportion value. The match rates between reported and genetically inferred
ancestries by this method were 88.3% for Non-Finnish European and 8.7% for Ashkenazi
98.7% for AA, 99.7% for Asian, and among Hispanic pa ancestries by this method were 88.3% for Non-Finnish European and 8.7% for Ashkena:
98.7% for AA, 99.7% for Asian, and among Hispanic participants, 45.5% for Latino, 15%
15% 98.7% for AA, 99.7% for Asian, and among Hispanic participants, 45.5% for Latino, 15% for AA
Repeated were and 8.7% for Asian, and 8.7% for Ashkenan and 8.7% for Latino, 15% for AA
Ashkenan and 8.7% for Asian and 8.7% for 98.7% for AA, 99.7% for Asian, and among Hispanic participants, 45.5% for Latino, 15% for AA

and 37.8% for remaining. The remaining subjects are most likely from Caribbean region (Figures
1B and 1D).
WGS sample quality assessment
We performed QC checks on all samples and 734 samples with quality issues (Method – A

was sample quality assessment

1B and 1B).
WGS sample
We perform
Sample leve
depth across |
|
|
| We performed QC checks on all samples and 734 samples with quality issues (Method - ADSP Sample level QC protocol) and reassessed the quality of the callset (N=36,361). The mean readepth across samples is 40.4x with 99% of samples having a coverage >30x (Figure 2A). The performed percentage of bases with the q Sample level QC protocol) and reassessed the quality of the callset (N=36,361). The mean read
depth across samples is 40.4x with 99% of samples having a coverage >30x (Figure 2A). The per
genome percentage of bases with th depth across samples is 40.4x with 99% of samples having a coverage 230x (Figure 2A). The per-
genome percentage of bases with the quality score greater than Q30 (sequencing error rate
less than 0.1%) is 90.18±2.43%. On av

less than 0.1%) is 90.18±2.43%. On average, 98.92±2.13% reads of samples are mapped, and
94.14±2.76% of paired-end reads have both ends mapped.
On average, each sample contained 4.3 million single nucleotide variants (SNV 1.14±2.76% of paired-end reads have both ends mapped.
On average, each sample contained 4.3 million single nucleotide variants (SNVs) and 999,000
short insertions and deletions (indels). AA samples have the highest number 94.14 On average, each sample contained 4.3 million single nucles
Short insertions and deletions (indels). AA samples have the
million) followed by Hispanic samples (4.3 million), Asian sa
samples (4.1 million) (Figures 2B Srs
N short insertions and deletions (indels). AA samples have the highest number of variants (4.9 million) followed by Hispanic samples (4.3 million), Asian samples (4.2 million), and NHW samples (4.1 million) (Figures 2B, Supp

million) followed by Hispanic samples (4.3 million), Asian samples (4.2 million), and NHW
samples (4.1 million) (**Figures 2B, Supplementary Figure 2a**).
We assessed if the number of called variants were affected by sequenc samples (4.1 million) (Figures 2B, Supplementary Figure 2a).
We assessed if the number of called variants were affected by sequencing configurations
as sequencing platform and use of PCR. We found that samples (regardless samples (4.1 million) (Figures 2B, Supplementary Figure 2a).
We assessed if the number of called variants were affected by
as sequencing platform and use of PCR. We found that sampl
sequenced by the PCRFree protocol tend t / U U U U U U U U as sequencing platform and use of PCR. We found that samples (regardless of ethnicities) as sequenced by the PCRFree protocol tend to yield more variants, with the combination of
NovaSeq + PCRFree returning the highest (Supplementary Figure 2b,2c).
Bi-allelic variants (SNVs and indels) in ADSP R4
We called gen

Bi-allelic variants (SNVs and indels) in ADSP R4

NovaSeq + PCRFree returning the highest (Supplementary Figure 2b,2c).
Bi-allelic variants (SNVs and indels) in ADSP R4
We called genotypes for all observed variants across all 36,361 samples and split the joint
called resu Novaseq + PCRFree returning the highest (Supplementary Figure 2b,2c).
Bi-allelic variants (SNVs and indels) in ADSP R4
We called genotypes for all observed variants across all 36,361 samples a
called results into two VCF f ■ \ く v v \overline{a} Rate>=80%, and is supported by <500 reads (Method - ADSP Variant QC protocol). Averag short indels. We identified 322,757,476 bi-allelic SNVs and 24,005,724 bi-allelic indels on
autosomes, comprising 83% of the original GATK output, which will be the focus of this article
R4 data also contains 54,425,255 mu short independent independent and state of the original GATK output, which will be the focus of this are R4 data also contains 54,425,255 multi-allelic SNVs and indels on autosomes. A variant particle in a QC if it has a G R4 data also contains 54,425,255 multi-allelic SNVs and indels on autosomes. A variant passes
QC if it has a GATK "FILTER" = PASS or is in tranche >= 99.8%, DP>10 and GQ>20, call
Rate>=80%, and is supported by <500 reads QC if it has a GATK "FILTER" = PASS or is in tranche >= 99.8%, DP>10 and GQ>20, call
Rate>=80%, and is supported by <500 reads (Method - ADSP Variant QC protocol). Average ca
rate of the variants is high (97.0%). Details f Rate>=80%, and is supported by <500 reads (Method - ADSP Variant QC protocol). A
rate of the variants is high (97.0%). Details for other quality of the variants can be for
Supplementary Table S3.
The ADSP quality control (

rate>=80%, and is supported by <500 reads (Method - ADSP Variant QC protocol). Average can
rate of the variants is high (97.0%). Details for other quality of the variants can be found in
Supplementary Table S3.
The ADSP qu rate of the variants is high (97.0%). Details for other quality of the variants can be found in
 Supplementary Table S3.

The ADSP quality control (QC) protocol flagged 92.94% of autosomal variants, 299,620,924

SNVs and Supplementary Table 33.

The ADSP quality control (

SNVs and 22,674,845 inde

(N=2,787), Hispanic (N=10

and 6,855,642 indels) for *1*

132,424,746 (123,631,863

(136,112,082,5NVs and 0.1 SNVs and 22,674,845 indels, as high quality. Of the four major ethnicities, AA (N=5,260), As (N=2,787), Hispanic (N=10,972), and NHW (N=13,969), there are 101,227,106 (94,371,761 S
and 6,855,642 indels) for AA, 69,338,361 (N=2,787), Hispanic (N=10,972), and NHW (N=13,969), there are 101,227,106 (94,371,761 SNV
and 6,855,642 indels) for AA, 69,338,361 (64,882,876 SNVs and 4,455,485 indels) for Asian,
132,424,746 (123,631,863 SNVs and 8,792,8 and 6,855,642 indels) for AA, 69,338,361 (64,882,876 SNVs and 4,455,485 indels) for Asian,
132,424,746 (123,631,863 SNVs and 8,792,883 indels) for Hispanic, and 135,672,855
(126,113,083 SNVs and 9,559,772 indels) for NHW (132,424,746 (123,631,863 SNVs and 8,792,883 indels) for Hispanic, and 135,672,855 $(126, 113, 083 \text{ SNVs and } 9,559, 772 \text{ indels}) \text{ for NHW (Table 1, top) respectively.}$ $(120,113,083$ SNVs and 9,335,772 indels) for NHW (Table 1, top) respectively.

Regarding allele frequency (AF), 52.53% of variants are singletons, followed by 40.62% rare variants with AF < 0.1%, 3.72% with AF 0.1-1%, 1.29% with AF 1-5%, and 1.84% with AF > 5%. The distribution of variants across AF The distribution of variants across AF ranges is consistent across ethnicities, with singletons
comprising close to or more than 50% of the variants. AA, Asian, and Hispanic groups (13.16%
11.61%, and 7.53%, respectively) comprising close to or more than 50% of the variants. AA, Asian, and Hispanic groups (13.16
11.61%, and 7.53%, respectively) have a higher proportion of variants with AF > 1% compare
to the overall dataset and the NHW grou

11.61%, and 7.53%, respectively) have a higher proportion of variants with AF > 1% compared
to the overall dataset and the NHW group (5.74% and 3.13%, respectively) (Table 1, top).
In terms of ethnic specific variants, th 11.61%, and 7.53%, respectively) have a higher proportion of variants with AF > 1% compared In terms of ethnic specific variants, there are 43,607,568 (40,417,070 SNVs and 3,190,498 indels), 41,397,885 (38,544,025 SNVs and 2,853,860 indels), 65,941,432 (61,362,910 SNVs 4,578,522 indels), 80,641,315 (7,4475,786 S Indels), 41,397,885 (38,544,025 SNVs and 2,853,860 indels), 65,941,432 (61,362,910 SNVs
4,578,522 indels), 80,641,315 (7,4475,786 SNVs and 6,165,529 indels) variants for AA, Asia
Hispanic, and NHW. Most of the ancestral sp 4,578,522 indels), 80,641,315 (7,4475,786 SNVs and 6,165,529 indels) variants for AA, Asian,
Hispanic, and NHW. Most of the ancestral specific variants are rare variants (AF < 0.1%,
Supplementary Table S4) (**Table 1, bot**

Hispanic, and NHW. Most of the ancestral specific variants are rare variants (AF < 0.1%,
Supplementary Table S4) (Table 1, bottom).
We compared the ADSP R4 bi-allelic variants to gnomAD [21] (Method — Comparison of
gnomAD Supplementary Table S4) (Table 1, bottom).
We compared the ADSP R4 bi-allelic variants to gnomAD [21] (Method – Comparison of gnomAD). Of the ADSP R4 variants, 62.39% of SNVs and 57.33% of indels are reported in gnomAD. I Supplementary Table S4) (Table 1, bottom).
We compared the ADSP R4 bi-allelic variants
gnomAD. In terms of allele frequencies, 99.89
5%, 99.74% with 0.1% < AF ≤ 1%, 82.76% with
in the gnomAD database. gnomAD). Of the ADSP R4 variants, 62.39% of SNVs and 57.33% of indels are reported in gnomAD. In terms of allele frequencies, 99.89% of variants with AF > 5%, 99.80% with 19
5%, 99.74% with 0.1% < AF ≤ 1%, 82.76% with AF gnomAD. Of the ADSP R4 variants, 62.39% of SNVs and 37.39% of mdels are reported in
gnomAD. In terms of allele frequencies, 99.89% of variants with AF > 5%, 99.80% with 1%
5%, 99.74% with 0.1% < AF ≤ 1%, 82.76% with AF ≤ 5%, 99.74% with 0.1% < AF ≤ 1%, 82.76% with AF ≤ 0.1%, and 41.08% of singletons are present in the gnomAD database.

Annotation of genetic variants

and indels) (Methods - Variant annotation protocol, Figure 3A). Functional impact of variants was accessed using snpEff [23] (Methods - LOF analyses). We identified 224,594 high-impact loss-of-function variants: frameshift (39%), stop gained (27%), splice donor (16%), splice $\begin{array}{c} \mathbf{1} \\ \mathbf{1} \\ \mathbf{2} \\ \mathbf{3} \\ \mathbf{4} \end{array}$ The official ADSP annotation pipeline [22] was used to annotate all 347 million variants (SNVs and indels) (Methods – Variant annotation protocol, Figure 3A). Functional impact of variants
was accessed using snpEff [23] (Methods – LOF analyses). We identified 224,594 high-impact
loss-of-function variants: frameshift and indels) (Methods – Variant annotation protocol, Figure 3A). Functional impact of variants
was accessed using snpEff [23] (Methods – LOF analyses). We identified 224,594 high-impact
loss-of-function variants: frameshift was accessed using sippen [23] (Methods – LOT analyses). We identified 224,394 high-impact
loss-of-function variants: frameshift (39%), stop gained (27%), splice donor (16%), splice
acceptor (12%), start lost (4%), and sto acceptor (12%), start lost (4%), and stop lost (2%) across 22,710 genes (**Figure 3B**). Amon
these, 1,295 genes were found to be intolerant to protein-truncating loss-of-function var
indicated by a Loss Intolerance Probabil acceptor (12%), start lost (4%), and stop lost (2%) across 22,710 genes (Figure 3B). Among
these, 1,295 genes were found to be intolerant to protein-truncating loss-of-function varia
indicated by a Loss Intolerance Probabi indicated by a Loss Intolerance Probability (pLI) score of 1, suggesting the critical importance of
these genes. We also provide annotation for all bi-allelic variants using FAVOR (Methods –
FAVOR annotation protocol). 27. indicate genes. We also provide annotation for all bi-allelic variants using FAVOR (Methods –
FAVOR annotation protocol). 27.45 million variants of such are with CADD (phred score) of 20
or above, with over 63.64 million v these genes. We also provide annotation for all bi-allelic variants using FAVOR (Methods
FAVOR annotation protocol). 27.45 million variants of such are with CADD (phred score) of
or above, with over 63.64 million variants

Structural variants in ADSP R4 samples

FAVOR annotation protocol). 27.45 million variants of such are with CADD (pined score) of 20
or above, with over 63.64 million variants lying in super-enhancer regions.
Structural variants in ADSP R4 samples
We applied the We applied the same protocol developed for the ADSP R3 dataset [24] to the R4 dataset
(Methods – Structural variant calling protocol). Individual Manta and Smoove callsets were
initially merged for each sample, and then al Structural variants in ADSP N4 samples
We applied the same protocol develope
(Methods – Structural variant calling pr
initially merged for each sample, and the
GraphTyper (v2.7) was subsequently app
joint genotyping. Notab (Methods – Structural variant calling protocol). Individual Manta and Smoove callsets winitially merged for each sample, and then all samples were combined using SVIMMER (v GraphTyper (v2.7) was subsequently applied to th (Methods – Structural variant calling protocol). Individual wanta and Shloove callsets were
initially merged for each sample, and then all samples were combined using SVIMMER (v0.1)
GraphTyper (v2.7) was subsequently appli GraphTyper (v2.7) was subsequently applied to the merged VCF for structural variant (SV)
joint genotyping. Notably, only SVs larger than 10 Mbp were filtered. The final callset consists
of a total of 6,796,267 SVs, includi joint genotyping. Notably, only SVs larger than 10 Mbp were filtered. The final callset cons
of a total of 6,796,267 SVs, including 4,101,354 deletions, 726,560 duplications, 558,860
insertions, and 1,409,493 inversions. S , of a total of 6,796,267 SVs, including 4,101,354 deletions, 726,560 duplications, 558,860
insertions, and 1,409,493 inversions. Since an SV can be associated with multiple joint
genotyping models, such as AGGREGATED, BRE insertions, and 1,409,493 inversions. Since an SV can be associated with multiple joint
genotyping models, such as AGGREGATED, BREAKPOINT, BREAKPOINT1, BREAKPOINT2, a
COVERAGE in GraphTyper2, some SVs appeared multiple tim genotyping models, such as AGGREGATED, BREAKPOINT, BREAKPOINT1, BREAKPOINT2
COVERAGE in GraphTyper2, some SVs appeared multiple times in the R4 SV VCF. After
COVERAGE in GraphTyper2, some SVs appeared multiple times in the COVERAGE in GraphTyper2, some SVs appeared multiple times in the R4 SV VCF. After COVERAGE in GraphTyper2, some SVs appeared multiple times in the R4 SV VCF. After

consolidating these models, the total number of unique SVs was reduced to 2,208,044,
comprising 1,367,118 deletions, 184,367 duplications, 186,290 insertions, and 470,269
inversions.
On average, 15,640 high-quality SVs wer comprisions.
Comprisions. Compresents 1,574 duplications, 6,246 insertions, and 7 inversions. Similar to the patterns observed
SNVs and indels, AA samples exhibited a higher number of SVs compared to other group
(Figure 4) On average
1,574 dupli
SNVs and ir
(**Figure 4**).)
)
| ا 1,574 duplications, 6,246 insertions, and 7 inversions. Similar to the patterns observed with
SNVs and indels, AA samples exhibited a higher number of SVs compared to other groups
(Figure 4).
<u>LD reference panel from ADSP </u> SNVs and indels, AA samples exhibited a higher number of SVs compared to other groups

LD reference panel from ADSP R4 data

(Figure 4).
LD reference panel from ADSP R4 data
Starting with the ADSP Integrated Phenotypes list which includes 32,236 samples (5,096 A
2,777 Asian, 10,438 Hispanic, 12,692 NHW, and 1,233 others), we constructed panels f (Figure 4).
LD referer
Starting w
2,777 Asia
Asian, Hisp $\begin{array}{c} \underline{\mathsf{I}} \\ \underline{\mathsf{I}} \\ \underline{\mathsf{I}} \\ \underline{\mathsf{I}} \end{array}$ S is the set of 2,777 Asian, 10,438 Hispanic, 12,692 NHW, and 1,233 others), we constructed panels for AA,
2,777 Asian, 10,438 Hispanic, 12,692 NHW, and 1,233 others), we constructed panels for AA,
Asian, Hispanic, and NHW separately. The Asian, Hispanic, and NHW separately. The panels were built using both SNVs and indels. The
number of variants included in each panel was as follows: 46,462,895 variants (43,494,096
SNVs and 2,968,799 indels) for AA, 25,779 number of variants included in each panel was as follows: 46,462,895 variants (43,494,096
SNVs and 2,968,799 indels) for AA, 25,779,737 variants (24,235,526 SNVs and 1,544,211 inde
for Asians, 57,683,258 variants (54,000,7 SNVs and 2,968,799 indels) for AA, 25,779,737 variants (24,235,526 SNVs and 1,544,211 inc
for Asians, 57,683,258 variants (54,000,742 SNVs and 3,682,516 indels) for Hispanics, and
44,629,226 variants (41,794,882 SNVs and 2

for Asians, 57,683,258 variants (54,000,742 SNVs and 3,682,516 indels) for Hispanics, and
44,629,226 variants (41,794,882 SNVs and 2,834,344 indels) for NHW.
We performed emeraLD [25] with the following parameters, --mac 44,629,226 variants (41,794,882 SNVs and 2,834,344 indels) for NHW.
We performed emeraLD [25] with the following parameters, --mac > 5, --threshold 0.2, an
window 5000000. Each segment was analyzed by 5Mb window with a 3M window 5000000. Each segment was analyzed by 5Mb window with a 3Mb overlapping, then
we concatenated all segments, removing duplicate records. As a result, we identified
3,153,513,864, 1,795,829,862, 4,990,587,680, and 3, ノノノミ ノモ window 5000000. Each segment was analyzed by 5Mb window with a 3Mb overlapping, then
we concatenated all segments, removing duplicate records. As a result, we identified
3, 153, 513, 864, 1, 795, 829, 862, 4, 990, 587, 68 we concatenated all segments, removing duplicate records. As a result, we identified
3, 153,5 13,864, 1,795,829,862, 4,990,587,680, and 3,205,008,552 pairs of variants for the AA,
Asian, Hispanic, and NHW, respectively. A 3, 153,513,864, 1,795,829,862, 4,990,587,680, and 3,205,008,552 pairs of variants for Asian, Hispanic, and NHW, respectively. Among these, the proportion of pairs with R^2 6.3% for AA, 11.0% for Asians, 5.7% for Hispani Asian, Hispanic, and NHW, respectively. Among these, the proportion of pairs with $R^2 > 0.8$ wa
6.3% for AA, 11.0% for Asians, 5.7% for Hispanics, and 9.2% for NHW. For $R^2 > 0.2$, the
proportions were 33.5%, 45.4%, 32.4% e
expressive
to onsure

Harmonized phenotypic data for 28,000+ participants with WGS

6.3% for AA, 11.0% for Asians, 5.7% for Hispanics, and 9.2% for NHW. For R^e > 0.2, the
proportions were 33.5%, 45.4%, 32.4%, and 40.9%, respectively.
Harmonized phenotypic data for 28,000+ participants with WGS
The AD **Harmonized phenotypic data for 28,000+ participants with W**
The ADSP Phenotype Harmonization Consortium (ADSP-PHC) was
endophenotype data from various cohort studies. The group collar
high-quality phenotype harmonization $\frac{1}{2}$ $\frac{1}{2}$ - 「 ()
- c (()
- c () endophenotype data from various cohort studies. The group collaborates with ADSP to ensure
high-quality phenotype harmonization across multiple domains, and document data availability
and harmonization processes. Currently endophenotype data from various consistent in group collaborates and document data availability
and harmonization processes. Currently, available phenotypes from the ADSP-PHC include
autopsy measures of neuropathology, flu mand harmonization processes. Currently, available phenotypes from the ADSP-PHC include
autopsy measures of neuropathology, fluid biomarkers of AD neuropathology, positron
emission tomography measures of amyloid and tau pa and hay measures of neuropathology, fluid biomarkers of AD neuropathology, positron
emission tomography measures of amyloid and tau pathology, structural brain imaging usin
magnetic resonance imaging, diffusion tensor imag emission tomography measures of amyloid and tau pathology, structural brain imaging using
magnetic resonance imaging, diffusion tensor imaging, longitudinal measures of cognition, and
cardiovascular risk factor data. Harmo magnetic resonance imaging, diffusion tensor imaging, longitudinal measures of cognition, an
cardiovascular risk factor data. Harmonization methods are detailed in the Methods - ADSP
Phenotype Harmonization Consortium (ADS

magnetic resonance imaging, annuation tend imaging, rengonal measures of cognition, and
cardiovascular risk factor data. Harmonization methods are detailed in the Methods - ADSP
Phenotype Harmonization Consortium (ADSP-PHC Phenotype Harmonization Consortium (ADSP-PHC) section.
In order to ensure the highest quality harmonization is conducted, the ADSP-PHC harmonize
available phenotypic data, regardless of sequencing status, which the researc Phenotype Harmonization Consortium (ADSP-PHC) section.
In order to ensure the highest quality harmonization is condu
available phenotypic data, regardless of sequencing status, w
access directly from each cohort. ADSP-PHC ー ここれ available phenotypic data, regardless of sequencing status, which the research community can
access directly from each cohort. ADSP-PHC deliverables are then subset to participants with
available sequencing data. To increa access directly from each cohort. ADSP-PHC deliverables are then subset to participants with
available sequencing data. To increase the value and usage of the ADSP dataset, the ADSP-PHC
has selected >9k phenotypes across 1 available sequencing data. To increase the value and usage of the ADSP dataset, the ADSP-PH
has selected >9k phenotypes across 10 domains, expanding on the three domains released
has selected >9k phenotypes across 10 domai has selected >9k phenotypes across 10 domains, expanding on the three domains released has selected >9k phenotypes across 10 domains, expanding on the three domains released

previously (NIAGADS ng00067.v9). Figure 5 summarizes the harmonized data availability for
more than 15,927 participants with whole genome sequencing and harmonized phenotypic
data in NIAGADS. A data availability and explor data in NIAGADS. A data availability and explorer tool are available online via

Summary of data files shared in this collection

https://vmacdata.org/adsp-phc.

Summary of data files shared in this collection

All the described R4 data have been released in NIAGADS Data Sharing Service

https://dss.niagads.org/). These include individual level CRAMs Summary of data files shared in
All the described R4 data have be
https://dss.niagads.org/). These
aggregated files including the join $\frac{S}{C}$ All the described R4 data have been released in NIAGADS Data Sharing Service (DSS, https://dss.niagads.org/). These include individual level CRAMs, gVCFs, and SV VCFs,
aggregated files including the joint-genotyped VCFs from 35,014 unique individuals.
alternative solutions for users who may not require V aggregated files including the joint-genotyped VCFs from 35,014 unique individuals. We offer
alternative solutions for users who may not require VCFs with detailed information for their
analyses. These options include geno alternative solutions for users who may not require VCFs with detailed information for their
analyses. These options include genotype information, full quality metrics, or ADSP QC details
and are organized by bi-allelic an analyses. These options include genotype information, full quality metrics, or ADSP QC detail
and are organized by bi-allelic and multi-allelic variants. Additionally, some of these files are
available in the CoreArray Gen and are organized by bi-allelic and multi-allelic variants. Additionally, some of these files are
available in the CoreArray Genomic Data Structure (GDS) [26], an alternative format to VCF
designed specifically for R users available in the CoreArray Genomic Data Structure (GDS) [26], an alternative format to VCF
designed specifically for R users. Besides we provided sequencing methods, quality data
metrics, variant metrics, phenotypes and re designed specifically for R users. Besides we provided sequencing methods, quality data
metrics, variant metrics, phenotypes and readmes along with these data files. Annotation a
LD reference panel files are available in o designed specifically for R users. Benotypes and readmes along with these data files. Annotatio
LD reference panel files are available in open access. We summarize these files by featur
file size in Table 2.
Browser of var LD reference panel files are available in open access. We summarize these files by features and file size in Table 2.

Browser of variants and annotations of diversified samples

File size in Table 2.

<u>Browser of variants and annotations of diversified samples</u>

To allow users to explore the ADSP R4 genotypes without downloading the data, we provide

users with two different visualization options. File size in Table 2.
Browser of variant
To allow users to e:
users with two difficallows users to che [erre]
|
|} To allow users to explore the ADSP R4 genotypes without downloading the data, we provide Users with two different visualization options. First, VariXam (https://varixam.niagads.org/)
allows users to check the alleles and QC quality of any genetic variants in any callsets (this
release R4 and earlier ADSP relea allows users to check the alleles and QC quality of any genetic variants in any callsets (this release R4 and earlier ADSP releases) (Figure 6A). Users can query by the SNV/gene (IDs or genomic coordinates) level or view a release R4 and earlier ADSP releases) (Figure 6A). Users can query by the SNV/gene (IDs or
genomic coordinates) level or view all variants in a particular genomic region (Methods –
VariXam).
Second, the NIAGADS Alzheimer's

release R4 and earlier ADSP releases) (Figure 6A). Users can query by the SNV/gene (IDs or
genomic coordinates) level or view all variants in a particular genomic region (Methods –
VariXam).
Second, the NIAGADS Alzheimer's VariXam).
VariXam).
Second, the NIAGADS Alzheimer's GenomicsDB [27] (https://www.niagads.org/genomics)
allows users to explore the variants and annotations in a broader genomic context (Method)
NIAGADS Alzheimer's Genomics VariXam).
Second, the
allows user
NIAGADS h
variant rec ミミ『『\く allows users to explore the variants and annotations in a broader genomic context (Methods - NIAGADS Alzheimer's GenomicsDB). It provides detailed reports of genetic associations from NIAGADS hosted GWAS summary statistic NIAGADS Alzheimer's GenomicsDB). It provides detailed reports of genetic associations from NIAGADS Alzheimer's GenomicsDBJ. It provides detailed reports of genetic associations from
NIAGADS hosted GWAS summary statistic datasets in the context of genes and annotated
variant records. ADSP R4 variants are flagged, variant records. ADSP R4 variants are flagged, and can be filtered by the most severe
consequence predicted by the ADSP annotation pipeline [22]. Besides, variant reports also
include the full ADSP annotation results (all consequence predicted by the ADSP annotation pipeline [22]. Besides, variant reports
include the full ADSP annotation results (all predicted consequences, CADD [28]
deleteriousness, loss of function [29], and FAVOR annotat consequence full ADSP annotation results (all predicted consequences, CADD [28]
deleteriousness, loss of function [29], and FAVOR annotations [29]), allele frequencies
(including, gnomAD [20], 1000Genomes [30]), and links deleteriousness, loss of function [29], and FAVOR annotations [29]), allele freque
(including, gnomAD [20], 1000Genomes [30]), and links out to related web-resou
GenomicsDB genome browser provides an ADSP R4 variant track (including, gnomAD [20], 1000Genomes [30]), and links out to related web-resources.
GenomicsDB genome browser provides an ADSP R4 variant track that can be recolored
on annotations (e.g., coding vs non-coding variants, con (including, given the party and the party), and the variant track that can be recolored based becommics DB genome browser provides an ADSP R4 variant track that can be recolored based on annotations (e.g., coding vs non-co on annotations (e.g., coding vs non-coding variants, consequence type). This track can be compared against the summary statistic data or other tracks in GenomicsDB (Figure 6B). compared against the summary statistic data or other tracks in GenomicsDB (Figure 6B).

Discussion

The R4 WGS dataset is the largest and most diverse whole genome data collection for Alzheimer's disease to date. Numerous studies utilizing the ADSP WGS data in a smaller scale
(previous R3 release, 46% of the current sample size, less diversified) have led to multiple
findings in AD genetics [31-48], inc (previous R3 release, 46% of the current sample size, less diversified) have led to multiple
findings in AD genetics [31-48], including but not limited to the discovery of 1) new AD genes
PLEC, UTRN, TP53, and POLD1 [33] u findings in AD genetics [31-48], including but not limited to the discovery of 1) new AD genes $PLEC$, UTRN, TP53, and $POLD1$ [33] using a novel approach, GeneEMBED, designed for studyin gene interactions; $DLG2$ and $DTNB$ via FIEC, UTRN, TP53, and POLD1 [33] using a novel approach, GeneEMBED, designed for studyin
gene interactions; DLG2 and DTNB via rvGWAS on family datasets [31]; 2) rare genetic variant
contributing to AD risk [32]; 3) sex-spe [44]; and (5) novel AD associations in Ashkenazi Jews with variants that are exceedingly rare or absent in other European ancestry populations [49]. The ADSP WGS data has also enable the further study of APOE in different gene interactions; DEO2 and DTNB via TVOWAS on family datasets [51]; 2) rare genetic variants
contributing to AD risk [32]; 3) sex-specific loci identified in family-based designs [36]; 4) novel
AD risk loci on 13q33.3 via contribution 13q33.3 via admixture mapping analyses in the Caribbean Hispanic populations [44]; and (5) novel AD associations in Ashkenazi Jews with variants that are exceedingly rare or absent in other European ancestry p [44]; and (5) novel AD associations in Ashkenazi Jews with variants that are exceedingly rare or
absent in other European ancestry populations [49]. The ADSP WGS data has also enable the
further study of APOE in different Entrin other European ancestry populations [49]. The ADSP WGS data has also enable the
further study of APOE in different ancestries [37, 38, 41, 46, 47] and identification of
associations of AD risk in multiple population Further study of APOE in different ancestries [37, 38, 41, 46, 47] and identification of
associations of AD risk in multiple population groups with human viruses detected among
unmapped reads in the WGS data [50]. Finally, associations of AD risk in multiple population groups with human viruses detected an
unmapped reads in the WGS data [50]. Finally, these data have led to new insights in
onset Alzheimer's (EOAD) [39], enabled the detection unmapped reads in the WGS data [50]. Finally, these data have led to new insights in Early-
onset Alzheimer's (EOAD) [39], enabled the detection of nuclear and mitochondrial copy
number variations and structural variants a number variations and structural variants and their association with LOAD [34, 40, 51], as well

und raced several challenges in this project. First, samples are recruited from independent as the generation of ancestry-specific polygenic risk score in the Amish population [35] and the
development of a pipeline for calling mitochondrial sequence variants and haplogroups [52].
We faced several challenges in th as the generator of ancestra polygenment of ancestra in the complete development of a pipeline for calling mitochondrial sequence variants and haplogroups [52].
We faced several challenges in this project. First, samples a We faced several challenges in this project. First, samples are recruited from independent
cohorts/projects of various study designs and sequencing experiments are funded during the
past 10 years, and challenges rise for G ヽ く ド く ト ド Cohorts/projects of various study designs and sequencing experiments are funded during t
past 10 years, and challenges rise for GCAD to process and cumulatively integrate all genoidata. Unlike AllofUS [53]UKBiobank [54] th past 10 years, and challenges rise for GCAD to process and cumulatively integrate all genomic
data. Unlike AllofUS [53]UKBiobank [54] the ADSP program has to address the issue of
heterogeneity in recruitment criteria, phen particulary of the Allof US [53] UKBiobank [54] the ADSP program has to address the issue of heterogeneity in recruitment criteria, phenotype data collection protocols, and sequencing platform and configuration changes as data. Unlike AllofUS [53]UKBiobank [54] the ADSP program has to address the issue of
heterogeneity in recruitment criteria, phenotype data collection protocols, and sequencing
platform and configuration changes as sequenci metring in an among production, protocollection process and analysis best practices
continue to evolve. Subtle batch effects may persist in our datasets even after we process a
sequence data and perform thorough quality ch plantinue to evolve. Subtle batch effects may persist in our datasets even after we process
sequence data and perform thorough quality checks during the data harmonization proce
sample, variant and phenotype levels.
Second

continue to evolve the first may perform the matter of the test in process a
sample, variant and phenotype levels.
Second, compared to previous releases (R1 and R3 for genomes), this ADSP R4 dataset more
than doubled the s sample, variant and phenotype levels.
Second, compared to previous releases (R1 and R3 for genomes), this ADSP R4 dataset more
than doubled the sample size (35,014 vs 16,285) with the biggest growth in the HL group
(Supple than doubled the sample size (35,014 vs 16,285) with the biggest growth in the HL group s t (εε than doubled the sample size (35,014 vs 16,285) with the biggest growth in the HL group
(Supplementary Figure 3), yet the Asian population is still significantly underrepresented. The
gaps must be addressed if we are to f (Supplementary Figure 3), yet the Asian population is still significantly underrepresented.
gaps must be addressed if we are to fully understand the distribution and effect of human
genetic variation in AD. Power analyses gaps must be addressed if we are to fully understand the distribution and effect of human
genetic variation in AD. Power analyses show we still need 18,500 cases and 18,500 contro
per ancestry group to gain enough power fo

genetic variation in AD. Power analyses show we still need 18,500 cases and 18,500 controls
per ancestry group to gain enough power for detecting variants with MAF of 0.005.
In conclusion, we have shown that the ADSP R4 re general carrier in all alternative analyses shown the calculation in AD. Power and September 20,005.
In conclusion, we have shown that the ADSP R4 resource offers researchers in the
neurodegenerative field multiple tools t per ances of the state of the SP Participate of the SP R4 resource offers researchers in the neurodegenerative field multiple tools to explore and analyze the genetic variations with these diseases. We anticipate that this | r ヽ
。 :<
ト neurodegenerative field multiple tools to explore and analyze the genetic variatior
with these diseases. We anticipate that this data will significantly influence nearly
studies on common and rare variants in Alzheimer's d with these diseases. We anticipate that this data will significantly influence nearly all ongoing
studies on common and rare variants in Alzheimer's disease (AD), with an even greater impact
as additional samples from dive studies on common and rare variants in Alzheimer's disease (AD), with an even greater impac
as additional samples from diverse populations become available. Moving forward, we plan to
provide annotations for the identified as additional samples from diverse populations become available. Moving forward, we plan to
provide annotations for the identified structural variants (SVs) further empowering researchers
rovide annotations for the identif provide annotations for the identified structural variants (SVs) further empowering researchers

researchers

The identified structural variants (SVs) further empowering researchers

The identified structural variants (SVs provide annotations for the identified structural variants $(1,1)$ further empowering researchers.

analyses.
The ADSP program needs to share data in a manner that supports the privacy and consent
preferences of participants. Members of the scientific community can access most ADSP
resources (individual level sequence fi The ADSP
preferenc
resources
platform
access, wl preferences of participants. Members of the scientific community can access most ADSP
resources (individual level sequence files, phenotype data, VCFs) through the NIAGADS DSS
platform (dbGAP-like). Some companion data (e. preferences (individual level sequence files, phenotype data, VCFs) through the NIAGADS D
platform (dbGAP-like). Some companion data (e.g. annotation LD reference panel) are op
access, while VariXam variant server and NIAG resources (dbGAP-like). Some companion data (e.g. annotation LD reference panel) are open
access, while VariXam variant server and NIAGADS genomicsDB provide a preview of the dat
without any restrictions.
The ADSP R4 datas placess, while VariXam variant server and NIAGADS genomicsDB provide a preview of the dat
without any restrictions.
The ADSP R4 dataset of over 347 million variants is available to qualified investigators at
https://dss.ni

access, while Variant server and NIAGADS generates provide a provide the and
without any restrictions.
The ADSP R4 dataset of over 347 million variants is available to qualified investigators at
https://dss.niagads.org/dat with any restriction
The ADSP R4 dataset of o
https://dss.niagads.org/c
Methods $\frac{1}{\sqrt{2}}$ METTE AT DRUCTER FRAMENTIFIED ADDITIONALLY AND ADDITION AT A HIS SEARCH IN ENGLISH AT AN AREA OF OUTSIDENTIFIE
https://dss.niagads.org/datasets/ng00067/.
Sequence existing ancestrally diverse cohorts via the ADSP Follow-Up

Methods

֦

<u>https://decentage.org/datasets/ng.org/07/2</u>
Methods
Sequence existing ancestrally diverse cohor
The ADSP-FUS is a NIA initiative focused on i The ADSP-FUS is a NIA initiative focused on identifying genetic risk and protective variants for Alzheimer Disease (AD) by expanding the ADSP cohorts beyond primarily participants with nor
Hispanic Whites of European Ancestry (NHW). Given the lack of ancestral diversity in the ADSF
the ADSP-FUS was designed to sequenc over 100,000 participants from diverse ancestries. See Supplementary method on cohort descriptions. Workflows within the FUS infrastructure (Supplementary Figure 1) include the ADSP-FUS was designed to sequence existing ancestrally diverse and unique cohorts. The
current phase for ADSP-FUS, ADSP- FUS 2.0: The Diverse Population Initiative, focuses on
Hispanic/Latino (HL), non-Hispanic Black w the ADSP-FUS, ADSP-FUS 2.0: The Diverse Population Initiative, focuses on
Hispanic/Latino (HL), non-Hispanic Black with African Ancestry (NHB-AA), and Asian populatio
(e.g., the Asian cohort for Alzheimer's disease). The A Hispanic/Latino (HL), non-Hispanic Black with African Ancestry (NHB-AA), and Asian popul
(e.g., the Asian cohort for Alzheimer's disease). The ADSP-FUS initiatives intend to sequen
over 100,000 participants from diverse an (e.g., the Asian cohort for Alzheimer's disease). The ADSP-FUS initiatives intend to sequence
over 100,000 participants from diverse ancestries. See **Supplementary method** on cohort
descriptions. Workflows within the FUS i (e.g., no como participants from diverse ancestries. See **Supplementary method** on cohort
descriptions. Workflows within the FUS infrastructure (**Supplementary Figure 1**) include
biospecimens being processed and DNA prepar over 100,000 participants from diverse ancestries. See Supplementary method on conort
descriptions. Workflows within the FUS infrastructure (Supplementary Figure 1) include
biospecimens being processed and DNA prepared and descriptions. Workhows within the FOS inhastracture (Supplementary Figure 1) include
biospecimens being processed and DNA prepared and allocated for WGS at USUHS and H
All raw sequence data is transferred to the GCAD for p biology and the GCAD for processing and harmonization following
quality control (QC) analysis at the University of Pennsylvania and University of Miami, resultir
in analysis-ready genotype and sequence data. All clinical, Mall required (QC) analysis at the University of Pennsylvania and University of Miami, resulti
in analysis-ready genotype and sequence data. All clinical, genotype and sequence data are
housed at the NIAGADS, which stores, in analysis-ready genotype and sequence data. All clinical, genotype and sequence data are
housed at the NIAGADS, which stores, manages, and distributes ASDP-FUS data to AD
researchers.

Dataflow and Sample processing protocol on SNVs and indels

NIAGADS receives genomes in FASTQ, BAM, or CRAM formats from national or international
investigators through secure FTP or S3 bucket. NIAGADS also received the ID information to moderations.
 Example 2018
 Example 2018

The data flow from receiving data to data sharing is shown in **Supplementary Figure 4**

NIAGADS receives genomes in FASTQ, BAM, or CRAM formats from national or internations

i **Dataflow and
The data flov
NIAGADS rec
investigators
generate a ur** Dataflow and Sample processing protocol on SNVs and indels
The data flow from mooning data to data shoring is shown in C The data flow from receiving data to data sharing is shown in Supplementary Figure 4.
NIAGADS receives genomes in FASTQ, BAM, or CRAM formats from national or internat
investigators through secure FTP or S3 bucket. NIAGADS investigators through secure FTP or S3 bucket. NIAGADS also received the ID information to
generate a unique ADSP ID, companion array data for checking concordance, ADSP minimal
phenotypes, Institutional Certification form invering the university. The secure FTP or S
phenotypes, Institutional Certification forms, and cohort and study information for these
genomes. Once the phenotypes, Institutional Certification forms, and cohort and study information for these
genomes. Once the unique ADSP IDs are received and generated by NIAGADS, GCAD can be_{
production on the genomes. phenotypes, Institutional Certification forms, and control that comp_{ress} increases the unique ADSP IDs are received and generated by NIAGADS, GCAD can l
production on the genomes.
The study is the sense of the sense of t generate the unique ADSP IDs are received and generated by NIAGADS, The Taxa again
production on the genomes. production on the generation of the generation of \mathcal{G}

All genomes are processed using the VCPA pipeline [17] (https://bitbucket.org/NIAGADS/vcpaposition and includes all steps from aligning raw sequence reads to variant calling
using GATK best practices. Depending on the file type received, the first few steps of the
pipeline are slightly different:
1) CRAM as inp using GATK best practices. Depending on the file type received, the first few steps of the
pipeline are slightly different:
1) CRAM as input: decompression to BAMs is required before running the pipeline as if the
input is

-
- 2) BAM as input: roll back to uBAM (i.e. FASTQ) format before running the pipeline as if the pipeline are slightly different:

1) CRAM as input: decompression to BAMs is required before running the pipeline as if the input is in BAM format

2) BAM as input: roll back to uBAM (i.e. FASTQ) format before running the 1) CRAM as input: decompressing the slightly different (1) BAM as input: roll back to input is in FASTQ format
3) FASTQ as input: roll back to input is in FASTQ format
3) FASTQ as input: no special
Using VCPA. FASTOs were ここしし
-

1) input is in BAM format

2) BAM as input: roll back to uBAM (i.e. FASTQ) format before running the pipeline as if the

1) FASTQ as input: no special steps

1) FASTQ as input: no special steps

1) Using VCPA, FASTQs were EXAM as input: roll back
input is in FASTQ forma
FASTQ as input: no spee
ig VCPA, FASTQs were f
7.15). Duplicated reads
el-realignment were doi 2) BAM as input is in FASTQ format

2) FASTQ as input: no special steps

2) FASTQ as input: no special steps

2) Sing VCPA, FASTQs were first mapped/remapped to hg38 reference genome using BWA-me

(v0.7.15). Duplicated rea FASTQ as input: no special
FASTQ as input: no special
ing VCPA, FASTQs were fir
7.15). Duplicated reads wel-realignment were done
stored. Genotype calling
pmosomes (1-22, X, Y, an 3) FAST AST AST OS were first mapp
Using VCPA, FAST QS were first mapp
(v0.7.15). Duplicated reads were the
indel-realignment were done by GAT
was stored. Genotype calling was the
chromosomes (1-22, X, Y, and M).
Joint gen | |
| |
| | | | (v0.7.15). Duplicated reads were then marked using BamUtil (v1.0.13). Base-recalibration and
indel-realignment were done by GATK (v4.1.1) and a CRAM that contains all above information
was stored. Genotype calling was then indel-realignment were done by GATK (v4.1.1) and a CRAM that contains all above information
was stored. Genotype calling was then done by GATK4.1.1, resulting in one gVCF across all
chromosomes (1-22, X, Y, and M).
Joint g

inder-realignment were done by Campach (value) and a Critical and a Containe in the gVCF across all
chromosomes (1-22, X, Y, and M).
Joint genotyping SNVs and indels – we followed the steps in GATK best practices (version
 chromosomes (1-22, X, Y, and M).
Joint genotyping SNVs and indels – we followed the steps in GATK best practices (version
GATK4.1.1) to create a joint genotyped called VCFs using gVCFs generated in step 2. VQSR
model was Joint genotyping SNVs and indels -
GATK4.1.1) to create a joint genoty
model was performed all chromos
GRAF-POP
GRAF-pop offers a novel method for
traditional approaches that require **リ(r** (1)

GRAF-POP

 $GATK4.1.1$) to create a joint genotyped called VCFs using gVCFs generated in step 2. VQSF
model was performed all chromosomes (1-22, X, Y, and M). This is the "Preview VCF".
GRAF-POP
GRAF-pop offers a novel method for glo model was performed all chromosomes (1-22, X, Y, and M). This is the "Preview VCF".

GRAF-POP

GRAF-pop offers a novel method for global ancestry inference [19], distinguishing itself fro

traditional approaches that requi GRAF-POP
GRAF-pop offers a novel method for global ancestry inference [19], distinguishing itsel
traditional approaches that require genotypes from reference populations or study pa
to determine an individual's ancestry. I **GRAF-POP**
GRAF-pop
traditional
to determin
selected SN
genotypes
genotypes.
Population Fraditional approaches that require genotypes from reference populations or study participar
to determine an individual's ancestry. Instead, GRAF-pop precomputes allele frequencies for
selected SNPs across reference popula traditional approaches that require general approaches the term of permitted to determine an individual's ancestry. Instead, GRAF-pop precomputes allele frequencies for selected SNPs across reference populations and infers to determine an individual's ancestry. Instead, GRAF-pop precomputes allele frequencies for
selected SNPs across reference populations and infers ancestry by directly comparing the tes
genotypes to these predefined frequen

Population substructure

SERET FOR ACT SNP ACT SNP ACT SNP ACT SNP (SPECIFY) and inference populations to these predefined frequencies, eliminating the need for other individual genotypes.
 Population substructure

We included bi-allelic vari genotypes to these predefined frequencies, summaring the need for our care individually
genotypes.
We included bi-allelic variants that have 1) passed ADSP QC, i.e. with VFLAG=0), 2) All
Ratio >0.3 or <0.7, 3) MAF>0.02, 4 Population

We include

We include

Ratio >0.3 o

larger than

Across all cl

(PCs)/genet

[55] along v We included bi-allelic variants that have 1) passed ADSP QC, i.e. with VFLAG=0), 2) Allelic Read Ratio >0.3 or <0.7, 3) MAF>0.02, 4) call rate >99.5% and with hardy-Weinberger test P value
larger than 0.0005. Then we applied LD pruning with R^2 cutoff at 0.05, window size of 500K bp.
Across all chromosomes, 146,964 larger than 0.0005. Then we applied LD pruning with R^2 cutoff at 0.05, window size of 500K b
Across all chromosomes, 146,964 pruned variants remained for principal components
(PCs)/genetic relationship matrix (GRM) cal larger than 0.0005. Then we applied LD pruning with R⁻ cutorf at 0.05, window size of 500K bp.
Across all chromosomes, 146,964 pruned variants remained for principal components
(PCs)/genetic relationship matrix (GRM) cal (PCs)/genetic relationship matrix (GRM) calculations. We used R package "GENESIS" (2
[55] along with GWASTools(1.36.0) [56] and SNPRelate(1.24.0) [26, 57] to calculate th
GRM. 5513 out of 36361 subjects were grouped as re (55) along with GWASTools(1.36.0) [56] and SNPRelate(1.24.0) [26, 57] to calculate the PCs and GRM. 5513 out of 36361 subjects were grouped as related at default threshold. Note, the threshold value on kinobj used for dec FORM. 5513 out of 36361 subjects were grouped as related at default threshold. Note, the threshold value on kinobj used for declaring each pair of participants as related or unrelated.
The default value is $2^x(-11/2) \approx 0.$

Threshold value on kinobj used for declaring each pair of participants as related or unrelate.
The default value is 2^(-11/2) ~ 0.022, corresponding to 4th degree relatives.
gnomAD 4.0 genome sites data provided allele fre The default value is $2^x(-11/2) \approx 0.022$, corresponding to 4th degree relatives.
gnomAD 4.0 genome sites data provided allele frequency at ancestry level, which can be used
for ancestry inference. There are 10 ancestries The definite interact value is 2001. The definition of the definition of the definition of ancestry level, when
for ancestry inference. There are 10 ancestries in the data: African/African-Ar
Amish (ami), Latino (amr), Ash }
f
/ gnomand and genome sites data provided and allegency at anticomplement, there is do
for ancestry inference. There are 10 ancestries in the data: African/African-American (afr),
Amish (ami), Latino (amr), Ashkenazi Jewish (for ancestry interested inference. There are there are the data-independent parameters (and),
Amish (ami), Latino (amr), Ashkenazi Jewish (asj), East Asian (eas), Finnish (fin), Middle East
The data: African-African-Afric A_{max} , A_{\text

matched between ADSP pruned file and gnomAD4.0 and passed to Software SCOPE
(https://github.com/sriramlab/SCOPE.git) to perform the ancestry proportion analysis. S
was assigned to ancestry group in which it has the highest

ADSP Sample level QC protocol

(https://github.com/sriramlab/SCOPE.git) to perform the ancestry proportion analys
was assigned to ancestry group in which it has the highest ancestry proportion value
ADSP Sample level QC protocol
Four checks were perform Was assigned to ancestry group in which it has the highest ancestry proportion value.

ADSP Sample level QC protocol

Four checks were performed to identify potentially low-quality samples for exclusion from the

VCFs: 1) MOSP Sample level QC protocol

Four checks were performed to identify potentially low-quality samples for exclusion

VCFs: 1) SNV concordance check with available GWAS genotypes; 2) sex check for pos

sample swaps or misre Four checks were performed to identify potentially low-quality samples for exclusion from the VCFs: 1) SNV concordance check with available GWAS genotypes; 2) sex check for possible
sample swaps or misreporting; 3) contamination check for possible sample swaps; and 4)
relatedness check to confirm known relationship

Sample swaps or misreporting; 3) contamination check for possible sample swaps; and 4)
relatedness check to confirm known relationships, identify unknown genetically identical
duplicates, and assess potential cryptic relat sample state of confirm known relationships, identify unknown genetically identical
duplicates, and assess potential cryptic relatedness.
1) SNV concordance check - this was done by comparing genotypes taken from existing duplicates, and assess potential cryptic relatedness.

1) SNV concordance check - this was done by comparing genotypes taken from existing (

array data and genotypes from the preview pVCFs on all overlapping genotypes. T 1) SNV concordance check - this was done by comp
array data and genotypes from the preview pVCI
this analysis is to ensure that samples and IDs ma
calling processes.
First, GWAS datasets were converted into VCF fo
pVCF an

1) Survey data and genotypes from the preview pVCFs on all overlapping genotypes. The goal of
this analysis is to ensure that samples and IDs match throughout the data management and
calling processes.
First, GWAS datasets this analysis is to ensure that samples and IDs match throughout the data management and
calling processes.
First, GWAS datasets were converted into VCF format. Then GATK was used to compare the
pVCF and GWAS file genotyp calling processes.
First, GWAS datasets were converted into VCF format. Then GATK was used to compare the
pVCF and GWAS file genotypes using the following parameters: Java -jar
GenomeAnalysisTK.jar -T GenotypeConcordance -First, GWAS datas
pVCF and GWAS f
GenomeAnalysis¹
[GWAS_VCF] -o
[PVCF] is the Previ
samples kept, the
Sex checks were p
following steps: pVCF and GWAS file genotypes using the following parameters: Java -jar
GenomeAnalysisTK.jar -T GenotypeConcordance -R hg38.fa -eval [PVCF] -comp
[GWAS_VCF] -o [OUTFILE] where [GWAS_VCF] is the VCF file converted from GWAS, pVCF and GWAS file genotypes using the following parameters: Java –jar
GenomeAnalysisTK.jar –T GenotypeConcordance –R hg38.fa –eval [F
[GWAS_VCF] -o [OUTFILE] where [GWAS_VCF] is the VCF file converted fro
[PVCF] is the Pr GenomeAnalysisTK.jar –T GenotypeConcordance –R hg38.fa -eval [PVCF] -comp [GWAS VCF] -o [OUTFILE] where [GWAS VCF] is the VCF file converted from GWAS, nd [OUTFILE] is the user-selected output filename. Of all
concordance of > 0.91.
sing BCFtools [58], VCFtools [59] and PLINK [60] with the
romosome X pvcf into plink format
oseudo-autosomal region (PAR)
:0.05, run 'impute-

- Entertainment of a concordance of > 0.91.

Sex checks were performed using BCFtools [58], VCFtools [59] and PLINK [60] with the

following steps:

a. Use bcftools to convert chromosome X pvcf into plink format

b. Filter o Sex checks were performed using BCFtools [58], VCF1
following steps:
a. Use bcftools to convert chromosome X pvcf into |
b. Filter out chromosome X pseudo-autosomal regic
c. Filter out SNVs with MAF<0.05, run 'impute-sex'
 2) Sex checks were performed using BCFtools [58], VCFtools [59] and PLINK [60] with the
	-
	-
	-
	-

Following steps:

a. Use beftools to convert chromosome X pvcf into plink format

b. Filter out chromosome X pseudo-autosomal region (PAR)

c. Filter out SNVs with MAF<0.05, run 'impute-sex'

d. Run sex-check using PLINK f Forecoming steps:
a. Use beftools
b. Filter out SN
d. Run sex-chec
Comparing the r
filtering on mino
pseudoautosom
1.918 samples w a. Filter out chromosome X pseudo-autosomal region (PAR)
c. Filter out SNVs with MAF<0.05, run 'impute-sex'
d. Run sex-check using PLINK for comparison
Comparing the results of 'impute-sex' in BCFtools and ''--sex-chec
fil c. Filter out SNVs with MAF<0.05, run 'impute-sex'
d. Run sex-check using PLINK for comparison
Comparing the results of 'impute-sex' in BCFtools and "--sex-
filtering on minor allele frequencies (MAFs) or without exclud
ps d. Run sex-check using PLINK for comparison
Comparing the results of 'impute-sex' in BCFtools an
filtering on minor allele frequencies (MAFs) or witho
pseudoautosomal region (PAR), the findings of the ty
1,918 samples with Comparing the results of 'impute-sex' in BCFto
filtering on minor allele frequencies (MAFs) or
pseudoautosomal region (PAR), the findings of
1,918 samples with *F*-statistic values between
0.7. A total of 74 samples were i filtering on minor allele frequencies (MAFs) or without excluding variants in the
pseudoautosomal region (PAR), the findings of the two approaches are the same. The
1,918 samples with *F*-statistic values between 0.2 and 0 pseudoautosomal region (PAR), the findings of the two approaches are the same
1,918 samples with *F*-statistic values between 0.2 and 0.8, but only 392 betweer
0.7. A total of 74 samples were identified to be the incorrect

parameter of the finding amples with *F*-statistic values between 0.2 and 0.8, but only 392 between 0.3 and 0.7. A total of 74 samples were identified to be the incorrect sex. The submitting centers agreed that these subje 0.7. A total of 74 samples were identified to be the incorrect sex. The submitting centers
agreed that these subjects needed to be dropped.
Sample-specific contamination checking was performed by using VerifyBamID [61] to
 agreed that these subjects needed to be dropped.

Sample-specific contamination checking was performed by using VerifyBamID [61] to

calculate the concordance estimate between the GWAS genotype data and the BAM file.

This Sample-specific contamination checking was perfo
calculate the concordance estimate between the G
This approach provides information that can be int
contamination or swapping using the GWAS-BAM (
modeling approach was used 3) Sample-specific contamination checking was performed by using VerifyBamID [61] to calculate the concordance estimate between the GWAS genotype data and the BAM f
This approach provides information that can be interpreted to identify potential samp
contamination or swapping using the GWAS-BAM contaminat This approach provides information that can be interpreted to identify potential sample
contamination or swapping using the GWAS-BAM contamination estimate. The 'FREEMIX
modeling approach was used in this analysis. We use This approach provides interpretent provides interpretent provides interpretent contamination or swapping using the GWAS-BAM contamination estimate. The 'FREEMIX modeling approach was used in this analysis. We used suggest contamination taken from the VerifyBamID website: a sample is potential contaminated if
contamination taken from the VerifyBamID website: a sample is potential contaminated if
the FREEMIX value is >0.05. The command line t montamination taken from the VerifyBamID website: a sample is potential contar
the FREEMIX value is >0.05. The command line to generate FREEMIX contaminat
estimates value is as follows: verifyBamID --vcf [INPUT.VCF] --bam the FREEMIX value is >0.05. The command line to generate FREEMIX contamination
estimates value is as follows: $verifyBandID --vcf$ [INPUT.VCF] --bam [INPUT.BAM] --
out [OUTPUT.PREFIX] --verbose --ignoreRG where [INPUT.VCF] is the VCF estimates value is as follows: $verifyBamID - -vcf$ [INPUT.VCF] --bam [INPUT.Bandle] out [OUTPUT.PREFIX] --verbose --ignoreRG where [INPUT.VCF] is the VCF file converted from GWAS, [INPUT.BAM] is the BAM file generated using VCPA1 estimates value is as follows: verifyBamID --vcf [INPUT.VCF] --bam [INPUT.BAM] --
out [OUTPUT.PREFIX] --verbose --ignoreRG where [INPUT.VCF] is the VCF file
converted from GWAS, [INPUT.BAM] is the BAM file generated using out [OUTPUT.PREFIX] --verbose --ignoreRG where [INPUT.VCF] is the VCF file e generated using VCPA1.1, and
appended to output files. [OUTPUT.PREFIX] is the user-selected prefix to be appended to output files.

The BAM files and the SAM files and VCPA1.1, and the SAM files and the USA files and SAM [OUTPUT.PREFIX] is the user-selected prefix to be appended to output files.

-
- Across all samples, all FREEMIX values obtained are <0.05.

Across all samples, all FREEMIX values obtained are <0.05.

Across checks were performed using PLINK as described:

Convert *vcf.gz files into PLINK binary format Across checks were performed using PLINK as described:

Convert *vcf.gz files into PLINK binary format - Run PLINK "

commands to convert "*.vcf.gz' files into "*.bed", "*.bim", at

Combine 22 chromosome-specific PLINK fil a. Convert *vcf.gz files into PLINK binary format - Run PLINK "
commands to convert '*.vcf.gz' files into '*.bed', '*.bim', ar
b. Combine 22 chromosome-specific PLINK filesets into one go
organize
• Clean and rename empty
	- - Clean and rename empty name "." snvs to "chr_pos" format to combine
	- commands to convert '*.vcf.gz' files into '*.bed', '*.bim', and '*.fam' files

	b. Combine 22 chromosome-specific PLINK filesets into one genome-wide set and

	organize

	 Clean and rename empty name "." snvs to "chr_pos" f Combine 22 chromosome-specific PLINK filesets into one genome-wide se

	organize

	• Clean and rename empty name "." snvs to "chr_pos" format to combine

	• Run PLINK "--merge-list" and "--make-bed" commands to combine 22

	• b. Clean and rename empty name "." snvs to "chr_pos" format to combine

	• Clean and rename empty name "." snvs to "chr_pos" format to combine

	• Run PLINK "--merge-list" and "--make-bed" commands to combine 22 binary

	into ● Clean
● Run P
into c
● Run P
Run P
IBD va
● Heing • Run PLINK "--merge-list" and "--make-bed" commands to combine 22 bir

	• Run PLINK "--extract" command to extract only overlapping ~21k SNVs

	Run PLINK to generate pairwise sample IBD values

	• Run PLINK "--genome" comman
		- Run PLINK "-extract" command to extract only overlapping ~21k SNVs
	- - •
	- into one

	Run PLINK "-extract" command to extract only overlapping ~21k SNVs

	PLINK to generate pairwise sample IBD values

	Run PLINK "--genome" command on combined file with ~21k SNVs to get pairwise

	IBD values

	Using t Run PLIN
PLINK to
Run PLIN
IBD value
Using the
PI_HAT i
each pair PLINK to generate pairwise sample IBD values
Run PLINK "--genome" command on combined file with ~21k SNVs to ge
IBD values
Using these IBD values, identify related pairs with IBD PI_HAT > 0.4
PI_HAT measurement [Proportion • Run PLINK "--genome" command on combined

	IBD values

	• Using these IBD values, identify related pairs w

	The PI_HAT measurement [Proportion IBD, i.e. P(IE

	for each pair of samples. All pairs with PI_HAT > 0.

	relatedn IBD values
Using these IBD values, identify related pairs with IBD PI_HAT > 0.4
PI_HAT measurement [Proportion IBD, i.e. P(IBD=2) + 0.5*P(IBD=1)] was reported
each pair of samples. All pairs with PI_HAT > 0.4 were evaluat Using these
PI_HAT me
each pair o'
tedness.
a total of 1,
i pairs of tv
times as tee • Using these IBD values, identify related pairs with IBD PI HAT > 0.4

PI_HAT measurement [Proportion IBD, i.e. $P(IBD=2) + 0.5*P(IBD=1)$
each pair of samples. All pairs with $PI_HAT > 0.4$ were evaluated for
tedness.
a total of 1,205 genetically unique participants identified with multip
is pairs o for each pair of samples. All pairs with $PI_HAT > 0.4$ were evaluated for known
relatedness.
was a total of 1,205 genetically unique participants identified with multiple samples
ng 5 pairs of twins) across the R4 dataset. T For each pair of samples. The pair of strate contract of this included for the samp
relatedness.
vas a total of 1,205 genetically unique participants identified with multiple samp
ng 5 pairs of twins) across the R4 dataset relatedness.

vas a total of

ng 5 pairs of

67 times as 1

res). After rer

genetically u

pased and ca.

e/control stu ヿ(t t …
f c f (including 5 pairs of twins) across the R4 dataset. This included 6 known replicates (sequen
total of 67 times as technical replicates), and 1,199 unintentional replicates (1,173 pairs an
triplicates). After removing the A (including 5 pairs as technical replicates), and 1,199 unintentional replicates (1,173 pairs and 26 triplicates). After removing the ADSP recommended duplicate samples, there were a total of 35,023 genetically unique sampl triplicates). After removing the ADSP recommended duplicate samples, there were a total of
35,023 genetically unique samples. NOTE: there are 24 subjects from the 36k listed in both the
family-based and case/control phenot 35,023 genetically unique samples. NOTE: there are 24 subjects from the 36k listed in both the Samily-based and case/control phenotype files. These samples were either sequenced 1) as p of a case/control study but were als 35,023 genetically unique samples. NOTE: there are 24 subjects from the 36k listed in both the

of a case/control study but were also part of a family so their phenotypes are provided in both
files, or 2) in both a case/control and family-based study. ADSP recommends using samples
from family-based sets.
After perfor files, or 2) in both a case/control and family-based study. ADSP recommends using samples
from family-based sets.
After performing sex check, GWAS concordance and contamination checks for each sample,
together with the me from family-based sets.

After performing sex check, GWAS concordance and contamination checks for each sample,

together with the metrics we collected per sample in VCPA, we designed a list of criteria to

drop or fail sa After performing sex choro
together with the metric
drop or fail samples base
sample is dropped if a) if
related sequencing metr
check, or sex check; or 4
in less than 3 sequencing ノ
t c c r c i After with the metrics we collected per sample in VCPA, we designed a list of criteria to
drop or fail samples based on this sample level QC process (**Supplementary Figure 5**) –. A
sample is dropped if a) its average genom to drop or fail samples based on this sample level QC process (**Supplementary Figure 5**) –. A
sample is dropped if a) its average genome coverage is lower than 20; or b) multiple of its n
related sequencing metrics are of drop or fail samples based on this sample level QC process (Supplementary Figure 3) –. A
sample is dropped if a) its average genome coverage is lower than 20; or b) multiple of its
related sequencing metrics are of bad qua

applied uniformly across all samples, regardless of cohort or sequencing information. Variants related sequencing in the late of badden, the spiritual contamination check, or sex check; or 4) it is an unexpected duplicate. Alternatively, a sample is flagged if fail
in less than 3 sequencing metrics.
ADSP Variant QC in less than 3 sequencing metrics.
 ADSP Variant QC protocol

Different filtering and quality-checking strategies were applied at each level (genotype-, and

variant-) [62]. The QC protocol was applied on the bi-allelic **ADSP Variant QC protocol**
Different filtering and quality-chec
variant-) [62]. The QC protocol wa
applied uniformly across all samplof low quality have been flagged b
the datasets. <u>/</u>
| |
| c
| ADSP Variant QC protocol
Different filtering and quality-checking strategies were applied at each level (genotype-, and variant-) [62]. The QC protocol was applied on the bi-allelic autosomal VCFs. All QC flags were
applied uniformly across all samples, regardless of cohort or sequencing information. Variant
of low quality have been flagged applied uniformly across all samples, regardless of cohort or sequencing information. Variants
of low quality have been flagged but these variants have not been excluded and filtered out of
the datasets. Filow quality have been flagged but these variants have not been excluded and filtered out of
the datasets.
The datasets. the datasets.

- set to missing ("./.") if either or both read depth ("DP") was less than 10 (DP<10) or
genotype quality ("GQ") score was less than 20 (GQ<20). All these censored genotypes
were excluded from subsequent QC steps, except fo
- genotype quality ("GQ") score was less than 20 (GQ<20). All these censored genoty
were excluded from subsequent QC steps, except for estimation of variant-level ave
depth ("AverageReadDepth") in variant-level QC.
Variant-l genously are excluded from subsequent QC steps, except for estimation of variant-level average
depth ("AverageReadDepth") in variant-level QC.
Variant-level QC was applied to all variants. Flags were applied in the followi depth ("AverageReadDepth") in variant-level QC.
Variant-level QC was applied to all variants. Flags were applied in the following order: a)
Variants in GATK low sequence quality tranches [variants without a FILTER value of Mariant-level QC was applied to all variants. Flags
Variant-level QC was applied to all variants. Flags
Variants in GATK low sequence quality tranches [v
that are above the 99.8% VQSR Tranche]; b) Mono
Variants with high m
- Variants in GATK low sequence quality tranches [variants without a FILTER value of "PASS
that are above the 99.8% VQSR Tranche]; b) Monomorphic variants were flagged; c)
Variants with high missing rate were flagged; d) Var that are above the 99.8% VQSR Tranche]; b) Monomorphic variants were flagged; c)
Variants with high missing rate were flagged; d) Variants with high read depth were flagged
Variants with excessive heterozygosity or departu Variants with high missing rate were flagged; d) Variants with high read depth were t
Variants with excessive heterozygosity or departure from Hardy-Weinberg equilibriu
(HWE) were evaluated within race/ethnic subgroup, how Variants with excessive heterozygosity or departure from Hardy-Weinberg equilibrium
(HWE) were evaluated within race/ethnic subgroup, however given the complexity around
race/ethnic subgroups, these were not flagged though (HWE) were evaluated within race/ethnic subgroup, however given the complexity arous
race/ethnic subgroups, these were not flagged though the measures have been made
available and can be implemented as user-defined filters (HTM) were evaluated with the measures have been made

available and can be implemented as user-defined filters if desired. Similarly, ("ABhet") was

computed among uncensored heterozygotes at each variant and provided in race/ethnic subgroups, and a weak and a wallable and can be implemented as user-defined filters if desired. Similarly, ("ABhet"
computed among uncensored heterozygotes at each variant and provided in the files k
not applie

Comparison with gnomAD

The gnomAD resource (version 4) were download from

computed among uncensored heterozygotes at each variant and provided in the files but
not applied in any filtering criteria.
no applied in any filtering criteria.
gnomAD resource (version 4) were download from
strains bi-a not applied in any filtering criteria.

nparison with gnomAD

gnomAD resource (version 4) were download from

ns://gnomad.broadinstitute.org/downloads#v4. This contains bi-allelic variants from multi-

nic samples. To comp nparison with gnomAD
gnomAD resource (version 4) were
ss://gnomad.broadinstitute.org/dov
nic samples. To compare with the A
n the VCF. We then broke the VCFs
ins and HHB-AA. Monomorphic vari
acted variants from both the gn https://gnomad.broadinstitute.org/downloads#v4</u>. This
ethnic samples. To compare with the ADSP R4 data, we
from the VCF. We then broke the VCFs down into sampl
Asians and HHB-AA. Monomorphic variants were exclud
extracted from the VCF. We then broke the VCFs down into samples of four ancestry groups – NHW, H
Asians and HHB-AA. Monomorphic variants were excluded from each of the VCFs. We then
extracted variants from both the gnomAD and ADSP From the VCFs. We then extracted variants were excluded from each of the VCFs. We then extracted variants from both the gnomAD and ADSP R4 data by MAF thresholds: <=0.1%, <=0.5%, <=1% and <=5%, and compared them at the sit extracted variants from both the gnomAD and ADSP R4 data by MAF thresholds: <=0.1%,

<=0.5%, <=1% and <=5%, and compared them at the site level. When compared across ance

groups, only sites that are observed in all ancest

λ https://gnomat.broadination.org/downloads/variants/variants from multi-allelic variants from multi-allel

<=0.5%, <=1% and <=5%, and compared them at the site level. When compared across an
groups, only sites that are observed in all ancestries were used for analyses.
Variant annotation protocol
All R4 bi-allelic variants have groups, only sites that are observed in all ancestries were used for analyses.
 Variant annotation protocol

All R4 bi-allelic variants have been annotated using the official ADSP annotation pipeline. First,

the QCed VC Variant annotation protocol
All R4 bi-allelic variants have been annotated using the official ADSP annotat
the QCed VCFs were processed using VEP103 [63] (with the --everything flag
formatted VEP output was processed so th **Variant annotation protocol**
All R4 bi-allelic variants have l
the QCed VCFs were processe
formatted VEP output was pr
gene were collapsed generate
process uses the ranking table
damaging' consequence and i
process which The QCed VCFs were processed using VEP103 [63] (with the --everything flag). Then the JSON-
formatted VEP output was processed so that variants affecting multiple transcripts of the same
gene were collapsed generate a 'mos CADD v1.6 scores [28]. Lastly, short indels not defined in CADD reference files were processed gene were collapsed generate a 'most damaging' consequence for each affected gene. This
process uses the ranking table specified in the file 'ranking_table.txt' to identify the 'most
damaging' consequence and to assign an gene were collapsed generate a 'most damaging' consequence for each affected gene. This
process uses the ranking table specified in the file 'ranking_table.txt' to identify the 'most
damaging' consequence and to assign an process uses the ranking table specified in the file 'ranking_table.txt' to identify the 'most
damaging' consequence and to assign an impact score, using a custom annotation ranking
process which down-weights consequences damaging' consequence and to assign an impact score, using a custom annotation ranking
process which down-weights consequences for non-sense mediated decay transcripts and
coding transcripts. Next the QCed VCFs were also p [23]. Variants are matched by chromosome, position, reference allele, and alternate allele to
CADD v1.6 scores [28]. Lastly, short indels not defined in CADD reference files were processed
by CADD and integrated into the d CADD v1.6 scores [28]. Lastly, short indels not defined in CADD reference files were processed
by CADD and integrated into the dataset. This resource is available at NIAGADS open access
https://dss.niagads.org/open-access-

<u>LOT analyses</u>
The functional impact of variants was assessed using SnnEff [23] Variants were annotated as CADD and integrated into the dataset. This resource is available at NIAGADS open access

https://dss.niagads.org/open-access-data-portal/

LOF analyses

The functional impact of variants was assessed using SnpEff [23]. Var by CAD analysis, the data integrated into the data in the data of the data of the data of the functional impact of variants was assessed using SnpEff [23]. Variants were annotated as 'MODERATE' and 'HIGH' when they were pr **LOF analyses**
The functional impact of variants was assessed usin
'MODERATE' and 'HIGH' when they were protein-al'
'MODIFIER' effects were considered non-protein alt $\begin{array}{c}\n\blacksquare \\
\blacksquare \\
\blacksquare \\
\blacksquare\n\end{array}$ The functional impact of variants was assessed using SnpEff [23]. Variants were annotated as
'MODERATE' and 'HIGH' when they were protein-altering, while variants with 'LOW' and
'MODIFIER' effects were considered non-prote 'MODIFIER' effects were considered non-protein altering. All variants categorized as 'HIG'
'MODIFIER' effects were considered non-protein-altering. All variants categorized as 'HIG 'MODIFIER' effects were considered non-protein altering. All variants categorized as 'HIGH's the United States
'HIGH's categorized as 'HIGH's categorized as 'HIGH's categorized as 'HIGH's categorized as 'HIGH's categorize

'MODERATE' effect were expected to be missense and splice region variants. Variants with
'MODIFIER' and 'LOW' effects were in non-coding regions or were non-disruptive to protein
functions. We focused on the LoF variants c

'MODIFIER' and 'LOW' effects were in non-coding regions or were non-disruptive to proteir
functions. We focused on the LoF variants categorized as 'HIGH' in the ADSP cohort.
FAVOR annotation protocol
We downloaded the FA 'MODIFIER' and 'LOW' effects were in non-coding regions or were non-disruptive to protein
functions. We focused on the LoF variants categorized as 'HIGH' in the ADSP cohort.
FAVOR annotation protocol
We downloaded the FAVO FAVOR annotation protocol
We downloaded the FAVOR [29] database annotations from <u>https://docs.genohub.o</u>
July 2023 and used that to annotate all the R4 bi-allelic variants. Then annotation on
were converted to GDS format FAVOR annotation protocol
We downloaded the FAVOR [29] database annotations from https://docs.genohub.org/data in July 2023 and used that to annotate all the R4 bi-allelic variants. Then annotation only VCF files
were converted to GDS format using the SeqArray package [26], containing 156 columns. This
resource is available at NIAGADS

Structural variant calling protocol Structural variant calling protocol

July 2023 and NDS format using the SeqArray package [26], containing 156 columns. This
resource is available at NIAGADS open access https://dss.niagads.org/open-access-data-portal/.
Structural variant calling protocol
The resource is available at NIAGADS open access https://dss.niagads.org/open-access-data-portal
resource is available at NIAGADS open access https://dss.niagads.org/open-access-data-portal
Structural variant calling protocol
 Structural variant calling protocol
The GCAD and ADSP SV workgroup together designs the production pipeline which includes
Manta [64] (v1.6.0) and Smoove (ref, v0.2.6) (https://github.com/brentp/smoove) for calling
delet Manta [64] (v1.6.0) and Smoove (ref, v0.2.6) (https://github.com/brentp/smoove) for calling
deletions and insertions. Individual Manta and Smoove callsets were first merged for each
sample and merged together with all samp deletions and insertions. Individual Manta and Smoove callsets were first merged for each
sample and merged together with all samples by SVIMMER (v0.1)
(https://github.com/DecodeGenetics/svimmer) . Then, GraphTyper (v2.7) sample and merged together with all samples by SVIMMER (v0.1)

(https://github.com/DecodeGenetics/svimmer) . Then, GraphTyper (v2.7) [65] was applied

the merged VCF for SV joint genotyping. Note that the only filter appl (https://github.com/DecodeGenetics/svimmer) . Then, GraphType
the merged VCF for SV joint genotyping. Note that the only filter a
other than that, there is no advance filter. The code used to gener
at: https://github.com/I

(https://github.com/llumina/manta, https://github.com/brenticles. SV size >10 Mbp;
other than that, there is no advance filter. The code used to generate these SV calls is available
at: https://github.com/lllumina/manta, h other than that, there is no advance filter. The code used to generate these SV calls is available at: https://github.com/lllumina/manta, https://github.com/brentp/smoove).

LD reference panel

We inferred LD separately at: <u>https://github.com/Illumina/manta, https://github.com/brentp/smoove</u>).
LD reference panel
We inferred LD separately for each of the four major ancestral groups (NHW, His, AA, and
Asian). All participants for both c **LD reference panel**
We inferred LD separately for each of the four major ancestral groups (NHW, I
Asian). All participants for both cases and controls were included but PSP and
not included, resulting in 32,236 participan <u>||</u> \
| / r \
c r <u>LD reference panel</u>
We informed LD cane Asian). All participants for both cases and controls were included but PSP and CBD samples
not included, resulting in 32,236 participants total). We calculated LD for all pairs of varian
with minor allele count (MAC) \geq Assumption parallel participants for the calculated LD for all pairs of variants
not included, resulting in 32,236 participants total). We calculated LD for all pairs of variants
with minor allele count (MAC) \geq 5 and with minor allele count (MAC) \geq 5 and within 5 Mbps of each other using emerald [25] using
the following parameters, --mac > 5, --threshold 0.2, and --window 5000000. Each segment v
analyzed by 5Mb window with a 3Mb o which minds and the count (MAC) \pm 5 and MAC) and subspaces the valid of the following parameters, $\frac{1}{2}$ -mac $>$ 5, $\frac{1}{2}$ -threshold 0.2, and $\frac{1}{2}$ -window 5000000. Each segment wanalyzed by 5Mb window with a analyzed by 5Mb window with a 3Mb overlapping, then we concatenated all segments and
analyzed by 5Mb window with a 3Mb overlapping, then we concatenated all segments and
removed duplicate records. Only variant pairs with Framaly 2018 of the metallical properties and the removed duplicate records. Only variant pairs with $R^2 > 0.2$ were then retained. For each variant pair, we reported variant genomic positions, reference and non-reference removed duplicate records. Only variant pairs with R
variant pair, we reported variant genomic positions, r
R and R² correlation, and D and D' statistics. This reso
https://dss.niagads.org/open-access-data-portal/.
ADSP eference and non-reference alleles,
urce is available at NIAGADS open a
PHC)
SP-PHC) harmonizes all available dat
ure the highest quality harmonizatio
cinants with available whole-genom

ADSP Phenotype Harmonization Consortium (ADSP-PHC)

R and R² correlation, and D and D' statistics. This resource is available at NIAGADS open access

https://dss.niagads.org/open-access-data-portal/.

ADSP Phenotype Harmonization Consortium (ADSP-PHC)

The ADSP Phenotype K and K
<u>https://d</u>
<u>ADSP Ph</u>
The ADS
each dor
harmoni
sequenc
accessed Soliding and Soliding and Correlation Consortium (ADSP-PHC)

<u>Analytics. Analytics. Penning and Discussical portal</u>

Discriming the permonization Consortium (ADSP-PHC)

Discriming the permonization Consortium (ADSP-PHC)

T ADSP Phenotype Harmonization Consortium (ADS
The ADSP Phenotype Harmonization Consortium (ADS
each domain, regardless of sequencing status, to e
harmonized phenotypic data are then subset to pa
sequencing. These data are r The ADSP Phenotype Harmonization Consortium (ADSP-PHC) harmonizes all available data from each domain, regardless of sequencing status, to ensure the highest quality harmonization. The
harmonized phenotypic data are then subset to participants with available whole-genome
sequencing. These data are released per harmonized phenotypic data are then subset to participants with available whole-genome
sequencing. These data are released per participant via NIAGADS. All harmonized data can be
accessed directly from each cohort. All ADS manutation. These data are released per participant via NIAGADS. All harmonized data can
accessed directly from each cohort. All ADSP phenotype data are harmonized by a multi-
disciplinary team that includes world experts sexual directly from each cohort. All ADSP phenotype data are harmonized by a multi-
disciplinary team that includes world experts in neuroimaging, neuropsychology, fluid
biomarkers, neuropathology, and vascular contributi discrimary team to provide the sum of the state includes to the processing and specific harmonization protocols are available in **Supplementary Methods - ADSP Phe Harmonization Consortium Protocol**. These files are availab specific harmonization protocols are available in Supplementary Methods - ADSP Phenotype
Harmonization Consortium Protocol. These files are available in https://dss.niagads.org/. Harmonization Consortium Protocol. These files are available in $\frac{f(t, p, s, f)}{f(t, p, s, f)}$.

Varikam
This is an
WGS/WE
allows us
GRCh38.
MIAGADS
This is an WGS/WES data of the ADSP. The database currently includes variants of all the R4 36K WGS and
allows users to search for genes or variants of interests. The human reference genome used is
GRCh38. It is available here: https

NIAGADS Alzheimer's GenomicsDB

allows users to search for genes or variants of interests. The human reference genome used is
GRCh38. It is available here: https://varixam.niagads.org/.
NIAGADS Alzheimer's GenomicsDB
This is an interactive knowledgebase GRCh38. It is available here: <u>https://varixam.niagads.org/.</u>
 NIAGADS Alzheimer's GenomicsDB

This is an interactive knowledgebase for AD genetics [27]. The resource provides unrestricted

access to GWAS summary statist MIAGADS Alzheimer's GenomicsDB
This is an interactive knowledgebase for AD genetics [27].
access to GWAS summary statistics datasets, variant annot
deposited at the NIAGADS. The platform allows users to se
interests, and i This is an interactive knowledgebase for AD genetics [27]. The resource provides unrestricted The constrained and the NIAGADS. The platform allows users to search for genes or variants of
deposited at the NIAGADS. The platform allows users to search for genes or variants of
interests, and interactively mine or visu access to Gwarten the NIAGADS. The platform allows users to search for genes or variants of
interests, and interactively mine or visually inspect datasets and annotated ADSP variant tra
on a genome browser. The GenomicsDB interests, and interactively mine or visually inspect datasets and annotated ADSP varian
on a genome browser. The GenomicsDB can be accessed at
https://www.niagads.org/genomics.
NIAGADS DSS
The NIAGADS Data Sharing Service

NIAGADS DSS

on a genome browser. The GenomicsDB can be accessed at

https://www.niagads.org/genomics.

NIAGADS DSS

The NIAGADS Data Sharing Service (DSS) was developed to facilitate the deposition and sharing

of whole-genome and who https://www.niagads.org/genomics.

NIAGADS DSS

The NIAGADS Data Sharing Service (DSS) was developed to f

of whole-genome and whole-exome sequencing data from *l*

studies to the research community at large. In keeping wi MIAGADS DSS
The NIAGADS Data Sharing Service (D
of whole-genome and whole-exome
studies to the research community at
(GDS) Policy, all genomic data are cla
Certification forms provided by the su
DSS distributed data throug The NIAGADS Data Sharing Service (DSS) was developed to facilitate the deposition and sharing of whole-genome and whole-exome sequencing data from ADSP and other NIA funded ADRD
studies to the research community at large. In keeping with the NIH Genomic Data Sharing
(GDS) Policy, all genomic data are classified as studies to the research community at large. In keeping with the NIH Genomic Data Sharing (GDS) Policy, all genomic data are classified as controlled access as outlined in the Institutiona Certification forms provided by th (GDS) Policy, all genomic data are classified as controlled access as outlined in the Institutio
Certification forms provided by the submitting institutions. Principal investigators can reque
DSS distributed data through t Certification forms provided by the submitting institutions. Principal investigators can request
DSS distributed data through the Data Access Request Management (DARM) system by logging
in using their eRA Commons ID. Once DSS distributed data through the Data Access Request Management (DARM) system by logging
in using their eRA Commons ID. Once an application is approved by the NIH-formed NIAGADS
ADRD Data Access Committee (NADAC) and Data in using their eRA Commons ID. Once an application is approved by the NIH-formed NIAGADS
ADRD Data Access Committee (NADAC) and Data Use Committee (DUC), the data can be
accessed through the Data Portal and downloaded dire ADRD Data Access Committee (NADAC) and Data Use Committee (DUC), the data can be
accessed through the Data Portal and downloaded directly or through Amazon EC2. DSS can b
found at <u>https://dss.niagads.org/</u> .
Released geno

Released genotyping files Released genotyping files

accessed through the Data Portal and downloaded directly or through Amazon EC2. DSS
found at <u>https://dss.niagads.org/</u>
Released genotyping files
Due to the sheer sizes of pVCFs, all pVCFs are split by chromosomes. We prov found at <u>https://dss.niagads.org/</u>
 Released genotyping files

Due to the sheer sizes of pVCFs, all pVCFs are split by chromosomes. We provide three versions

of pVCFs for users to choose from: (1) "Preview pVCF"; (2) " Released genotyping files
Due to the sheer sizes of pVCFs, all
of pVCFs for users to choose from:
(genotype) of each sample for each
with replacing low-quality genotype
chromosome, then split into bi-alle
fully OCed VCF (o of pVCFs for users to choose from: (1) "Preview pVCF"; (2) "Compact pVCF": only reserved GT
(genotype) of each sample for each variant; (3) "Compact filtered pVCF": a compact version
with replacing low-quality genotypes to of pVCFs for users to choose from: (1) "Preview pVCF"; (2) "Compact pVCF": only reserved GT
(genotype) of each sample for each variant; (3) "Compact filtered pVCF": a compact version
with replacing low-quality genotypes to (genotype) of each sample for each sample in the results of purifical set of purific results are divided by
chromosome, then split into bi-allelic and multi-allelic variant files. Besides, we also created
fully QCed VCF (o chromosome, then split into bi-allelic and multi-allelic variant files. Besides, we also create
fully QCed VCF (output from the "ADSP Variant QC protocol") in Genomic Data Structure (
format to facilitate analysts using R Fully QCed VCF (output from the "ADSP Variant QC protocol") in Genomic Data Structure (GDS)
format to facilitate analysts using R for downstream association analyses. These files are
available in https://dss.niagads.org/.

Format to facilitate analysts using R for downstream association analyses. These files are
available in https://dss.niagads.org/.
Authors' contribution
YYL, W-PL, LAF, RPM, MAP-V, JLH, EM, ACN, XZ, BNV, BWK, WSB, GDS, an available in https://dss.niagads.org/
Authors' contribution
YYL, W-PL, LAF, RPM, MAP-V, JLH, EM, ACN, XZ, BNV, BWK, WSB, GDS, and L-SW concept
the study. CD, LAF, RPM, MAP-V, JLH, EM, AT, MC, TJH, GDS, and L-SW secures t **Authors' contribution**

YYL, W-PL, LAF, RPM, MAP-V, JLH, EM

the study. CD, LAF, RPM, MAP-V, JLH,

ABK, HN, HW, ZK, LC, PRM, CD, THJ, ai

the project. YYL, OV, PG, LQ, YZ, P-LC,

PL, PPK, TIG, CZ, HW, JJ, and WSB per ノート ちょうしょう しょうしょう YYL, W-PL, LAF, RPM, MAP-V, JLH, EM, ACN, XZ, BNV, BWK, WSB, GDS, and L-SW conceptualized the study. CD, LAF, RPM, MAP-V, JLH, EM, AT, MC, TJH, GDS, and L-SW secures the funding. YYL,
ABK, HN, HW, ZK, LC, PRM, CD, THJ, and BWK performed data management and coordination of
the project. YYL, OV, PG, LQ, YZ, P-LC, ABK, HN, HW, ZK, LC, PRM, CD, THJ, and BWK performed data management and coordination of
the project. YYL, OV, PG, LQ, YZ, P-LC, HN, LA, NS, TI, and NRW generated the resources. YYL, W-
PL, PPK, TIG, CZ, HW, JJ, and WSB pe the project. YYL, OV, PG, LQ, YZ, P-LC, HN, LA, NS, TI, and NRW generated the resources. YYL, W-
PL, PPK, TIG, CZ, HW, JJ, and WSB performed the analyses. YYL, W-PK, PPK, SLT, TIG, CAN, XZ,
BNV, TJH, HW, JJ, BWK, and WSB d PL, Press, P
BNV, TJH, HW, JJ, BWK, and WSB drafted the manuscript. BNV, TJH, HW, JJ, BWK, and WSB drafted the manuscript.

Acknowledgements

Data for this study were prepared, archived, and distributed by the National Institute on Aging Alzheimer's Disease Data Storage Site (NIAGADS) at the University of Pennsylvania (U24-AG041689), funded by the National Institute on Aging.

Alzheimer's Disease Sequencing Project (sa000001) data:

Alz**heimer's Disease Sequencing Project (sa000001) data:**
The Alzheimer's Disease Sequencing Project (ADSP) is com_l
genetics consortia and three National Human Genome Res genetics consortia and three National Human Genome Research Institute (NHGRI) funded Large S
Sequencing and Analysis Centers (LSAC). The two AD genetics consortia are the Alzheimer's Disea genetics construction and the National Human Genome Research Institute (NHSIII), military is Disease
Sequencing and Analysis Centers (LSAC). The two AD genetics consortia are the Alzheimer's Disease
Genetics Consortium (AD Sequencing and Analysis Centers (Later). The two Analysis Centers are the Cohorts for Heart and Aging
Genetics Consortium (ADGC) funded by NIA (U01 AG032984), and the Cohorts for Heart and Aging
Research in Genomic Epidemi Research in Genomic Epidemiology (CHARGE) funded by NIA (R01 AG033193), the National Heart, I
and Blood Institute (NHLBI), other National Institute of Health (NIH) institutes and other foreign and Blood Institute (NHLBI), other National Institute of Health (NIH) institutes and other foreign
governmental and non-governmental organizations. The Discovery Phase analysis of sequence data is governmental and non-governmental organizations. The Discovery Phase analysis of sequence data is supported through UF1AG047133 (to Drs. Schellenberg, Farrer, Pericak-Vance, Mayeux, and Haines); U01AG049505 to Dr. Seshadri; U01AG049506 to Dr. Boerwinkle; U01AG049507 to Dr. Wijsman; and U01AG049508 to Dr. Goate and the Discovery Extension Phase analysis is supported through U01AG049508 to Dr. Goate and the Discovery Extension Phase analysis is supported through
U01AG052411 to Dr. Goate, U01AG052410 to Dr. Pericak-Vance and U01 AG052409 to Drs. Seshadri U01AG052411 to Dr. Goate, U01AG052410 to Dr. Pericak-Vance and U01 AG052409 to Drs. S
Fornage United States, United State

Fornage.
Sequencing for the Follow Up Study (FUS) is supported through U01AG057659 (to Drs. PericakVance, Sequencing for the Forman Specialty (FUS) is supported through USD to Dramato (CD 2011 District through
Mayeux, and Vardarajan) and U01AG062943 (to Drs. Pericak-Vance and Mayeux). Data generation an
harmonization in the Fo Mayeux, and Vardarajan, and UD112012 (to Letter US4AG052427 (to Drs. Schellenberg and Wang).
harmonization in the Follow-up Phase is supported by U54AG052427 (to Drs. Schellenberg and Wang).
The FUS Phase analysis of seque harmonization in the Follow-up Phase is supported by The Follow-US-2022-2022.
The FUS Phase analysis of sequence data is supported through U01AG058589 (to Drs. Destefano,
Boerwinkle, De Jager, Fornage, Seshadri, and Wijsma The Fushell Champer Secrets and The Fugges and Difference data is supported to Drs. Haines, Bush, Farre
Boerwinkle, De Jager, Fornage, Seshadri, and Wijsman), U01AG058654 (to Drs. Haines, Bush, Farre
Martin, and Pericak-Va Martin, and Pericak-Vance), U01AG058635 (to Dr. Goate), RF1AG058066 (to Drs. Haines, Pericak-Vance, and Scott), RF1AG057519 (to Drs. Farrer and Jun), R01AG048927 (to Dr. Farrer), and RF1AG054074 (to Drs. Pericak-Vance and Beecham).

The ADGC cohorts include: Adult Changes in Thought (ACT) (U01 AG006781, U19 AG066567), the Alzheimer's Disease Research Centers (ADRC) (P30 AG062429, P30 AG066468, P30 AG062421, P30 AG066509, P30 AG066514, P30 AG066530, P30 AG066507, P30 AG066444, P30 AG066518, P30 AG066512, P30 AG066462, P30 AG072979, P30 AG072972, P30 AG072976, P30 AG072975, P30 AG072978, P30 AG072977, P30 AG066519, P30 AG062677, P30 AG079280, P30 AG062422, P30
AG066511, P30 AG072946, P30 AG062715, P30 AG072973, P30 AG066506, P30 AG066508, P30 AG066515, P30 AG072947, P30 AG072931, P30 AG066546, P20 AG068024, P20 AG068053, P20 ...
AG066515, P30 AG072947, P30 AG072931, P30 AG066546, P20 AG068024, P20 AG068053, P20
AG068077, P20 AG068082, P30 AG072958, P30 AG072959), the Chicago Health and Aging Proje AG068077, P20 AG068082, P30 AG072958, P30 AG072959), the Chicago Health and Aging Proje
(CHAP) (R01 AG11101, RC4 AG039085, K23 AG030944), Indiana Memory and Aging Study (IMAS (CHAP) (R01 AG11101, RC4 AG039085, K23 AG030944), Indiana Memory and Aging Study (IMAS) (R01 AG019771), Indianapolis Ibadan (R01 AG009956, P30 AG010133), the Memory and Aging Project (MAP) (
R01 AG17917), Mayo Clinic (MAYO) (R01 AG032990, U01 AG046139, R01 NS080820, RF1 AG051504, P50 AG019771), Indianapolis Ibadan (R01 AG009956, P30 AG010133), the Memory and Aging Project (MAP) (R01 AG17917), Mayo Clinic (MAYO) (R01 AG032990, U01 AG046139, R01 NS080820, RF1 AG051504, P50 Miami (R01 AG027944, R01 AG028786, R01 AG019085, IIRG09133827, A2011048), the Multi-
Institutional Research in Alzheimer's Genetic Epidemiology Study (MIRAGE) (R01 AG09029, R01 AG025259), the National Centralized Repository for Alzheimer's Disease and Related Dementias Institutional Research in Alexandria Process and Research in Alachemer's Genetic Epidemiology (MIRAGO25259), the National Centralized Repository for Alzheimer's Disease and Related Dementias
(NCRAD) (U24 AG021886), the Nat , the Hago (U24 AG021886), the National Institute on Aging Late Onset Alzheimer's Disease Family
(NIA- LOAD) (U24 AG056270), the Religious Orders Study (ROS) (P30 AG10161, R01 AG15819), th (NIA-LOAD) (U24 AG056270), the Religious Orders Study (ROS) (P30 AG10161, R01 AG15819), the Texas (NIA- LOAD) (U24 AG056270), the Religious Orders Study (ROS) (P30 AG10161, R01 AG10161, R01 AG15819), the Texas Study (ROS) (P30 AG10161, R01 AG10161, R01 AG15819), the Texas Study (ROS) (P30 AG15819), the Texas Study (RO

Alzheimer's Research and Care Consortium (TARCC) (funded by the Darrell K Royal Texas Alzheimer's
Initiative), Vanderbilt University/Case Western Reserve University (VAN/CWRU) (R01 AG019757, R01
AG021547, R01 AG027944, R01 Washington Heights-Inwood Columbia Aging Project (WHICAP) (RF1 AG054023), the University of Washington Heights-Inwood Columbia Aging Project (WHICAP) (RF1 AG054023), the Universit
Washington Families (VA Research Merit Grant, NIA: P50AG005136, R01AG041797, NINDS: Washington Families (VA Research Merit Grant, NIA: P50AG005136, R01AG041797, NINDS:
R01NS069719), the Columbia University Hispanic Estudio Familiar de Influencia Genetica de Alzhei Washington Families (Wither Brent, Merit Grant, Indian Brent, NIA: P50AG0069719), the Columbia University Hispanic Estudio Familiar de Influencia Genetica de
(EFIGA) (RF1 AG015473), the University of Toronto (UT) (funded b resided) (RF1 AG015473), the University Amperica Columbia Influencia Genetica delical Research
(EFIGA) (RF1 AG015473), the University of Toronto (UT) (funded by Wellcome Trust, Medical Research
Council, Canadian Institutes Council, Canadian Institutes of Health Research), and Genetic Differences (GD) (R01 AG007584). The CHARGE cohorts are supported in part by National Heart, Lung, and Blood Institute (NHLBI) CHARGE cohorts are supported in part by National Heart, Lung, and Blood Institute (NHLBI)
infrastructure grant HL105756 (Psaty), RC2HL102419 (Boerwinkle) and the neurology working group infrastructure grant HL105756 (Psaty), RC2HL102419 (Boerwinkle) and the neurology workii
supported by the National Institute on Aging (NIA) R01 grant AG033193. supported by the National Institute on Aging (NIA) R01 grant AG033193.

The CHARGE cohorts participating in the ADSP include the following: Austrian Stroke Prevention Study (ASPS), ASPS-Family study, and the Prospective Dementia Registry-Austria (ASPS/PRODEM-Aus), the lasmus
Atherosclerosis Risk in Communities (ARIC) Study, the Cardiovascular Health Study (CHS), the Erasmus ,
Atherosclerosis Risk in Communities (ARIC) Study, the Cardiovascular Health Study (CHS), the Erasmu
Rucphen Family Study (ERF), the Framingham Heart Study (FHS), and the Rotterdam Study (RS). ASPS Rucphen Family Study (ERF), the Framingham Heart Study (FHS), and the Rotterdam Study (RS). ASPS is
funded by the Austrian Science Fond (FWF) grant number P20545-P05 and P13180 and the Medical funded by the Austrian Science Fond (FWF) grant number P20545-P05 and P13180 and the Medical
University of Graz. The ASPS-Fam is funded by the Austrian Science Fund (FWF) project 1904), the EU funded by the Austrian Science Forms, programmed by the Austrian Science Fund (FWF) project 1904), the EU
University of Graz. The ASPS-Fam is funded by the Austrian Science Fund (FWF) project 1904), the EU
Joint Programme University of Grazing Craz. The Austrian Science Funded by the Austrian Science Fund (FWF) project project
Joint Programme – Neurodegenerative Disease Research (JPND) in frame of the BRIDGET project
(Austria, Ministry of S (Austria, Ministry of Science) and the Medical University of Graz and the Steiermärkische
Krankenanstalten Gesellschaft. PRODEM-Austria is supported by the Austrian Research Promotio ,
Krankenanstalten Gesellschaft. PRODEM-Austria is supported by the Austrian Research Pr
agency (FFG) (Project No. 827462) and by the Austrian National Bank (Anniversary Fund, _I agency (FFG) (Project No. 827462) and by the Austrian National Bank (Anniversary Fund, project 15435.
ARIC research is carried out as a collaborative study supported by NHLBI contracts agency (FFG) (Project No. 827–1932) and by the Austrian National Bank (Anniversary Fund, project 25–1633)
ARIC research is carried out as a collaborative study supported by NHLBI contracts
(HHSN268201100005C, HHSN268201100 ARIC research is carried out as a collaborative study supported by NHLBI contracts HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C).
Neurocognitive data in ARIC is collected by U01 2U01HL096812, 2U01HL096814, 2U01HL096899, 2U01HL096902, 2U01HL096917 from the NIH (NHLBI, NINDS, NIA and NIDCD), and with previous brain MRI examinations funded by R01-HL70825 from the NHLBI. CHS research was supported by contracts MRI examinations funded by R01-HL70825 from the NHLBI. CHS research was supported by contracts
HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC8508
N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the NHLI N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the NHLBI with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Noth additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS).
Additional support was provided by R01AG023629, R01AG15928, and R01AG20098 from the NIA. FHS Additional support was provided by R01AG023629, R01AG15928, and R01AG20098 from the NIA. FHS
research is supported by NHLBI contracts N01-HC-25195 and HHSN268201500001I. This study was also research is supported by NHLBI contracts N01-HC-25195 and HHSN268201500001I. This study was also
supported by additional grants from the NIA (R01s AG054076, AG049607 and AG033040 and NINDS supported by additional grants from the NIA (R01s AG054076, AG049607 and AG033040 and NINDS
(R01 NS017950). The ERF study as a part of EUROSPAN (European Special Populations Research supported by additional grants from the NIA (NIADSPAN (European Special Populations Research
(R01 NS017950). The ERF study as a part of EUROSPAN (European Special Populations Research
Network) was supported by European Com (R01 NS01969). The ERF study as a part of EUROSPAN (ERP) pand power of B9947 (LSHG-CT-20
Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-20
01947) and also received funding from the Euro ork) was supported by European Community's Seventh Framework Programme
(FP7/2007-2013)/grant agreement HEALTH-F4- 2007-201413 by the European Commission under th) (FP7/2007-2013)/grant agreement HEALTH-F4- 2007-201413 by the European Commission under the the Spangist Rep
1997- programme "Quality of Life and Management of the Living Resources" of 5th Framework Programme ,
programme "Quality of Life and Management of the Living Resources" of 5th Framework Programme
(no. QLG2-CT-2002- 01254). High-throughput analysis of the ERF data was supported by a joint grant programme "Change of Life" and Management of the Life and Management of the Life and Management (no. QLG2-CT-2002- 01254). High-throughput analysis of the ERF data was supported by a joint grant
from the Netherlands Organi (no. Queen Concept and Concept and Concept analysis of the ERF data was supported by a june grant
from the Netherlands Organization for Scientific Research and the Russian Foundation for Basic
Research (NWO-RFBR 047.017.04 Research (NWO-RFBR 047.017.043). The Rotterdam Study is funded by Erasmus Medical Center and
Erasmus University, Rotterdam, the Netherlands Organization for Health Research and Development Erasmus University, Rotterdam, the Netherlands Organization for Health Research and Development
(ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture IZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Sci
Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the (Zondin), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Diseases in Santa Pana
Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the
municipality of Ro Science, the Ministry for Health, Welfare, the Sports, the European Commission (DC 2001), and the
Scientific Research NWO Investments (175.010.2005.011, 911-03-012), the Genetic Laboratory of Scientific Research NWO Investments (175.010.2005.011, 911-03-012), the Genetic Laboratory of the Scientific Research NWO Investments (175.010.2005.010.2005.011.2005.011, 911-03-012), the Genetic Laboratory o
The Genetic Laboratory of the Genetic Laboratory of the Genetic Laboratory of the Genetic Laboratory of the Ge

Department of Internal Medicine, Erasmus MC, the Research Institute for Diseases in the Elderly (014-
93-015; RIDE2), and the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA), project 050-060-810. All studies are
grateful to their participants, faculty and staff. The content of these manuscripts is solely the grateful to their participants, faculty and staff. The content of these manuscripts is solely the
responsibility of the authors and does not necessarily represent the official views of the National .
responsibility of the authors and does not necessarily represent the official views of the Natic
Institutes of Health or the U.S. Department of Health and Human Services. Institutes of Health or the U.S. Department of Health and Human Services.

The FUS cohorts include: the Alzheimer's Disease Research Centers (ADRC) (P30 AG062429, P30 AG066468, P30 AG062421, P30 AG066509, P30 AG066514, P30 AG066530, P30 AG066507, P30 AG066444, P30 AG066518, P30 AG066512, P30 AG066462, P30 AG072979, P30 AG072972, P30
AG072976, P30 AG072975, P30 AG072978, P30 AG072977, P30 AG066519, P30 AG062677, P30 AG072976, P30 AG072975, P30 AG072978, P30 AG072977, P30 AG066519, P30 AG062677, P30
AG079280, P30 AG062422, P30 AG066511, P30 AG072946, P30 AG062715, P30 AG072973, P30 AG072976, P30 AG072975, P30 AG072978, P30 AG072977, P30 AG066519, P30 AG062677, P30 AG079280, P30 AG062422, P30 AG066511, P30 AG072946, P30 AG062715, P30 AG072973, P30 ...
AG068024, P20 AG068053, P20 AG068077, P20 AG068082, P30 AG072958, P30 AG072959), Alzł
Disease Neuroimaging Initiative (ADNI) (U19AG024904), Amish Protective Variant Study (RF1AG Disease Neuroimaging Initiative (ADNI) (U19AG024904), Amish Protective Variant Study (RF1AG058066), Cache County Study (R01AG11380, R01AG031272, R01AG21136, RF1AG054052), Case Western Reserve
University Brain Bank (CWRUBB) (P50AG008012), Case Western Reserve University Rapid Decline (CWRURD) (RF1AG058267, NU38CK000480), CubanAmerican Alzheimer's Disease Initiative (CuAADI) University Brain Bank (CWRUBB) (P50AG008012), Case Western Reserve University Rapid Decline (CUOLAG052410), Estudio Familiar de Influencia Genetica en Alzheimer (EFIGA) (5R37AG015473,
(3U01AG052410), Estudio Familiar de Influencia Genetica en Alzheimer (EFIGA) (5R37AG015473,
RF1AG015473, R56AG051876), Genetic and .
RF1AG015473, RS6AG051876), Genetic and Environmental Risk Factors for Alzheimer Disease Am
African Americans Study (GenerAAtions) (2R01AG09029, R01AG025259, 2R01AG048927), Gwang African Americans Study (GenerAAtions) (2R01AG09029, R01AG025259, 2R01AG048927), Gwangju
Alzheimer and Related Dementias Study (GARD) (U01AG062602), Hillblom Aging Network (2014-A-004 African Americans Study (Senemaation) (2002–002–0025029), Hillblom Aging Network (2014-A-G
Alzheimer and Related Dementias Study (GARD) (U01AG062602), Hillblom Aging Network (2014-A-G
NET, R01AG032289, R01AG048234), Hussma Alzheimer and Relations Studies Study (Gamer, Cardinal Press Study American Gardinal Cardinal Press).
NET, R01AG032289, R01AG048234), Hussman Institute for Human Genomics Brain Bank (HIHGBB)
(R01AG027944, Alzheimer's Assoc (R01AG027944, Alzheimer's Association "Identification of Rare Variants in Alzheimer Disease"), Ibadan Study of Aging (IBADAN) (5R01AG009956), Longevity Genes Project (LGP) and LonGenity (R01AG042188, R01AG044829, R01AG046949, R01AG057909, R01AG061155, P30AG038072), Mexican Health and Aging
Study (MHAS) (R01AG018016), Multi-Institutional Research in Alzheimer's Genetic Epidemiology (MIRAGE) (2R01AG09029, R01AG025259, 2R01AG048927), Northern Manhattan Study (NOMAS) Study (MHAS) (R01AG018016), Multi-Institutional Research in Alzheimer's Genetic Epidemiology ,
(R01NS29993), Peru Alzheimer's Disease Initiative (PeADI) (RF1AG054074), Puerto Rican 1066 (PF
(Wellcome Trust (GR066133/GR080002), European Research Council (340755)), Puerto Rican Alzh (Wellcome Trust (GR066133/GR080002), European Research Council (340755)), Puerto Rican Alzheimer
Disease Initiative (PRADI) (RF1AG054074), Reasons for Geographic and Racial Differences in Stroke (Wellcome Trust (CRADI) (RF1AG054074), European Research Council (310767615161516151615161516151615)
Disease Initiative (PRADI) (RF1AG054074), Reasons for Geographic and Racial Differences in Stroke
(REGARDS) (U01NS041588) (REGARDS) (U01NS041588), Research in African American Alzheimer Disease Initiative (REAAADI)
(U01AG052410), the Religious Orders Study (ROS) (P30 AG10161, P30 AG72975, R01 AG15819, R01 (U01AG052410), the Religious Orders Study (ROS) (P30 AG10161, P30 AG72975, R01 AG15819, R01 AG42210), the RUSH Memory and Aging Project (MAP) (R01 AG017917, R01 AG42210Stanford Extreme Phenotypes in AD (R01AG060747), University of Miami Brain Endowment Bank (MBB), University of Miami/Case Western/North Carolina A&T African American (UM/CASE/NCAT) (U01AG052410, Phenotypes in AD (R01AG1), University of Millen and Prench (MM/CASE/NCAT) (U01AG052410,
Miami/Case Western/North Carolina A&T African American (UM/CASE/NCAT) (U01AG052410,
R01AG028786), and Wisconsin Registry for Alzheimer R01AG028786), and Wisconsin Registry for Alzheimer's Prevention (WRAP) (R01AG027161 and
R01AG054047). R01AG054047), and Wisconsin Registry for Alzheimer's Prevention (WRAP) (R01AG054047). R01AG054047).
The four LSACs are: the Human Genome Sequencing Center at the Baylor College of Medicine (U54

The four Later are: the Four Later are: the Human Genome Center at the Good 273), the Broad Institute Genome Center at the Baylor Center at the Broad Institute Genome Center at the Broad Institute Genome Center at the Broa Hood 22, 2021, the Broad Interview of the Health Sciences (U01AG057659), and the Washington University
Uniformed Services University of the Health Sciences (U01AG057659), and the Washington University
Genome Institute (U54 Genome Institute (U54HG003079). Genotyping and sequencing for the ADSP FUS is also conducted at
John P. Hussman Institute for Human Genomics (HIHG) Center for Genome Technology (CGT). John P. Hussman Institute for Human Genomics (HIHG) Center for Genome Technology (CGT).
John P. Hussman Institute for Human Genomics (HIHG) Center for Genome Technology (CGT).

Biological samples and associated phenotypic data used in primary data analyses were stored at Study Investigators institutions, and at the National Centralized Repository for Alzheimer's Disease and Investigators institutions, and at the National Centralized Repository for Alzheimer's Disease and

Related Dementias (NCRAD, CORAD, Dementias Corporation Democratic Phenotypics Corporation Democratic Phenote
Data used in primary and secondary data analyses were provided by Study Investigators, the NIA funded
Alzheimer's Data used in primary and secondary and secondary secondary terms in providing the University Orion Primary of
Alzheimer's Disease Centers (ADCs), and the National Alzheimer's Coordinating Center (NACC,
U24AG072122) and the U24AG072122) and the National Institute on Aging Genetics of Alzheimer's Disease Data Stora
(NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA. Harmonized phen (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA. Harmonized phenotypes
were provided by the ADSP Phenotype Harmonization Consortium (ADSP-PHC), funded by NIA (U24 (MITHER) CHRISTER (NIMED MINITEM) at the University of Pennsylvania, funded by NIA (U24
Were provided by the ADSP Phenotype Harmonization Consortium (ADSP-PHC), funded by NIA (U24
AG074855, U01 AG068057 and R01 AG059716) a were provided by the ADSP Phenotype Harmonization Consortium (ADSP) harmonized phenotype AGO
AGO74855, U01 AG068057 and R01 AG059716) and Ultrascale Machine Learning to Empower Discov
in Alzheimer's Disease Biobanks (Al4AD AGD AGD FIGURE AGO FIGURE AGO 6857. IN THE RESERVED THE RESERVED BY THE POST OF LATER IN THE LEATH IS NOT THE
Intramural Research Program of the National Institutes of health, National Library of Medicine. Intramural Research Program of the National Institutes of health, National Library of Medicine.
Contributors to the Genetic Analysis Data included Study Investigators on projects that were individu Contributors to the Genetic Analysis Data included Study Investigators on projects that were individually funded by NIA, and other NIH institutes, and by private U.S. organizations, or foreign governmental or nongovernmental organizations.

The ADSP Phenotype Harmonization Consortium (ADSP-PHC) is funded by NIA (U24 AG074855, U01 AG068057 and R01 AG059716). The harmonized cohorts within the ADSP-PHC include Dthe Anti-Amyloid
Treatment in Asymptomatic Alzheimer's study (A4 Study), a secondary prevention trial in preclinical Alzheimer's disease, aiming to slow cognitive decline associated with brain amyloid accumulation in clinically normal older participants. The A4 Study is funded by a public-private-philanthropic partnership, including funding from the National Institutes of Health-National Institute on Aging, Eli Lilly and Company, Alzheimer's Association, Accelerating Medicines Partnership, GHR Foundation, an anonymous Company, Alzheimer's Association, Accelerating Medicines Partnership, GHR Foundation, an anonymous
foundation and additional private donors, with in-kind support from Avid and Cogstate. The companion
observational Longitud foundation and additional private donors, with in-tional prepart from Avid and Condom Avid Consponse.
Observational Longitudinal Evaluation of Amyloid Risk and Neurodegeneration (LEARN) Study is funded
by the Alzheimer's A observational Longitudinal Longitudinal Congression, Controlled Risk and LEARN Studies are led by Dr. Reisa
by the Alzheimer's Association and GHR Foundation. The A4 and LEARN Studies are led by Dr. Reisa
Sperling at Brigh Sperling at Brigham and Women's Hospital, Harvard Medical School and Dr. Paul Aisen at the Alzheimer's Therapeutic Research Institute (ATRI), University of Southern California. The A4 and LEARN
Studies are coordinated by ATRI at the University of Southern California, and the data are made available through the Laboratory for Neuro Imaging at the University of Southern California. The available through the Laboratory for Neuro Imaging at the University of Southern California. The
participants screening for the A4 Study provided permission to share their de-identified data in o participants screening for the A4 Study provided permission to share their de-identified data in o
advance the quest to find a successful treatment for Alzheimer's disease. We would like to ackno participants of the quest to find a successful treatment for Alzheimer's disease. We would like to acknowledge
the dedication of all the participants, the site personnel, and all of the partnership team members who
continu the dedication of all the participants, the site personnel, and all the partnership team list is available
continue to make the A4 and LEARN Studies possible. The complete A4 Study Team list is available
on: <u>a4study.org/a</u> on: <u>a4study.org/a4-study-team.</u>; the Adult Changes in Thought study (ACT), U01 AG006781, U19
AG066567; Alzheimer's Disease Neuroimaging Initiative (ADNI): Data collection and sharing for this AG066567; Alzheimer's Disease Neuroimaging Initiative (ADNI): Data collection and sharing for this
project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of
Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2
0012). ADNI is funded by the Nati Health Grant Grant Grant Grant Grant Universe and Donald Institute of Biomedical Imaging
1991). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging
1991 and Bioengineering, and t and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's
Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristoland Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's
Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bris
Myers Squibb Company; CereS Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-
Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and
Company; EuroImm Myers Squibb Company; Cerespir, Inc.; Cepture; Eisai Inc.; Cummerceutical; Inc.; Eli Lili, Inc.; Fujir
Company; Eurolmmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujir
GE Healthcare; IXICO Ltd Company; Euro Immun; F. Hoffmann-La Roche Ltd and its affiliated company; Forthwest, Inc.; Fujirebio;
GE Healthcare; IXICO Ltd.;Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson &
Johnson Pharmaceutical Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.;Meso
Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals
Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Sorporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transit
Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI cli Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and

Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the
University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at
the University of Southe (EFIGA): 5R37AG015473, RF1AG015473, R56AG051876; Memory & Aging Project at Knight Alzheimer's (EFIGA): 5R37AG015473, RF1AG015473, R56AG051876; Memory & Aging Project at Knight
Disease Research Center (MAP at Knight ADRC): The Memory and Aging Project at the Knig .
Disease Research Center (MAP at Knight ADRC): The Memory and Aging Project at the Knight-ADRC
(Knight-ADRC). This work was supported by the National Institutes of Health (NIH) grants R01AG064614 - Disease Martin Center (Mark Center Center (March Center Center): The Memory and Aging Project at the Memory
(Knight-ADRC). This work was supported by the National Institutes of Health (NIH) grants R01AG064
R01AG044546, R (Knight-Adria). This work was supported by the National Institutes of Adria). This was supported by the National Institutes of Health (NIH) grants RD1AG064877 to Carlos Cruchaga. The recruitment and clinical characterizati recruitment and clinical characterization of research participants at Washington University was
supported by NIH grants P30AG066444, P01AG03991, and P01AG026276. Data collection and sharing for this project was supported by NIH grants RF1AG054080, P30AG066462, R01AG064614 and U01AG052410. We thank the contributors who collected samples used in this study, as well as patients for the project was supported by NIH grants III and to support the supported by NIH 110.
U01AG052410. We thank the contributors who collected samples used in this study, as well as
and their families, whose help and partic The contributors whose help and participation made this work possible. This work was supported by
access to equipment made possible by the Hope Center for Neurological Disorders, the Neurogenomics access to equipment made possible by the Hope Center for Neurological Disorders, the Neurogenomics and Informatics Center (NGI: https://neurogenomics.wustl.edu/) and the Departments of Neurology and
Psychiatry at Washington University School of Medicine; National Alzheimer's Coordinating Center p
And Informative Center (Nashington University School of Medicine; National Alzheimer's Coordinating Center
MACC): The NACC database is funded by NIA/NIH Grant U24 AG072122. NACC data are contributed by Psychiatry at Washington University School of Medicine; National Albert October 2014, 2016;
(NACC): The NACC database is funded by NIA/NIH Grant U24 AG072122. NACC data are contributed
the NIA-funded ADRCs: P30 AG062429 (P (MPPP), MPP MPP PROS: P30 AG062429 (PI James Brewer, MD, PhD), P30 AG066468 (PI Oscar Lopez,
the NIA-funded ADRCs: P30 AG062429 (PI James Brewer, MD, PhD), P30 AG066468 (PI Oscar Lopez,
MD), P30 AG062421 (PI Bradley Hyman, the NIA-funded ADRCs: P30 AG062429 (PI James Brewer, MD, PhD), P30 AG066468 (PI Oscar Lopez, AG066514 (PI Mary Sano, PhD), P30 AG066530 (PI Helena Chui, MD), P30 AG066507 (PI Marilyn Albert,
PhD), P30 AG066444 (PI John Morris, MD), P30 AG066518 (PI Jeffrey Kaye, MD), P30 AG066512 (PI PhD), P30 AG066444 (PI John Morris, MD), P30 AG066518 (PI Jeffrey Kaye, MD), P30 AG066512 (PI
Thomas Wisniewski, MD), P30 AG066462 (PI Scott Small, MD), P30 AG072979 (PI David Wolk, MD), P30 PhD), P30 AG066444 (PI John Morris, MD), P30 AG066518 (PI Jeffrey Kaye, MD), P30 AG066512 (PI AG072972 (PI Charles DeCarli, MD), P30 AG072976 (PI Andrew Saykin, PsyD), P30 AG072975 (PI David
Bennett, MD), P30 AG072978 (PI Neil Kowall, MD), P30 AG072977 (PI Robert Vassar, PhD), P30 AG066519 (PI Frank LaFerla, PhD), P30 AG062677 (PI Ronald Petersen, MD, PhD), P30 AG079280 (PI Eric Bennett, MD), P30 AG072978 (PI Neil Kowall, MD), P30 AG072977 (PI Robert Vassar, PhD), P30 AG066519 (PI Frank LaFerla, PhD), P30 AG062677 (PI Ronald Petersen, MD, PhD), P30 AG079280 (PI Eric AG072946 (PI Linda Van Eldik, PhD), P30 AG062715 (PI Sanjay Asthana, MD, FRCP), P30 AG072973 (PI
Russell Swerdlow, MD), P30 AG066506 (PI Todd Golde, MD, PhD), P30 AG066508 (PI Stephen Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suzanne Craft, Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suza
PhD), P30 AG072931 (PI Henry Paulson, MD, PhD), P30 AG066546 (PI Sudha Seshadri, MD), P2 Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suzanne Craft, AG068024 (PI Erik Roberson, MD, PhD), P20 AG068053 (PI Justin Miller, PhD), P20 AG068077 (PI Gary
Rosenberg, MD), P20 AG068082 (PI Angela Jefferson, PhD), P30 AG072958 (PI Heather Whitson, MD), Rosenberg, MD), P20 AG068082 (PI Angela Jefferson, PhD), P30 AG072958 (PI Heather Whitson, MD),
P30 AG072959 (PI James Leverenz, MD); National Institute on Aging Alzheimer's Disease Family Based P30 AG072959 (PI James Leverenz, MD); National Institute on Aging Alzheimer's Disease Family Based
Study (NIA-AD FBS): U24 AG056270; Religious Orders Study (ROS): P30AG10161,R01AG15819, Study (NIA-AD FBS): U24 AG056270; Religious Orders Study (ROS): P30AG10161,R01AG15819,
R01AG42210; Memory and Aging Project (MAP - Rush): R01AG017917, R01AG42210; Minority Aging study (NIC-AD From the From Project (MAP - Rush): R01AG017917, R01AG42210; Minority
Ro1AG42210; Memory and Aging Project (MAP - Rush): R01AG017917, R01AG42210; Minority
Research Study (MARS): R01AG22018, R01AG42210; Washin Research Study (MARS): R01AG22018, R01AG42210; Washington Heights/Inwood Columbia Aging
Project (WHICAP): RF1 AG054023;and Wisconsin Registry for Alzheimer's Prevention Project (WHICAP): RF1 AG054023;and Wisconsin Registry for Alzheimer's Prevention
(WRAP): R01AG027161 and R01AG054047. Additional acknowledgments include the National Instit (WRAP): R01AG027161 and R01AG054047. Additional acknowledgments include the
on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS, U24AG041689 on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA.

Alzheimer's Disease Neuroimaging Initiative (sa000002) data:

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan

Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.;
Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda
Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is Technologies; Novartis Pharmaceuticals Corporation; Pharmaceutical Corporation; Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research
providing funds to support ADNI clinical sites providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by
the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the providing funds to support ADNI clinical sites of Health (www.fnih.org). The grantee organization is the
the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the
Northern Califor Northern California Institute for Research and Education, and the study is coordinated by the
Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Additional information to include in an acknowledgment statement can be found on the LONI site: https://adni.loni.usc.edu/wp-content/uploads/how to apply/ADNI Data Use Agreement.pdf.

Alzheimer's Disease Genetics Consortium (sa000003) data:

Use the following for use of any ADGC generated data:

The Alzheimer's Disease Genetics Consortium (ADGC) supported sample preparation, sequencing and The Alzheimer's Disease Disease Consortium (ADDC) supported sample preparation, sequencing and
data processing through NIA grant U01AG032984. Sequencing data generation and harmonization is
supported by the Genome Center f data processing through through the Carl Peterson Processing and generation and harmonization is
supported by the Genome Center for Alzheimer's Disease, U54AG052427, and data sharing is suppor
by NIAGADS, U24AG041689. Samp supported by the Genome Center formation of the Genome Center for Alzheimer's Disease,
by NIAGADS, U24AG041689. Samples from the National Centralized Repository for Alzheimer's Disease,
and Related Dementias (NCRAD), which , ,
and Related Dementias (NCRAD), which receives government support under a cooperative agreement
grant (U24 AG021886) awarded by the National Institute on Aging (NIA), were used in this study. We grant (U24 AG021886) awarded by the National Institute on Aging (NIA), were used in this study. We thank contributors who collected samples used in this study, as well as patients and their families, whose help and participation made this work possible.

For use with the ADGC-TARCC-WGS ($\frac{\text{snd10030}}{\text{data}}$) data:

This study was made possible by the Texas Alzheimer's Research and Care Consortium (TARCC) funded by the state of Texas through the Texas Council on Alzheimer's Disease and Related Disorders and the Darrell K Royal Texas Alzheimer's Initiative.

The Familial Alzheimer Sequencing Project (sa000004) data:

This work was supported by grants from the National Institute
RF1AG053303. R01AG058501. U01AG058922. RF1AG058501 RF1AG053303, R01AG058501, U01AG058922, RF1AG058501 and R01AG057777). The recruitment and
clinical characterization of research participants at Washington University were supported by NIH P50 clinical characterization of research participants at Washington University were supported by NIH P50
AG05681, P01 AG03991, and P01 AG026276. This work was supported by access to equipment made AG05681, P01 AG03991, and P01 AG026276. This work was supported by access to equipment made
possible by the Hope Center for Neurological Disorders, and the Departments of Neurology and possible by the Hope Center for Neurological Disorders, and the Departments of Neurology and Psychiatry at Washington University School of Medicine.

We thank the contributors who collected samples used in this study, as well as patients and their families, whose help and participation made this work possible. This work was supported by access to families, whose help and participation made this work possible. This work was supported by access to

equipment made possible by the Hope Center for Neurological Disorders, and the Departments of
Neurology and Psychiatry at Washington University School of Medicine

Charles F. and Joanne Knight Alzheimer's Disease Research Center (sa000008) data:

This work was supported by grants from the National Institutes of Health (R01AG044546, P01AG003991, RF1AG058501, U01AG058922, RF1AG058501, and R01AG058501, and Mathamagnetic and RF1AG053303, R01AG058501, U01AG058922, RF1AG058501 and R01AG057777). The recruitment and
clinical characterization of research participants at Washington University were supported by NIH P50 clinical characterization of research participants at Washington University were supported by NIH P50
AG05681, P01 AG03991, and P01 AG026276. This work was supported by access to equipment made clinical characterization of research participants at Washington University Washington participants at AG05681,
AG05681, P01 AG03991, and P01 AG026276. This work was supported by access to equipment made
possible by the Ho AGOSSIDE BY THE FIGURE CHAGOSTIC PROTOCOLS INTO THE UNDER THE POLITION OF PRINCIPLE IN AGN.
Possible by the Hope Center for Neurological Disorders, and the Departments of Neurology and
Psychiatry at Washington University S Psychiatry at Washington University School of Medicine.

We thank the contributors who collected samples used in this study, as well as patients and their families, whose help and participation made this work possible. This work was supported by access to equipment made possible by the Hope Center for Neurological Disorders, and the Departments of Neurology and Psychiatry at Washington University School of Medicine.

For use of the ADSP-PHC harmonized phenotypes deposited within dataset, ng00067, use the following $\mathsf{Statement:}\quad\mathsf{Output}$

statement:
The Memory and Aging Project at the Knight-ADRC (Knight-ADRC), supported by NIH grants R01AG064614, R01AG044546, RF1AG053303, RF1AG058501, U01AG058922 and R01AG064877 to Carlos Cruchaga. The recruitment and clinical characterization of research participants at Washington University was supported by NIH grants P30AG066444, P01AG03991, and P01AG026276. Data collection
and sharing for this project was supported by NIH grants RF1AG054080, P30AG066462, R01AG064614 and sharing for this project was supported by NIH grants RF1AG054080, P30AG066462, R01AG064614
and U01AG052410. This work was supported by access to equipment made possible by the Hope Center and U01AG052410. This work was supported by access to equipment made possible by the Hope Cente
for Neurological Disorders, the Neurogenomics and Informatics Center and Deurological Disorders, the Neurogenomics and Informatics Center
for Neurological Disorders, the Neurogenomics and Informatics Center
(NGI: <u>https://neurogenomics.wustl.edu/</u>) and the Departments of Neurology and Psych for the Neurogenomics with Neurogenomic and Informatic Center
(NGI: <u>https://neurogenomics.wustl.edu/</u>) and the Departments of Neur
Washington University School of Medicine. Washington University School of Medicine.

Accelerating Medicines Partnership-Alzheimer's Disease (AMP-AD) (sa000011) data:

Mayo RNAseq Study- Study data were provided by the following sources: The Mayo Clinic Alzheimer's Disease Genetic Studies, led by Dr. Nilufer Ertekin-Taner and Dr. Steven G. Younkin, Mayo Clinic, **Mayo RNAseq Study-** Study data were provided by the following sources: The Mayo Clinic Alzheimer's
Disease Genetic Studies, led by Dr. Nilufer Ertekin-Taner and Dr. Steven G. Younkin, Mayo Clinic,
Jacksonville, FL using s Disease Cenetic Studies, 2008, 2008, 2008, 2008, 2008, 2008, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2
Discussion Steven Genetic Steven Grand Dr. Steven G. Youngin, Mayo Clinic, Alzheimer's D. Research Center, and Research Center, and the Mayo Clinic Brain Bank. Data collection was supported through funding by NIA grants P50 AG016574, R01 AG032990, U01 AG046139, R01 AG018023, U01 AG006576, U01 AG006786,
R01 AG025711, R01 AG017216, R01 AG003949, NINDS grant R01 NS080820, CurePSP Foundation, and support from Mayo Foundation. Study data includes samples collected through the Sun Health Research support from Mayo Foundation. Study data includes samples collected through the Sun Health Researc
Institute Brain and Body Donation Program of Sun City, Arizona. The Brain and Body Donation Program support from Mayo Foundation. The Mayo Foundation. Study to the Sun Health Research Marson Program.
Institute Brain and Body Donation Program of Sun City, Arizona. The Brain and Body Donation Program
is supported by the Na Institute Brain and Brain and Body Donation Program of Sun City, Arizona. The Brain and Brain Body Donational
Is supported by the National Institute of Neurological Disorders and Stroke (U24 NS072026 National
Brain and Tis is in and Tissue Resource for Parkinson's Disease and Related Disorders), the National Institute on
Aging (P30 AG19610 Arizona Alzheimer's Disease Core Center), the Arizona Department of Health Brain and Tissue Resource for Parkinson's Disease and Related Disorders), the National Institute on
Aging (P30 AG19610 Arizona Alzheimer's Disease Core Center), the Arizona Department of Health
Services (contract 211002, A Aging (P30 AG19610 Arizona Alzheimer's Disease Core Center), the Arizona Department of Health
Services (contract 211002, Arizona Alzheimer's Research Center), the Arizona Biomedical Research
Commission (contracts 4001, 001 Services (contract 211002, Arizona Alzheimer's Research Center), the Arizona Biomedical Research
Commission (contracts 4001, 0011, 05-901 and 1001 to the Arizona Parkinson's Disease Consortiun
the Michael J. Fox Foundation Commission (contracts 4001, 0011, 05-901 and 1001 to the Arizona Parkinson's Disease Consortium) and
the Michael J. Fox Foundation for Parkinson's Research
. the Michael J. Fox Foundation for Parkinson's Research

ROSMAP- We are grateful to the participants in the Religious Order Study, the Memory and Aging
Project. This work is supported by the US National Institutes of Health [U01 AG046152, R01 AG043617, R01 AG042210, R01 AG036042, R01 AG036836, R01 AG032990, R01 AG18023, RC2 AG036547, P50 AG016574, U01 ES017155, KL2 RR024151, K25 AG041906-01, R01 AG30146, P30 AG10161, R01 AG17917, R01 AG15819, K08 AG034290, P30 AG10161 and R01 AG11101.

Mount Sinai Brain Bank (MSBB)- This work was supported by the grants R01AG046170, RF1AG054014. Mount Sinai Brain Bank (MSBB)- This work was supported by the grand method of the component of the AMP-AD Target Discovery and Preclinical Validation Project. Brain tissue collection
component of the AMP-AD Target Discover Management of the AMP-AD Target Discovery and Preclinical Validation Project. Brain tissue collection
and characterization was supported by NIH HHSN271201300031C. and characterization was supported by NIH HHSN271201300031C.

University of Pittsburg-Kamboh (sa000012) data:

This study was supported by the National Institute on Aging (NIA) grants $A₀30653$, AG041718, AG064877 and P30-AG066468.

NACC Genentech Study (sa000013) data:

We would like to thank study participants, their families, and the sample collectors for their invaluable
contributions. This research was supported in part by the National Institute on Aging grant contributions. This research was supported in part by the National Institute on Aging grant
U01AG049508 (PI Alison M. Goate). This research was supported in part by Genentech, Inc. (PI Alison N U01AG049508 (PI Alison M. Goate). This research was supported in part by Genentech, Inc.
Goate, Robert R. Graham). Goate, Robert R. Graham). This research was supported in part by Genentech, Inc. (PI Alison M. Contractor M. C
Inc. (PI Alison M. Graham). Goate, Robert R. Graham).
The NACC database is funded by NIA/NIH Grant U01 AG016976. NACC data are contributed by these

The NACC database is funded by NIA/NIH Grant U01 AG016976. NACC data are contributed by these AG025688 (PI Allan Levey, MD, PhD), P30 AG010133 (PI Andrew Saykin, PsyD), P50 AG005146 (F
Marilyn Albert, PhD), P50 AG005134 (PI Bradley Hyman, MD, PhD), P50 AG016574 (PI Ronald Pe Marilyn Albert, PhD), P50 AG005134 (PI Bradley Hyman, MD, PhD), P50 AG016574 (PI Ronald Pete
MD, PhD), P30 AG013854 (PI M. Marsel Mesulam, MD), P30 AG008017 (PI Jeffrey Kaye, MD), P30 MD, PhD), P30 AG013854 (PI M. Marsel Mesulam, MD), P30 AG008017 (PI Jeffrey Kaye, MD), P30
AG010161 (PI David Bennett, MD), P30 AG010129 (PI Charles DeCarli, MD), P50 AG016573 (PI Frank AG010161 (PI David Bennett, MD), P30 AG010129 (PI Charles DeCarli, MD), P50 AG016573 (PI Frank LaFerla, PhD), P50 AG005131 (PI Douglas Galasko, MD), P30 AG028383 (PI Linda Van Eldik, PhD), P30
AG010124 (PI John Trojanowski, MD, PhD), P50 AG005142 (PI Helena Chui, MD), P30 AG012300 (PI Roger Rosenberg, MD), P50 AG005136 (PI Thomas Grabowski, MD), P50 AG005681 (PI John Morris, MD), Roger Rosenberg, MD), P50 AG005136 (PI Thomas Grabowski, MD), P50 AG005681 (PI John Morris,
P30 AG028377 (Kathleen Welsh-Bohmer, PhD), and P50 AG008671 (PI Henry Paulson, MD, PhD). P30 AG028377 (Kathleen Welsh-Bohmer, PhD), and P50 AG008671 (PI Henry Paulson, MD, PhD). $\overline{\hspace{0.2cm}}^{''}$

Samples from the National Cell Repository for Alzheimer's Disease (NCRAD), which receives government support under a cooperative agreement grant (U24 AG21886) awarded by the National Institute on Aging (NIA), were used in this study. We thank contributors who collected samples used in this study, as well as patients and their families, whose help and participation made this work possible.

well as patients and their families.
The Alzheimer's Disease Genetics Consortium supported the collection of samples used in The Alzheimer's Disease Genetics Consortium supported the collection of samples used in this study
through National Institute on Aging (NIA) grants U01AG032984 and RC2AG036528.
. through National Institute on Aging (NIA) grants U01AG032984 and RC2AG036528.
Cache County Study (sa000014) data:

We acknowledge the generous contributions of the Cache County Memory Study participants. Sequencing for this study was funded by RF1AG054052 (PI: John S.K. Kauwe). Sequencing for this study was funded by RF1AG054052 (PI: John S.K. Kauwe).

NIH, CurePSP and Tau Consortium PSP WGS (sa000015) data:
This project was funded by the NIH grant UG3NS104095 and supported by grants U54NS100693 and U54AG052427. Queen Square Brain Bank is supported by the Reta Lila Weston Institute for Neurological
Studies and the Medical Research Council UK. The Mayo Clinic Florida had support from a Morris K. - Cambias and the Medical Research Council UK. The Mayo Clinic Florida had support from a Morris K.
Udall Parkinson's Disease Research Center of Excellence (NINDS P50 #NS072187), CurePSP and the Tau Udall Parkinson's Disease Research Center of Excellence (NINDS P50 #NS072187), CurePSP and the 1
Consortium. The samples from the University of Pennsylvania are supported by NIA grant P01AG017 Consortium. The samples from the University of Pennsylvania are supported by NIA grant P01AG017586.

CurePSP and Tau Consortium PSP WGS (sa000016) data:

This project was funded by the Tau Consortium, Rainwater Charitable Foundation, and CurePSP. It was
also supported by NINDS grant U54NS100693 and NIA grants U54NS100693 and U54AG052427. Queen This project was function, the Tau Consortium, Admirator Charlies Management, 2014, 2017, Queen
also supported by NINDS grant U54NS100693 and NIA grants U54NS100693 and U54AG052427. Queen
Square Brain Bank is supported by square Brain Bank is supported by the Reta Lila Weston Institute for Neurological Studies and the
Medical Research Council UK. The Mayo Clinic Florida had support from a Morris K. Udall Parkinson's Medical Research Council UK. The Mayo Clinic Florida had support from a Morris K. Udall Parkinson's
Disease Research Center of Excellence (NINDS P50 #NS072187), CurePSP and the Tau Consortium. Th
samples from the Universit samples from the University of Pennsylvania are supported by NIA grant P01AG017586. Tissues were
received from the Victorian Brain Bank, supported by The Florey Institute of Neuroscience and Mental samples from the University of Pennsylvania are supported by the Spanish Calverns from the University of Penns
The University of Pennsylvania are supported by The Florey Institute of Neuroscience and Menta
Health, The Alfr Health, The Alfred and the Victorian Forensic Institute of Medicine and funded in part by Parkinson's
Victoria and MND Victoria. We are grateful to the Sun Health Research Institute Brain and Body Victoria and MND Victoria. We are grateful to the Sun Health Research Institute Brain and Body
Donation Program of Sun City, Arizona for the provision of human biological materials (or specific Donation Program of Sun City, Arizona for the provision of human biological materials (or specifi
description, e.g. brain tissue, cerebrospinal fluid). The Brain and Body Donation Program is supp description, e.g. brain tissue, cerebrospinal fluid). The Brain and Body Donation Program is supported by the National Institute of Neurological Disorders and Stroke (U24 NS072026 National Brain and Tissue
Resource for Parkinson's Disease and Related Disorders), the National Institute on Aging (P30 AG19610 Arizona Alzheimer's Disease Core Center), the Arizona Department of Health Services (contract 211002, Resource for Parkinson's Disease and Related Disorders), the National Institute on Aging (P30 AG19610
Arizona Alzheimer's Disease Core Center), the Arizona Department of Health Services (contract 211002
Arizona Alzheimer' Arizona Alzheimer's Disease Core Center), the Arizona Department of Health Services (contract 211002,
Arizona Alzheimer's Research Center), the Arizona Biomedical Research Commission (contracts 4001,
0011, 05-901 and 1001 Arizona Alzheimer's Research Center), the Arizona Biomedical Research Commission (contracts 4001,
0011, 05-901 and 1001 to the Arizona Parkinson's Disease Consortium) and the Michael J. Fox
Foundation for Parkinson's Resea 0011, 05-901 and 1001 to the Arizona Parkinson's Disease Consortium) and the Michael J. Fox
Foundation for Parkinson's Research. Biomaterial was provided by the Study Group DESCRIBE
theClinical Research of the German Cente Foundation for Parkinson's Research. Biomaterial was provided by the Study Group DESCRIBE of
theClinical Research of the German Center for Neurodegenerative Diseases (DZNE).
. theClinical Research of the German Center for Neurodegenerative Diseases (DZNE).
UCLA Progressive Supranuclear Palsy (sa000017) data:

If data are used for a publication, "on behalf of the AL-108-231 investigators" should be included in the
authorship list. \mathcal{I} authorship list. The AL-108-231 investigators in the AL-108-231 investigators in the AL-108-231 investigators \mathcal{I}

The Diagnostic Assessment of Dementia for the Longitudinal Aging Study of India (LASI-DAD) The Diagnostic Assessment of Demontation for the Demandial Aging Study of India (DAT Diagnostic Assessment of
In text: "The Longitudinal Aging Study in India. Diagnostic Assessment of Dementia data is spon

、
In text: "The Long
the National Instit the National Institute on Aging (grant numbers R01AG051125 and U01AG065958) and is conducted by the University of Southern California."

In references: "The Longitudinal Aging Study in India, Diagnostic Assessment of Dementia Study. Produced and distributed by the University of Southern California with funding from the National Institute on Aging (grant numbers R01AG051125 and U01AG065958), Los Angles, CA."

Dissecting the Genomic Etiology of non-Mendelian Early-Onset Alzheimer Disease (EOAD) and Related Discremi_c the Generate Etiology of non-Mendelian Early-Onset Alzheimer Disease (EOAD) and Related
Phenotypes (sa000023) data: Phenotypes (sa000023) data:

This work was supported by U01AG057659.
Supported by U01AG057659. supported by U01AG057659.
The National Institutes of Health, National Institute on Aging (NIH-NIA) supported this work through the

following grants: ADGC, U01 AG032984, RC2 AG036528; samples from the National Centralized Repository for Alzheimer's Disease and Related Dementias (NCRAD), which receives government support under a cooperative agreement grant (U24 AG21886) awarded by the National Institute on
Aging (NIA), were used in this study. Sequencing data generation and harmonization is supported by the support under a cooperative agreement grant (U24 AG21886) and also provided by the National Institute on
Aging (NIA), were used in this study. Sequencing data generation and harmonization is supported by
Genome Center for Aging (NIA), were the later interesting the process case of the interest in this study. The completed by NIAGADS,
Genome Center for Alzheimer's Disease, U54AG052427, and data sharing is supported by NIAGADS,
U24AG041689. W U24AG041689. We thank contributors who collected samples used in this study, as well as patients and their families, whose help and participation made this work possible.

NIH grants supported enrollment and data collection for the individual studies including the Alzheimer's Disease Centers (ADC, P30 AG062429 (PI James Brewer, MD, PhD), P30 AG066468 (PI Oscar Lopez, MD), P30
P30 AG062421 (PI Bradley Hyman, MD, PhD), P30 AG066509 (PI Thomas Grabowski, MD), P30 AG066514 (PI Mary Sano, PhD), P30 AG066530 (PI Helena Chui, MD), P30 AG066507 (PI Marilyn Albert, AG066514 (PI Mary Sano, PhD), P30 AG066530 (PI Helena Chui, MD), P30 AG066507 (PI Marily
PhD), P30 AG066444 (PI John Morris, MD), P30 AG066518 (PI Jeffrey Kaye, MD), P30 AG06651. PhD), P30 AG066444 (PI John Morris, MD), P30 AG066518 (PI Jeffrey Kaye, MD), P30 AG066512 (PI
Thomas Wisniewski, MD), P30 AG066462 (PI Scott Small, MD), P30 AG072979 (PI David Wolk, MD), P30 Thomas Wisniewski, MD), P30 AG066462 (PI Scott Small, MD), P30 AG072979 (PI David Wolk, MD), P30 AG072972 (PI Charles DeCarli, MD), P30 AG072976 (PI Andrew Saykin, PsyD), P30 AG072975 (PI David
Bennett, MD), P30 AG072978 (PI Neil Kowall, MD), P30 AG072977 (PI Robert Vassar, PhD), P30 AG072972 (PI Charles DeCarli, MD), P30 AG072976 (PI Andrew Saykin, PsyD), P30 AG072975 (PI David AG066519 (PI Frank LaFerla, PhD), P30 AG062677 (PI Ronald Petersen, MD, PhD), P30 AG07928
Reiman, MD), P30 AG062422 (PI Gil Rabinovici, MD), P30 AG066511 (PI Allan Levey, MD, PhD), I Reiman, MD), P30 AG062422 (PI Gil Rabinovici, MD), P30 AG066511 (PI Allan Levey, MD, PhD), P30
AG072946 (PI Linda Van Eldik, PhD), P30 AG062715 (PI Sanjay Asthana, MD, FRCP), P30 AG072973 (PI Russell Swerdlow, MD), P30 AG066506 (PI Todd Golde, MD, PhD), P30 AG066508 (PI Stephen Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suzanne Craft, در المساحل العربية التي تعدد المساحل المساحل المساحل المساحل المساحل المساحل المساحل المساحل المساحل المساحل
Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suza
PhD), P30 AG072931 (PI Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suzanne Craft, PhD), P30 AG072931 (PI Henry Paulson, MD, PhD), P30 AG066546 (PI Sudha Seshadri, MD), P20 Rosenberg, MD), P20 AG068082 (PI Angela Jefferson, PhD), P30 AG072958 (PI Heather Whitson, MD),
P30 AG072959 (PI James Leverenz, MD). The Miami ascertainment and research were supported in pa P30 AG072959 (PI James Leverenz, MD). The Miami ascertainment and research were supported in pa
through: RF1AG054080, R01AG027944, R01AG019085, R01AG028786-02, RC2AG036528. The Columb through: RF1AG054080, R01AG027944, R01AG019085, R01AG028786-02, RC2AG036528. The Columbia
ascertainment and research were supported in part through: R37AG015473 and U24AG056270. The ascertainment and research were supported in part through: R37AG015473 and U24AG056270. The University of Washington ascertainment and research were supported in part through R01AG044546, RF1AG053303, RF1AG058501, U01AG058922 and R01AG064877. ϵ ₂₃ ϵ ₂₃ ϵ ₂₅ ϵ ₃ ϵ ₃₅ ϵ ₃₅ ϵ ₃₅ ϵ ₃₅ ϵ ₃₅

- Code availability
• VCPA code https://bitbucket.org/NIAGADS/vcpa-pipeline/src/master/ •
- - SV related code

	o https://github.com/DecodeGenetics/svimmer. o https://
https:// o
		- mater , names , names
- QC code https://bitbucket.org/Taha Iqbal UPenn/gcad-vcf-qc public/ • ζ code <u>https://bitbucket.org/Taha_Iqbal_UPenn/gcad-vcf-qc_public/</u>
Tahawa

Data availability

- Complete list of files showing the NIAGADS accession number together with a description of the files • is available at **Supplementary Table S5**.
NIAGADS DSS is accessible at <u>https://dss.niagads.org/</u>
- NIAGADS DSS is accessible at <u>https://dss</u>
-• NIAGADS DSS is accessible at https://doi.org/
niagads.org/windows/2010/01/2010

- •
- VariXam is accessible at <u>https://varixam.niagads.org/</u>
GenomicsDB is accessible at https://www.niagads.org/genomics/app • accessible at the head of $\frac{m_{\text{max}}}{m_{\text{max}}}$ and $\frac{m_{\text{max}}}{m_{\text{max}}}$

\overline{a} Figures and Tables

|
|
| **.**
Non-Hispanic Black with African Ancestry (NHB-AA) samples are from Africa, and North America
(Canada, USA); Asian and Asian American are from Asia, and North America (Canada, USA)**;** (Canada, USA); Asian and Asian American are from Asia, and North America (Canada, USA);
Hispanic/Latino (HL) are from North America (Mexico/Caribbean, Canada, USA), and South Amer Hispanic/Latino (HL) are from North America (Mexico/Caribbean, Canada, USA), and South America; non-Hispanic white (NHW) are from Europe and North America (Canada, USA). Lastly, some samples categorized as others or unknown and they are from Australia. B) Comparison of reported ethnicity against those inferred by GRAF-POP and SCOPE based methods. C) Estimated GRAF-pop ancestral components Pe, Pf, and Pa for all participants. D) PCA plot on R4 participants colored by reported ethnicity (top) or SCOPE. components Pe, Pf, and Pa for all participants. D) PCA plot on R4 participants colored by reported

ethnicity (top) or SCOPE.
 Figure 2 WGS sample quality. A) Coverage (30x) for the ADSP R4 data. Red dotted line indicates

coverage value at 30. 99% of samples pass this threshold. **B)** Number of SNVs called per sample i Figure 2 WGS sample quality. A) Coverage (30x) for the ADSP R4 data. Red dotted line indicates l
C Coverage value at 30. 99% of samples pass this threshold. **B)** Number of SNVs called per sample in coverage (30x) for the ADS reported ethnic group. Line in each displayed boxplot denotes the mean value where each dot is reported ethnic group. Line in each displayed boxplot denotes the mean value where each dot is a
sample. reported ethnic group. Line in each displayed boxplot denotes the mean value where each dot is a
sample.

Figure 3 The distribution of variant types across the genome, with a specific focus on high-risk loss-of-|
|
| genome, categorized by genomic annotation as follows: insertions and deletions, loss-of-function function variants. A) Bar chart depicting the breakdown of the total number of character increasing
genome, categorized by genomic annotation as follows: insertions and deletions, loss-of-function
variants, upstream gene v genome.
variants, upstream gene variants, synonymous variants, non-coding transcript exon variants, misse
variants, intron variants, intragenic variants, intergenic variants, downstream variants, 5 prime UT variants, intron variants, intragenic variants, intergenic variants, downstream variants, 5 prime UTR
variants, and 3 prime UTR variants. **B**) The distribution of 224,594 loss-of-function variants is further variants, and 3 prime UTR variants. **B**) The distribution of 224,594 loss-of-function variants is further
broken down into the following categories: frameshift (39%), stop gained (27%), splice donor (16%), broken down into the following categories: frameshift (39%), stop gained (27%), splice donor (16%), splice acceptor (12%), start lost (4%), and stop lost (2%).

Figure 4 – Comparison of the number of SVs called in the ADSP R4 dataset across different reported |
|
| ethnicities. SVs can be categorized into four different types of SVs: deletion (DEL), duplication (DUP), inversion (INV), and insertion (INS).

Figure 5-ADSP-PHC Release ($ng00067.v11$) Sample sizes ("N" on the right) reflect individuals with ADSP |
|
| Figure 5 – ADSP-PHC Release (ngl. 1) Sample sizes (ng. 1) Sample sizes (with ADSP-PHC right) reflect in the right) reflect in the right) reflect in the right of the right sequencing data in R4.
Figure 6 Browser of variants and annotations of diversified samples. A) VariXam interface. A variant

|
|
| browser displaying all genomic variants identified in the ADSP whole genome and exome data across
releases. The figure below shows the search results of APOE. Accessible at: https://varixam.niagads.org/. releases. The figure below shows the search results of *APOE*. Accessible at: <u>https://varixam.niagads.o</u>
B) The R4 Variants can be visually inspected as a track on the NIAGADS Genome Browser. The track **B)** The R4 Variants can be visually inspected as a track on the NIAGADS Genome Browser. The track displays annotated short INDELS and SNVs that passed the biallelic QC criteria. Track annotations include the most severe v displays annotated short INDELS and SNVs that passed the biallelic QC criteria. Track annotations
include the most severe variant consequences and consequence impacts predicted by the ADSP displays and success and some that passed the passed the biallelic QC criteria. Track and the biallelic and in
include the most severe variant consequences and consequence impacts predicted by the ADSP
annotation pipeline annotation pipeline and mappings to dbSNP refSNP identifiers. The track settings menu can be used to recolor the variants based on various annotations; the legend (made available by clicking on the track reforma
name) will update accordingly. Users can zoom into regions of interest (here, the green rectangle name) will update accordingly. Users can zoom into regions of interest (here, the green rectangle
highlights the region displayed in the close-up inset) to view sequence information and click on highlights the region displayed in the close-up inset) to view sequence information and click on highlights the region displayed in the close-up inset) to view sequence information and click on

It is made available under a CC-BY-ND 4.0 International license. medRxiv preprint doi: [https://doi.org/10.1101/2024.12.03.24317000;](https://doi.org/10.1101/2024.12.03.24317000) this version posted December 6, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted

individual variants for a brief summary of the annotations. Full annotation results can be browsed by following the link to the GenomicsDB record for the variant.

Table 1 - Number of variants (SNVs and indels) identified in the four major ethnic groups in ADSP R4 $\overline{}$ data, broken down allele frequency (AF). Shown in top is the variant count and percentages per
ethnicity. Instead of the total number of variants identified, we showed at the bottom the ethnic data, broken down allele frequency (AF). Shown in top is the variant count and percentages per
ethnicity. Instead of the total number of variants identified, we showed at the **bottom** the ethnic
specific variants. ethnicity. Instead of the total number of variants identified, we show the the three than the ethnics is the ethnic
specific variants. specific variants.
Table 2 – ADSP R4 released file set. All files are available under

]
| https://dss.niagads.org/datasets/ng00067/. Both individual (CRAMs, gVCFs, SV VCFs and phenotypes) and summary level files (VCFs, GDS, annotation files) are available. Annotation and LD reference panel files are also available in NIAGADS Open Access Data Portal https://dss.niagads.org/open-access-data-
portal/. portal/.

It is made available under a CC-BY-ND 4.0 International license. medRxiv preprint doi: [https://doi.org/10.1101/2024.12.03.24317000;](https://doi.org/10.1101/2024.12.03.24317000) this version posted December 6, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted

Figure 1 Participants in ADSP R4 dataset. A) Worldwide cohorts assembled for this ADSP R4 dataset.
Non-Hispanic Black with African Ancestry (NHB-AA) samples are from Africa, and North America
(Canada, USA); Asian and Asian non-Hispanic State March Miller, Miller March Willer and North America, Clamber, USA);
(Canada, USA); Asian and Asian American are from Asia, and North America (Canada, USA);
Hispanic/Latino (HL) are from North America (Me (Canada, Canada, Canada, Canada, Canada, USA); Asian American Asian American American (Canada, USA), and South
hon-Hispanic white (NHW) are from Europe and North America (Canada, USA). Lastly, some non-Hispanic white (NHW) are from Europe and North America (Canada, USA). Lastly, some samples categorized as others or unknown and they are from Australia. B) Comparison of reported ethnicity against those inferred by GRAF-POP and SCOPE based methods. C) Estimated GRAF-pop ancestral components Pe, Pf, and Pa for all participants. D) PCA plot on R4 participants colored by reported ethnicity (top) or SCOPE. components Pe, Pf, and Pa for all participants. D) PCA plot on R4 participants colored by reported components Pey, J, and Patricipants Pey, Participants Colored by reported by r ethnicity (top) or SCOPE.

It is made available under a CC-BY-ND 4.0 International license. medRxiv preprint doi: [https://doi.org/10.1101/2024.12.03.24317000;](https://doi.org/10.1101/2024.12.03.24317000) this version posted December 6, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted

Figure 2 WGS sample quality. A) Coverage (30x) for the ADSP R4 data. Red dotted line indicates coverage value at 30. 99% of samples pass this threshold. B) Number of SNVs called per sample in each reported ethnic group. Li reported ethnic group. Line in each displayed boxplot denotes the mean value where each dot is a reproduce the mean value of the mean value where the mean value where each dot is a mean value where each dot i
control to the mean value where each dot is a mean value where each dot is a mean value where where where we h sample.

Figure 3 The distribution of variant types across the genome, with a specific focus on high-risk loss-ofgenome, categorized by genomic annotation as follows: insertions and deletions, loss-of-function
variants, upstream gene variants, synonymous variants, non-coding transcript exon variants, misse genome, categorized by genomic annotation as follows: insertions and deletions, loss-of-function genome.
variants, upstream gene variants, synonymous variants, non-coding transcript exon variants, misse
variants, intron variants, intragenic variants, intergenic variants, downstream variants, 5 prime UT variants, apen earn gene variants, synonymous variants, non-colling transcript exonomics, missense utility var
variants, intron variants, intragenic variants, intergenic variants, downstream variants, 5 prime UTR
variants, variants, interest variants, intergents can there genie variants, at the contract meaning suppresentation vari
Variants, and 3 prime UTR variants. B) The distribution of 224,594 loss-of-function variants is further
broken broken down into the following categories: frameshift (39%), stop gained (27%), splice donor (16%), splice acceptor (12%), start lost (4%), and stop lost (2%). splice acceptor (12%), start lost (4%), and stop lost (2%).

Figure 4 – Comparison of the number of SVs called in the ADSP R4 dataset across different reported
ethnicities. SVs can be categorized into four different types of SVs: deletion (DEL), duplication (DUP), inversion (INV), and insertion (INS). $\sum_{i=1}^{N}$

It is made available under a CC-BY-ND 4.0 International license. medRxiv preprint doi: [https://doi.org/10.1101/2024.12.03.24317000;](https://doi.org/10.1101/2024.12.03.24317000) this version posted December 6, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted

Figure 5 – ADSP-PHC RELEASE (ng00067.11) Sample sizes ("N" on the right) reflect individuals with ADSP.
sequencing data in R4. sequencing data in R4.

It is made available under a CC-BY-ND 4.0 International license. medRxiv preprint doi: [https://doi.org/10.1101/2024.12.03.24317000;](https://doi.org/10.1101/2024.12.03.24317000) this version posted December 6, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted

A)

VariXam is an aggregated database and a variant browser that shows genomic variants detected on whole-genome/wholeexome sequence (WGS/WES) data of the Alzheimer's Disease Sequencing Project (ADSP). The variants were processed by
VCPA (Variant Calling Pipeline and data management tool) and released by NIAGADS. The database currently in variants of R1 5K WGS, R2 20K WES, R3 17K WGS, and R4 36K WGS. The human reference genome used is GRCh38.

ADSP and NIAGADS are sponsored by the National Aging Institute (NIA) and aim to improve our understanding and provide insight of genetic factors of Alzheimer's Disease.

Q

Examples - Gene: APOE - Chromosome Position: 5:17100035 - Chromosome Region: 19:1593000-1594000 rsID: rs146621479

Showing results for: APOE

- Enter the region, gene or rsID and click the search icon to see autosomal varia

B)

Figure 6 Browser of variants and annotations of diversified and annotations, then and exome data across
browser displaying all genomic variants identified in the ADSP whole genome and exome data across
releases. The figure releases. The figure below shows the search results of *APOE*. Accessible at: <u>https://varixam.niagads.o</u>
B) The R4 Variants can be visually inspected as a track on the NIAGADS Genome Browser. The track **B)** The R4 Variants can be visually inspected as a track on the NIAGADS Genome Browser. The track displays annotated short INDELS and SNVs that passed the biallelic QC criteria. Track annotations include the most severe v displays annotated short INDELS and SNVs that passed the biallelic QC criteria. Track annotations
include the most severe variant consequences and consequence impacts predicted by the ADSP displays and success and some that passed the passed the biallelic QC criteria. Track and the biallelic and in
include the most severe variant consequences and consequence impacts predicted by the ADSP
annotation pipeline include the most severe variant consequences and consequences impact predicted by the most
annotation pipeline and mappings to dbSNP refSNP identifiers. The track settings menu can be t
recolor the variants based on variou recolor the variants based on various annotations; the legend (made available by clicking on the track
name) will update accordingly. Users can zoom into regions of interest (here, the green rectangle name) will update accordingly. Users can zoom into regions of interest (here, the green rectangle
highlights the region displayed in the close-up inset) to view sequence information and click on highlights the region displayed in the close-up inset) to view sequence information and click on
individual variants for a brief summary of the annotations. Full annotation results can be browsed individual variants for a brief summary of the annotations. Full annotation results can be browsed by following the link to the GenomicsDB record for the variant. following the link to the GenomicsDB record for the variant.

It is made available under a CC-BY-ND 4.0 International license. medRxiv preprint doi: [https://doi.org/10.1101/2024.12.03.24317000;](https://doi.org/10.1101/2024.12.03.24317000) this version posted December 6, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted

Table 1 – Number of variants (SNVS and indeed) in the four major ethnic groups interest in the data, broken down allele frequency (AF). Shown in top is the variant count and percentages per ethnic ethnic ethnic approach et data, broken down allele frequency (AF). Shown in top is the variant count and percentages per
ethnicity. Instead of the total number of variants identified, we showed at the **bottom** the ethnic
specific variants. ethnicity. Instead of the total number of the total number of variants is a the ethnicity of variants in t

Table 2 – ADSP R4 released file set. All files are arounder the set.
https://dss.niagads.org/datasets/ng00067/. Both individual (CRA
and summary level files (VCFs, GDS, annotation files) are availab https://decepter.org/datasets/ndottps://doi.org/datasets/ng000631.
and summary level files (VCFs, GDS, annotation files) are available. Annotation and LD reference panel
files are also available in NIAGADS Open Access Data files are also available in NIAGADS Open Access Data Portal https://dss.niagads.org/open-access-data-
portal/.

- $\begin{array}{c} \n\frac{1}{2} & \frac{1}{2} \\ \n\frac{1}{2} & \frac{1}{$ References
1. Strittmatter, W.J., et al., Binding of human apolipoprotein E to synthetic amyloid beta peptide:
isoform-specific effects and implications for late-onset Alzheimer disease. Proc Natl Acad Sci U S 1. Stritter, Stritter, Stritter, M.J., 1993. 90(17): p. 8098-102.

2. Harold, D., et al., Genome-wide association study identifies variants at CLU and PICALM

1. associated with Alzheimer's disease. Nat Genet, 2009. 41(10)
- A, 1993. **90**(17): p. 8098-102.
Harold, D., et al., *Genome-wide association study identifies variants at CLU and PICALM*
associated with Alzheimer's disease. Nat Genet, 2009. 41(10): p. 1088-93.
Lambert, J.C., et al., *Ge* A, 1993. 90(17): p. 8098-102.
Harold, D., et al., *Genome-wide association study identifies variants at CLU and PICALM*
associated with Alzheimer's disease. Nat Genet, 2009. 41(10): p. 1088-93.
Lambert, J.C., et al., *Ge*
-
-
- 2. Harold With Alzheimer's disease. Nat Genet, 2009. 41(10): p. 1088-93.

2. Lambert, J.C., et al., *Genome-wide association study identifies variants at CLU and CR1*

associated with Alzheimer's disease. Nat Genet, 2009. 3. associated with Alzheimer's disease. Nat Genet, 2009. 41(10): p. 1094-9.

3. Bellenguez, C., et al., New insights into the genetic etiology of Alzheimer's disease and related

3. Annel Cenet, 2022. 54(4): p. 412-436.

5 4. Bellenguez, C., et al., New insights into the genetic etiology of Alzheimer's disease and related
dementias. Nat Genet, 2022. **54**(4): p. 412-436.
5. Kunkle, B.W., et al., *Genetic meta-analysis of diagnosed Alzheimer's* dementias. Nat Genet, 2022. **54**(4): p. 412-436.
Kunkle, B.W., et al., *Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk*
loci and implicates Abeta, tau, immunity and lipid processing. Nat Genet 5. Kunkle, B.W., 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): p. 414-

430.

Lambert, J.C.,
- Alzheimer's disease. Nat Genet, 2013. 45(12): p. 1452-8.
- Lambert, J.C., et al., *Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease.* Nat Genet, 2013. 45(12): p. 1452-8.
Wightman, D.P., et al., *Author Correction: A genome-wide asso*
- Alzheimer's disease. Nat Genet, 2013. 45(12): p. 1452-8.

7. Wightman, D.P., et al., Author Correction: A genome-wide association study with 1,126,563

individuals identifies new risk loci for Alzheimer's disease. Nat Gene Alzheimer's disease. Nat Genet, 2013. **45**(12): p. 1452-8.
Wightman, D.P., et al., *Author Correction: A genome-
individuals identifies new risk loci for Alzheimer's disease
Bergem, A.L., K. Engedal, and E. Kringlen, <i>The* 1. Wightman, University Schools individuals identifies new risk loci for Alzheimer's disease. Nat Genet, 2021. 53(12): p. 1722.

8. Bergem, A.L., K. Engedal, and E. Kringlen, The role of heredity in late-onset Alzheimer di individuals identifies new risk loci for Alzheimer's disease. Nat Genet, 2021. **53**(12): p. 1722.
Bergem, A.L., K. Engedal, and E. Kringlen, *The role of heredity in late-onset Alzheimer disease*
vascular dementia. A twin vascular dementia. A twin study. Arch Gen Psychiatry, 1997. 54(3): p. 264-70.
National, A. Engel
Alzheimer disease and E. S. Alzheimer disease and the role of the vascular dementia. A twin study. Arch Gen Psychiatry, 1997. 54(3): p. 264-70.

-
- Psychiatry, 2006. **63**(2): p. 168-74.
Popejoy, A.B. and S.M. Fullerton, *Genomics is failing on diversity.* Nature, 2016. **538**(7624): p.
161-164.
- 9. Gatz, M., 2006. 63(2): p. 168-74.

9. Gataxienty, 2006. 63(2): p. 168-74.

10. Popejoy, A.B. and S.M. Fullerton, *Genomics is failing on diversity.* Nature, 2016. 538(7624): p.

161-164.

Farrer, L.A., et al., *Effects* 161-164.

11. Farrer, L.A., et al., *Effects of age, sex, and ethnicity on the association between apolipoprotein E*

11. Farrer, L.A., et al., *Effects of age, sex, and ethnicity on the association between apolipoprotein* Farrer, L.*I*
g*enotype*
- 11. Farrer, L.A., 1997. 1999. The age, A., and the analysis. APOE and Alzheimer Disease Meta Analysis

Consortium. JAMA, 1997. 278(16): p. 1349-56.

12. Crean, S., et al., Apolipoprotein E epsilon4 prevalence in Alzheimer' genotype and Alaxheimer 2008 (2013)
Consortium. JAMA, 1997. 278(16): p. 1349-56.
Crean, S., et al., *Apolipoprotein E epsilon4 prevalence in Alzheimer's disease patients varies*
across global populations: a systematic lite Crean, S., et al., *Apolipoprotein E epsilon4 pacross global populations: a systematic liters*
Cogn Disord, 2011. 31(1): p. 20-30.
Fortea, J., et al., *APOE4 homozygozity represe*
Nat Med, 2024. 30(5): p. 1284-1291.
Miyash Crean, S., et al., Apolipoprotein E epsilon4 prevalence in Alzheimer's disease patients varies
across global populations: a systematic literature review and meta-analysis. Dement Geriatr
Cogn Disord, 2011. **31**(1): p. 20-3
- Cogn Disord, 2011. **31**(1): p. 20-30.
Fortea, J., et al., *APOE4 homozygozity represents a distinct genetic form of Alzheimer's disease.*
Nat Med, 2024. **30**(5): p. 1284-1291.
- Miyashita, A., et al., SORL1 is genetically associated with late-onset Alzheimer's disease in
Japanese, Koreans and Caucasians. PLoS One, 2013. 8(4): p. e58618.
- Cogn Disord, 2011. **31**(1): p. 20-30.
Fortea, J., et al., *APOE4 homozygozity represents a distinct genetic form of Alzheimer's disease.*
Nat Med, 2024. **30**(5): p. 1284-1291.
Miyashita, A., et al., *SORL1 is genetically a* 13. Fortea, J., et al., *APOE4 homozygozity represents a distinct genetic form of Alzheimer's disease.*

Nat Med, 2024. **30**(5): p. 1284-1291.

14. Miyashita, A., et al., *SORL1 is genetically associated with late-onset Al* Miyashita, A., et al., *SORL1* is genetically associated with late-onset Alzheimer's disease in

Japanese, Koreans and Caucasians. PLoS One, 2013. **8**(4): p. e58618.

15. Reitz, C., et al., Variants in the ATP-binding cass Reitz, C., et al., *Variants in the ATP-binding cassette transporter*

24, and the risk of late-onset Alzheimer disease in African Americar

1483-92.

Meng, Y., et al., Association of polymorphisms in the Angiotensin-co

A Reitz, C., et al., Variants in the ATP-binding cassette transporter (ABCA7), apolipoprotein E

24, and the risk of late-onset Alzheimer disease in African Americans. JAMA, 2013. **309**(14): p.

1483-92.

Meng, Y., et al., A
- --------
Meng, Y.,
Alzheimei
- 1483-92.

Meng, Y., et al., Association of polymorphisms in the Angiotensin-converting enzyme gene with

Alzheimer disease in an Israeli Arab community. Am J Hum Genet, 2006. **78**(5): p. 871-877.

Leung, Y.Y., et al., *VCP* Alzheimer disease in an Israeli Arab community. Am J Hum Genet, 2006. **78**(5): p. 871-877.

17. Leung, Y.Y., et al., *VCPA: genomic variant calling pipeline and data management tool for

Alzheimer's Disease Sequencing Proj*
-
- Leung, Y.Y., et al., *VCPA: genomic variant calling pipeline and data management to*
Alzheimer's Disease Sequencing Project. Bioinformatics, 2019. **35**(10): p. 1768-1770.
Mersha, T.B. and T. Abebe, *Self-reported race/ethn* Alzheimer's Disease Sequencing Project. Bioinformatics, 2019. 35(10): p. 1768-1770.

Alzheimer's Disease Sequencing Project. Bioinformatics, 2019. 35(10): p. 1768-1770.

18. Mersha, T.B. and T. Abebe, Self-reported race/et 18. Mercential impact on understanding health disparities. Hum Genomics, 2015. 9(1): p. 1.

19. Jin, Y., et al., GRAF-pop: A Fast Distance-Based Method To Infer Subject Ancestry from Multiple

6 Genotype Datasets Without P potential impact on understanding in the process contours of the Subject Ancestry from
Jin, Y., et al., *GRAF-pop: A Fast Distance-Based Method To Infer Subject Ancestry from*
Genotype Datasets Without Principal Components 19. 19. June 20. Chen, S., et al., A genomic mutational components Analysis. G3 (Bethesda), 2019. 9(8): p. 2447-

19. S., et al., A genomic mutational constraint map using variation in 76,156 human genomes.

19. Nature, 20
- Chen,
Natur
- 2461.
Chen, S., et al., A genomic mutational constraint map using variation in 76,156 human genomes.
Nature, 2024. 625(7993): p. 92-100.
Karczewski, K.J., et al., *The mutational constraint spectrum quantified from variati* Nature, 2024. 625(7993): p. 92-100.
Karczewski, K.J., et al., *The mutational constraint spectrum quantified from variation in 141,456*
humans. Nature, 2020. 581(7809): p. 434-443.
Butkiewicz, M., et al., *Functional annot*
-
- Nature, 2024. 625(7993): p. 92-100.

21. Karczewski, K.J., et al., *The mutational constraint spectrum quantified from variation in 141,456*
 humans. Nature, 2020. 581(7809): p. 434-443.

22. Butkiewicz, M., et al., *Fun* 22. Butkiewicz, M., et al., Functional annotation of genomic variants in studies of late-onset

22. Butkiewicz, M., et al., Functional annotation of genomic variants in studies of late-onset

23. Cingolani, P., et al., A p Butkiewicz, M., et al., *Functional annotatio.*
Alzheimer's disease. Bioinformatics, 2018. **34**(1
Cingolani, P., et al., A program for annotatin
polymorphisms, SnpEff: SNPs in the genome of
3. Fly (Austin), 2012. **6**(2): p Alzheimer's disease. Bioinformatics, 2018. **34**(16): p. 2724-2731.

23. Cingolani, P., et al., A program for annotating and predicting the effects of single nucleotide

polymorphisms, SnpEff: SNPs in the genome of Drosophi 23. Chingolani, P. Shipsim, P. Shipsim for any the effects of Drosophila melanogaster strain w1118; iso-2; iso-
23. Fly (Austin), 2012. 6(2): p. 80-92.
24. Wang, H., et al., *Structural Variation Detection and Association*
-
-
- calls. Bioinformatics, 2017. 33(15): p. 2251-2257. 3. Fly (Austin), 2012. 6(2): p. 80-92.

Wang, H., et al., *Structural Variation Detection and Association Analysis of Whole-Genome-*

Sequence Data from 16,905 Alzheimer's Diseases Sequencing Project Subjects. medRxiv, 202 Zheng, X., et al., SegArray-a storage-efficient high-performance data format for WGS variant
- Quick, C., et al., *emeraLD: rapid linkage disequilibrium estimation with massive datasets.*

26. Zheng, X., et al., *SeqArray-a storage-efficient high-performance data format for WGS variant*

26. Zheng, X., et al., *SeqA* Sequence Data from 16,905 Alzheimer's Diseases Sequencing Project Subjects. medRxiv, 2023.
Quick, C., et al., emeraLD: rapid linkage disequilibrium estimation with massive datase
Bioinformatics, 2019. **35**(1): p. 164-166.
 26. Zheng, X., et al., *SeqArray-a storage-efficient high-performance data format for WGS variant*
26. Zheng, X., et al., *SeqArray-a storage-efficient high-performance data format for WGS variant*
27. Greenfest-Allen, E., 27. Greenfest-Allen, E., et al., NIAGADS Alzheimer's GenomicsDB: A resource for exploring
27. Greenfest-Allen, E., et al., NIAGADS Alzheimer's GenomicsDB: A resource for exploring
28. Schubach, M., et al., CADD v1.7: using Greenfest-Allen, E., et al., *NIAGADS Alzheim-Alzheimer's disease genetic and genomic knowled*
Alzheimer's disease genetic and genomic knowled
1136.
Schubach, M., et al., CADD v1.7: using protein
nucleotide-level scores to
- 27. Greenfest-Allen, E., et al., *NIAGADS Alzheimer's GenomicsDB: A resource for exploring Alzheimer's disease genetic and genomic knowledge. Alzheimers Dement, 2024. 20(2): p. 1123-1136.
Schubach, M., et al., <i>CADD v1.7:* Alzheimer's disease genetic and genomic knowledge. Alzheimers Dement, 2024. **20**(2): p. 1123-
1136.
Schubach, M., et al., *CADD v1.7: using protein language models, regulatory CNNs and other
nucleotide-level scores to impr* 1136.
Schubach, M., et al., CADD v1.7: using protein language models, regulatory CNNs and other
nucleotide-level scores to improve genome-wide variant predictions. Nucleic Acids Res, 2024. nucleotide-level scores to improve genome-wide variant predictions. Nucleic Acids Res, 2024.
52(D1): p. D1143-D1154.
52(D1): p. D1143-D1154. $52(D1): p. D1143-D1154.$ $\frac{1}{2}$

It is made available under a CC-BY-ND 4.0 International license. medRxiv preprint doi: [https://doi.org/10.1101/2024.12.03.24317000;](https://doi.org/10.1101/2024.12.03.24317000) this version posted December 6, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted

-
-
- 2022. 29. The same of all photons of the buman genome. Nucleic Acids Res, 2023. 51(D1): p. D1300-D1311.

29. Genomes Project, C., et al., A global reference for human genetic variation. Nature, 2015.

29. S26(7571): p. 68variation across the human generation across the denomes Project, C., et al., A global reference for human genetic variation. Natu
526(7571): p. 68-74.
Prokopenko, D., et al., Region-based analysis of rare genomic variants 31. Prokopenko, D., et al., Region-based analysis of rare genomic variants in whole-genome
sequencing datasets reveal two novel Alzheimer's disease-associated genes: DTNB and DLG2.
Mol Psychiatry, 2022. 27(4): p. 1963-196 rokopenko, D., et
sequencing datasets
- Sequencing datasets reveal two novel Alzheimer's disease-associated genes: DTNB and DLG2.

Mol Psychiatry, 2022. 27(4): p. 1963-1969.

32. Prokopenko, D., et al., Whole-genome sequencing reveals new Alzheimer's disease-ass Mol Psychiatry, 2022. 27(4): p. 1963-1969.
Prokopenko, D., et al., Whole-genome sequencing reveals new Alzheimer's disease-associated
rare variants in loci related to synaptic function and neuronal development. Alzheimers
- 2022. 2022. 2022. 2022. 2022. 2022. 2023. 2023. 2023. 2023. 2023. 2023. 2023. 2023. 2023. 2023. 2023. 2023. 20
- Frokopenko, D., et al., Whole-genome sequencing reveals new Alzheimer's disease-associated

rare variants in loci related to synaptic function and neuronal development. Alzheimers Dement,

2021. 17(9): p. 1509-1527.

Lagis 2021. **17**(9): p. 1509-1527.
Lagisetty, Y., et al., *Identification of risk genes for Alzheimer's disease by gene embedding. Cell*
Genom, 2022. **2**(9).
Lee, W.P., et al., *Copy Number Variation Identification on 3,800 Alzh* Genom, 2022. **2**(9).
Lee, W.P., et al., *Copy Number Variation Identification on 3,800 Alzheimer's Disease Whole
Genome Sequencing Data from the Alzheimer's Disease Sequencing Project. Front Genet, 2021.
12: p. 752390.*
- Lagisetty, Y., et al., *Identification of risk genes for Alzheimer's disease by gene embedding*. Cell
Genom, 2022. **2**(9).
34. Lee, W.P., et al., *Copy Number Variation Identification on 3,800 Alzheimer's Disease Whole*
6 Lee, W.P., et al., Copy Number Variation Identification on 3,800 Alzheimer's Disease Whole
Genome Sequencing Data from the Alzheimer's Disease Sequencing Project. Front Genet, 2021.
12: p. 752390.
Osterman, M.D., et al., F Genome Sequencing Data from the Alzheimer's Disease Sequencing Project. Front Genet, 2021.
12: p. 752390.
Osterman, M.D., et al., Founder population-specific weights yield improvements in performance
of polygenic risk sc ------------
Osterman, M.C
of polygenic ris 35. Osterman, M. Hans, Scholars (M. D., Hans, P. P. M. Press, P. P. M. Press, P. P. P. P. 2023. 4(4): p.
36. Prokopenko, D., et al., Identification of Novel Alzheimer's Disease Loci Using Sex-Specific Family-
35. Bhattarai
- Prokopenko, D., et al., Identification of Novel Alzheimer's Disease Loci Using Sex-Specific Family-
Based Association Analysis of Whole-Genome Sequence Data. Sci Rep, 2020. 10(1): p. 5029.
- 100241.
Prokopenko, D., et al., *Identification of Novel Alzheimer's Disease Loci Using Sex-Specific Family-Based Association Analysis of Whole-Genome Sequence Data. Sci Rep, 2020. 10(1): p. 5029.
Bhattarai, P., et al., <i>R* Frokopenko, D., et al., *Identification of Novel Alzheimer's Disease Loci Using Sex-Specific Family-*
 Based Association Analysis of Whole-Genome Sequence Data. Sci Rep, 2020. **10**(1): p. 5029.
 Bhattarai, P., et al., Bhattarai, P., et al., *Rare genetic variation in fibronectin 1 (FN1) protects against APOEepsilo*
Alzheimer's disease. Acta Neuropathol, 2024. **147**(1): p. 70.
Lopatko Lindman, K., et al., *PILRA polymorphism modifies the*
- 27. Bhattaraires and Salzheimer's disease. Acta Neuropathol, 2024. 147(1): p. 70.

38. Lopatko Lindman, K., et al., *PILRA polymorphism modifies the effect of APOE4 and GM17 on*
 Alzheimer's disease risk. Sci Rep, 2022.
- 39. Lopather's disease risk. Sci Rep, 2022. 12(1): p. 13264.

Mantyh, W.G., et al., *Early-onset Alzheimer's disease explained by polygenic risk of late-onset

disease? Alzheimers Dement (Amst), 2023. 15(4): p. e12482.

Ma*
- Mantyh, W.G., et al., *Early-onset Alzheimer's disease explained by polygenic risk of late-onset disease?* Alzheimers Dement (Amst), 2023. **15**(4): p. e12482.

40. Malamon, J.S., et al., *A comparative study of structural* disease? Alzheimers Dement (Amst), 2023. **15**(4): p. e12482.

Malamon, J.S., et al., *A comparative study of structural variant calling in WGS from Alzheimer's*

disease families. Life Sci Alliance, 2024. **7**(5).

Le Guen,
-
-
- Malamon, J.S., et al., A comparative study of structural variant calling in WGS from Alzheimer's
disease families. Life Sci Alliance, 2024. **7**(5).
41. Le Guen, Y., et al., Association of Rare APOE Missense Variants V236E Le Guen, Y., et al., Association of Rare APOE Missense Variants V236E and R251G With Risk of
Alzheimer Disease. JAMA Neurol, 2022. **79**(7): p. 652-663.
Le Guen, Y., et al., A novel age-informed approach for genetic associa Alzheimer Disease. JAMA Neurol, 2022. **79**(7): p. 652-663.

42. Le Guen, Y., et al., *A novel age-informed approach for genetic association analysis in Alzheimer's*

disease. Alzheimers Res Ther, 2021. **13**(1): p. 72.

43. Le Guen, Y., et al., A novel age-informed approach for gen
disease. Alzheimers Res Ther, 2021. **13**(1): p. 72.
He, Z., et al., *GhostKnockoff inference empowers identif*
genome-wide association studies. Nat Commun, 2022. Le Guen, Y., et al., *A novel age-informed approach for genetic association analysis in Alzheimer's*
disease. Alzheimers Res Ther, 2021. **13**(1): p. 72.
43. He, Z., et al., *GhostKnockoff inference empowers identification*
- genome-wide association studies. Nat Commun, 2022. 13(1): p. 7209.
Horimoto, A., et al., Admixture mapping implicates 13q33.3 as ancestry-of-origin locus for
Alzheimer disease in Hispanic and Latino populations. HGG Adv, 2
-
- 143. Horimoto, A., et al., *Admixture mapping implicates 13q33.3 as ancestry-of-origin locus for Alzheimer disease in Hispanic and Latino populations. HGG Adv, 2023. 4(3): p. 100207.
45. Lee, S., et al., <i>A comparison betw* genome-wide association superinter mapping implicates 13q33.3 as a
Horimoto, A., et al., Admixture mapping implicates 13q33.3 as a
Alzheimer disease in Hispanic and Latino populations. HGG Adv, 2023.
Lee, S., et al., A com 44. Horimoto, A., et al., *Admixture mapping implicates 13q33.3 as ancestry-of-origin locus for Alzheimer disease in Hispanic and Latino populations. HGG Adv, 2023. 4(3): p. 100207.
45. Lee, S., et al., <i>A comparison betwe* Lee, S., et al., A comparison between similarity matrices for principal component a
assess population stratification in sequenced genetic data sets. Brief Bioinform, 2023. 2
Belloy, M.E., et al., APOE Genotype and Alzheime assess population stratification in sequenced genetic data sets. Brief Bioinform, 2023. 24(1).

46. Belloy, M.E., et al., APOE Genotype and Alzheimer Disease Risk Across Age, Sex, and Population

47. Belloy, M.E., et al., assessive, M.E., et al., *APOE Genotype and Alzheimer Disease Risk Across Age, Sex, and Populcancestry.* JAMA Neurol, 2023. **80**(12): p. 1284-1294.
Belloy, M.E., et al., *Challenges at the APOE locus: a robust quality cont*
- Ancestry. JAMA Neurol, 2023. 80(12): p. 1284-1294.

47. Belloy, M.E., et al., Challenges at the APOE locus: a robust quality control approach for accurate

APOE genotyping. Alzheimers Res Ther, 2022. 14(1): p. 22.

48. Bel Belloy, M.E., et al., *Challenges at the APOE locus: a*
APOE genotyping. Alzheimers Res Ther, 2022. **14**(1):
Belloy, M.E., et al., *A Fast and Robust Strategy to l*
Disease Sequencing Project Data. Neurol Genet, 202.
Li, D
- Belloy, M.E., et al., A Fast and Robust Strategy to Remov
Disease Sequencing Project Data. Neurol Genet, 2022. 8(5)
Li, D., et al., Novel loci for Alzheimer's disease identified b
Ashkenazi Jews. Alzheimers Dement, 2023. 1
- APOE genotyping. Alzheimers Res Ther, 2022. 14(1): p. 22.

48. Belloy, M.E., et al., A Fast and Robust Strategy to Remove Variant-Level Artifacts in Alzheimer

Disease Sequencing Project Data. Neurol Genet, 2022. 8(5): p. Disease Sequencing Project Data. Neurol Genet, 2022. 8(5): p. e200012.

49. Li, D., et al., Novel loci for Alzheimer's disease identified by a genome-wide association study in

4shkenazi Jews. Alzheimers Dement, 2023. 19(1 Disease Sequencing Project Data. Neurol Genet, 2022. **8**(5): p. e200012.
Li, D., et al., Novel loci for Alzheimer's disease identified by a genome-wide association study in
Ashkenazi Jews. Alzheimers Dement, 2023. **19**(12)
- 49. Li, D., et al., *Novel loci for Alzheimer's disease identified by a genome-wide association study in*
Ashkenazi Jews. Alzheimers Dement, 2023. **19**(12): p. 5550-5562.
50. Tejeda, M., et al., *DNA from multiple viral sp* Tejeda, M., et al., DNA from multiple viral species is associated with Alzheimer's disease risk. 50. Tejeda, M., et al., *DNA from multiple viral species is associated with Alzheimer's disease risk.*
Alzheimers Dement, 2024. **20**(1): p. 253-265.
Fig. 253-265.
Alzheimers Dement, 2024. **20(**1): p. 253-265. $A = \{P\}$: p. 253-264. 2024. 2024. 2024. 2024. 2024. 2024. 2024. 2024. 2024. 2024. 2024. 2024. 2024. 2025. 2024. 2025. 2025. 2025. 2025. 2025. 2025. 2025. 2025. 2025. 2025. 2025. 2025. 2025. 2025. 2025. 2025. 2025. 2025.

-
- disease, Alzheimer-related biomarkers and serum metabolites. Alzheimers Res Ther, 2024. 16(1):

p. 234.

Zhu, C., et al., MitoH3: Mitochondrial Haplogroup and Homoplasmic/Heteroplasmic Variant

Calling Pipeline for Alzheim p. 234.
2008) 234.
2008, C., et al., *MitoH3: Mitochondrial Haplogroup and Homoplasmic/Heteroplasmic Variant*
2011 Calling Pipeline for Alzheimer's Disease Sequencing Project. J Alzheimers Dis Rep, 2024. **8**(1): p.
2024. 2 .
Zhu, C.
Calling 52. Zhu, C., et al., MitoH3: Mitochondrial Haplogroup and Homoplasmic/Heteroplasmic Variant

Calling Pipeline for Alzheimer's Disease Sequencing Project. J Alzheimers Dis Rep, 2024. **8**(1): p.

575-587.

All of Us Research
- All of Us Research Program Genomics, I., Genomic data in the All of Us Research Program.
- Calling Pipeline for Alzheimer's Disease Sequencing Project. J Alzheimers Dis Rep, 2024. **8**(1): p.
575-587.
All of Us Research Program Genomics, I., *Genomic data in the All of Us Research Program.*
Nature, 2024. **627**(80 Halldorsson, B.V., et al., *The sequend*
607(7920): p. 732-740.
- Nature, 2024. 627(8003): p. 340-346.

Halldorsson, B.V., et al., The sequences of 150,119 genomes in the UK Biobank. Nature, 2022.

607(7920): p. 732-740.

55. Gogarten, S.M., et al., Genetic association testing using the 608 Contract Co
- 55. Gogarten, S.M., et al., *Genetic association testing using the GENESIS R/Bioconductor package.*

Bioinformatics, 2019. **35**(24): p. 5346-5348.

Gogarten, S.M., et al., *GWASTools: an R/Bioconductor package for quality* For Bioinformatics, 2019. **35**(24): p. 5346-5348.

Gogarten, S.M., et al., GWASTools: an R/Bioconductor package for quality control and analysis of

genome-wide association studies. Bioinformatics, 2012. **28**(24): p. 3329-Bioinformatics, 2019. **35**(24): p. 5346-5348.
Gogarten, S.M., et al., *GWASTools: an R/Bioconductor package for quality control and analysis of*
genome-wide association studies. Bioinformatics, 2012. **28**(24): p. 3329-31
-
- genome-wide association studies. Bioinformatics, 2012. 28(24): p. 3329-31.

57. Zheng, X., et al., A high-performance computing toolset for relatedness and principal component

analysis of SNP data. Bioinformatics, 2012. 2 Sheng, X., et al., *A high-performance computing toolset for relatedness and*
analysis of SNP data. Bioinformatics, 2012. **28**(24): p. 3326-8.
Li, H., *A statistical framework for SNP calling, mutation discovery, assoc*
po 57. Examples of SNP data. Bioinformatics, 2012. 28(24): p. 3326-8.

58. Li, H., A statistical framework for SNP calling, mutation discovery, association mapping and

population genetical parameter estimation from sequencin analysis of Suitsical framework for SNP calling, mutation di

Li, H., A statistical framework for SNP calling, mutation di

population genetical parameter estimation from sequencing

p. 2987-93.

Danecek, P., et al., The v 18. Li, H., A statistical framework for SNP calling, mutation discovery, association mapping and

18. p. 2987-93.

19. 2987-93.

2987-93.

2987-93.

2987-93.

2987-93.

2987-93.

2987-93.

2010. The variant call format and
- Danecek, P., et al., The variant call format and VCFtools. Bioinformatics, 2011. 27(15): p. 2156-8.
-
- p. 2987-93.
Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-8.
Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage*
analyse
- For Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage*

analyses. Am J Hum Genet, 2007. **81**(3): p. 559-75.

Jun, G., et al., *Detecting and estimating contamination of human* Franchine States. Am J Hum Genet, 2007. **81**(3): p. 559-75.

61. Jun, G., et al., *Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet, 2012. 91(5): p.* analyses. Am Detecting and estimating contamination, G., et al., Detecting and estimating contaminand array-based genotype data. Am J Hum Genet, 2008.
Naj, A.C., et al., Quality control and integration of g
genome sequence Fig. 1911. The same of details and array-based genotype data. Am J Hum Genet, 2012. 91(5): p. 839-48.

62. Naj, A.C., et al., Quality control and integration of genotypes from two calling pipelines for whole

genome sequen Naj, A.C., et al., *Quality control and integration of genotypes from two calling pipelines for whole*
genome sequence data in the Alzheimer's disease sequencing project. Genomics, 2019. 111(4): p.
808-818.
McLaren, W., et genome sequence data in the Alzheimer's disease sequencing project. Genomics, 2019. 111(4): p.

808-818.

McLaren, W., et al., *The Ensembl Variant Effect Predictor*. Genome Biol, 2016. 17(1): p. 122.

Chen, X., et al., *M*
-
- Chen, X., et al., Manta: rapid detection of structural variants and indels for germline and cancer 64. Chen, X., et al., *Manta: rapid detection of structural variants and indels for germline and cosequencing applications.* Bioinformatics, 2016. **32**(8): p. 1220-2.
Eggertsson, H.P., et al., *GraphTyper2 enables populati*
- genome sequence data in the Alzheimer's disease sequencing project. Genomics, 2019. **111**(4): p.
808-818.
McLaren, W., et al., *The Ensembl Variant Effect Predictor*. Genome Biol, 2016. **17**(1): p. 122.
Chen, X., et al., sequencing applications. Bioinformatics, 2016. 32(8): p. 1220-2.
65. Eggertsson, H.P., et al., *GraphTyper2 enables population-scale genotyping of structural variation*
using pangenome graphs. Nat Commun, 2019. 10(1): p. 5 Eggertsson, H.P., et al., *GraphTyper2 enables population-scale g*
using pangenome graphs. Nat Commun, 2019. **10**(1): p. 5402.
Wing pangenome graphs. Nat Commun, 2019. **10**(1): p. 5402. $65.$ Eggertas one graphts. Nat Commun, 2019. $10(1)$: p. 5402. using pangenome graphs. Nat Commun, 2019. 10(1): p. 5402.