RESEARCH ARTICLE

NeuroBooster Array: A Genome-Wide Genotyping Platform to Study Neurological Disorders Across Diverse Populations

Sara Bandres-Ciga, PhD,[1*] Faraz Faghri, PhD,[1,2] Elisa Majounie, PhD,[3] Mathew J. Koretsky, BSc,[1] Jeffrey Kim, BSc,[1,4] Kristin S. Levine, MSc,[1,2] Hampton Leonard, MSc,[1,2] Mary B. Makarious, PhD,[4,5] Hirotaka Iwaki, MD,[1,2] Peter Wild Crea, BSc,[1,4] Dena G. Hernandez, PhD,[4] Sampath Arepalli, BSc,[4] Kimberley Billingsley, PhD,[1,4] Katja Lohmann, PhD,[6] Christine Klein, MD, PhD,[6] Steven J. Lubbe, PhD,[7,8] Edwin Jabbari, MD, PhD,[5] Paula Saffie-Awad, MD,[9,10,11] Derek Narendra, MD, PhD,[12] Armando Reyes-Palomares, PhD,[13] John P. Quinn, PhD,[14] Claudia Schulte, PhD,[15] Huw R. Morris, MD, PhD,[5,16] Bryan J. Traynor, MD, PhD,[17,18] Sonja W. Scholz, MD, PhD,[19] Henry Houlden, MD, PhD,[16,20] John Hardy, PhD,[16,21] Sonya Dumanis, PhD,[16] Ekemini Riley, PhD,[16] Cornelis Blauwendraat, PhD,[1,4] Andrew Singleton, PhD,[1,4] Mike Nalls, PhD,[1,2*] Janina Jeff, PhD,[3] and Dan Vitale, MSc,[1,2*] on behalf of the Global Parkinson's Genetics Program (GP2) and the Center for Alzheimer's and Related Dementias (CARD)

[1]*Center for Alzheimer's and Related Dementias, National Institute on Aging and National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland, USA*
[2]*Data Tecnica, Washington, District of Columbia, USA*
[3]*Illumina Inc, San Diego, California, USA*
[4]*Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, Maryland, USA*
[5]*Department of Clinical and Movement Neurosciences, Queen Square Institute of Neurology, University College London, London, United Kingdom*
[6]*Institute of Neurogenetics, University of Lübeck, Lübeck, Germany*
[7]*Ken and Ruth Davee Department of Neurology, Northwestern University, Feinberg School of Medicine, Chicago, Illinois, USA*
[8]*Simpson Querrey Center for Neurogenetics, Northwestern University, Feinberg School of Medicine, Chicago, Illinois, USA*
[9]*Programa de Pós-Graduação em Ciências Médicas, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil*
[10]*Centro de Trastornos del Movimiento, Santiago, Chile*
[11]*Clínica Santa María, Santiago, Chile*
[12]*Inherited Movement Disorders Unit, Neurogenetics Branch, Division of Intramural Research, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland, USA*
[13]*Department of Molecular Biology and Biochemistry, Faculty of Sciences, University of Málaga, Málaga, Spain*
[14]*Department of Pharmacology & Therapeutics, University of Liverpool, Liverpool, United Kingdom*
[15]*Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tuebingen and German Center for Neurodegenerative Diseases, University of Tuebingen, Tuebingen, Germany*
[16]*Aligning Science Across Parkinson's Collaborative Research Network, Chevy Chase, Maryland, USA*
[17]*Department of Neurology, Johns Hopkins University Medical Center, Baltimore, Maryland, USA*

***Correspondence to:** Dan Vitale, Dr. Mike Nalls, and Dr. Sara Bandres-Ciga, Center for Alzheimer's Disease and Related Dementias, National Institute on Aging, National Institutes of Health, 9000 Rockville Pike, NIH Building T44, Bethesda, MD 20892, USA; E-mails: E-mail: dan@datatecnica.com (D.V.), E-mail: mike@datatecnica.com (M.N.), and E-mail: sara.bandresciga@nih.gov (S.B.-C.)

GP2 members, including their names and affiliations, are provided within Supporting Information Table S9.

Funding agencies: This research was supported by the GP2 and the Intramural Research Program at the National Institute on Aging, National Institutes of Health, Department of Health and Human Services (projects ZO1 AG000535 and ZIA AG000949), as well as the National Institute of Neurological Disorders and Stroke (program ZIA NS003154) and the National Human Genome Research Institute. GP2 was supported by the Aligning Science Across Parkinson's initiative and implemented by The Michael J. Fox Foundation for Parkinson's Research through grant MJFF-009421/17483.

Relevant conflicts of interest/financial disclosures: D.V., F.F., H.L., H.I., K.S.L., and M.N. declare that they are consultants employed by Data Tecnica International, whose participation in this is part of a consulting agreement between the National Institutes of Health and said company. M.N. is also an advisor to Neuron23 Inc and Character Biosciences. S.W.S. serves on the Scientific Advisory Council of the Lewy Body Dementia Association and the Multiple System Atrophy Coalition. S.W.S. and B.J.T. receive research support from Cerevel Therapeutics. H.R.M. is employed by UCL and reports paid consultancy from Roche, Aprinoia, AI Therapeutics, and Amylyx; lecture fees/honoraria from BMJ, Kyowa Kirin, and Movement Disorders Society; research grants from Parkinson's UK, Cure Parkinson's Trust, PSP Association, Medical Research Council, and The Michael J. Fox Foundation; and is a coapplicant on a patent application related to C9ORF72—method for diagnosing a neurodegenerative disease (PCT/GB2012/052140). C.K. is a Medical Advisor to Centogene and Retromer Therapeutics and received speakers' honoraria from Desitin and Bial.

Received: 28 March 2024; **Revised:** 28 May 2024; **Accepted:** 6 June 2024

Published online 16 September 2024 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/mds.29902

[18]Neuromuscular Diseases Research Section, Laboratory of Neurogenetics, National Institute on Aging, Bethesda, Maryland, USA
[19]Neurodegenerative Diseases Research Unit, National Institute of Neurological Disorders and Stroke, Bethesda, Maryland, USA
[20]Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London, United Kingdom
[21]Department of Neurodegenerative Disease, UCL Queen Square Institute of Neurology, London, United Kingdom

ABSTRACT: Background: Commercial genome-wide genotyping arrays have historically neglected coverage of genetic variation across populations.
Objective: We aimed to create a multi-ancestry genome-wide array that would include a wide range of neuro-specific genetic content to facilitate genetic research in neurological disorders across multiple ancestral groups, fostering diversity and inclusivity in research studies.
Methods: We developed the Illumina NeuroBooster Array (NBA), a custom high-throughput and cost-effective platform on a backbone of 1,914,934 variants from the Infinium Global Diversity Array and added custom content comprising 95,273 variants associated with more than 70 neurological conditions or traits, and we further tested its performance on more than 2000 patient samples. This novel platform includes approximately 10,000 tagging variants to facilitate imputation and analyses of neurodegenerative disease–related genome-wide association study loci across diverse populations.
Results: In this article, we describe NBA's potential as an efficient means for researchers to assess known and novel disease genetic associations in a multi-ancestry framework. The NBA can identify rare genetic variants and accurately impute more than 15 million common variants across populations. Apart from enabling sample prioritization for further whole-genome sequencing studies, we envisage that NBA will play a pivotal role in recruitment for interventional studies in the precision medicine space.
Conclusions: From a broader perspective, the NBA serves as a promising means to foster collaborative research endeavors in the field of neurological disorders worldwide. Ultimately, this carefully designed tool is poised to make a substantial contribution to uncovering the genetic etiology underlying these debilitating conditions.

Key Words: Centre for Alzheimer's and Related Dementias; diversity; genetic screening; genotyping; Global Parkinson's Genetics Program; NeuroBooster array; neurological diseases

Historically, commercial genome-wide genotyping arrays have been designed by selecting tag single nucleotide polymorphisms (SNPs) (SNPs in a region of the genome with high linkage disequilibrium that represent a haplotype) from European or Asian populations. Little attention has been paid to designing arrays that can globally cover genetic variation across populations. The impetus behind our endeavor was to develop a comprehensive genome-wide array capable of encompassing a broad spectrum of neuro-specific content. This array serves the objectives of two key initiatives: the Global Parkinson's Genetics Program (GP2) and the Center for Alzheimer's and Related Dementias (CARD).

Much has been done to unravel the genetic landscape of brain disorders, identifying both causal and genetic contributors underlying disease risk and progression in a wide range of neurological diseases. To this end, there is a growing understanding of the importance of globally representative and diverse genetics.[1] Despite these recent efforts, there is an unmet need to address: The vast majority of genetic studies have focused on populations of European ancestry.[2,3] The genetic architecture of other populations, such as Africans, Asians, and Latinos, remains largely unexplored in neurology and other areas of research with only a few studies being published.[4-9]

In this article, we present the NeuroBooster Array v.1.0 (NBA), a cost-efficient platform to perform ancestry-inclusive genome-wide association studies (GWAS) and low frequency variant detection. The NBA represents a valuable screening tool to rapidly identify risk variants and disease modifiers that contribute to neurological conditions across ancestries, as well as known causal pathogenic variants with almost eight times the coverage of previous platforms such as NeuroChip[10] and NeuroX arrays.[11] The NBA is a reliable platform to accelerate participant enrollment for target-specific clinical trials according to underlying genetics, as well as sample prioritization for more costly approaches, such as short and long whole-genome sequencing (WGS).

In this article, we describe the NBA custom design, due diligence, and content. Furthermore, we discuss coverage and imputation accuracy for common and rare variation in the context of previous arrays (NeuroChip[10] and NeuroX[11]). We comprehensively evaluate the NBA's performance by assessing SNP imputation via mean imputed r^2 and minor allele frequency (MAF) across 1000 Genomes populations data, including American Admixed (AMR), African Admixed (AAC), Africans (AFR), East Asians (EAS), South Asians (SAS), and Europeans (EUR), using several population reference panels, such as the TOPMed Imputation Server (https://imputation.biodatacatalyst.nhlbi.nih.gov), the Haplotype Reference Consortium r1.1 2016 (http://www.haplotype-reference-consortium.org), the Genome Asia Pilot Project (GAsP) (https://genomeasia100k.org/), and the Consortium on Asthma among African-Ancestry Populations in the Americas (CAAPA) Panel https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001123.v2.p1).

Finally, we discuss the benefits and limitations of using this platform to study polygenic inheritance in neurological conditions.

Materials and Methods

NBA Design

The NBA contains a backbone of 1,914,934 genetic markers from the Infinium Global Diversity Array-8 v1.0 (GDA) complemented with custom content of 95,273 disease-associated variants involved in a wide range of neurological conditions. The custom content on the NBA was purposely designed to complement the base content of the GDA series of arrays from Illumina (Table 1).

The GDA is built on a high-density SNP global backbone optimized for cross-population imputation coverage of the genome (https://www.illumina.com/products). The GDA contains a robust genome-wide scaffold designed to tag both common and low-frequency variants (MAF > 1%) in global populations. This scaffold was designed through collaborations with the CAAPA and the Population Architecture using Genomics and Epidemiology study (https://www.pagestudy.org/). For GDA design, more than 400,000 markers of exome content were gathered from 36,000 individuals of diverse ethnic groups, including AAC, Hispanics, Pacific Islanders, EAS, and individuals of mixed ancestry. The GDA includes content not found in the 1000 Genomes Project, with more than 1000 whole-genome sequences of AFR ancestry and populations throughout the Americas, including the United States, Caribbean, and Latin and South America. The GDA is designed to contain diverse exonic content from the Genome Aggregation Database version 2.1 (https://gnomad.broadinstitute.org/news/2018-10-gnomad-v2-1/, RRID: SCR_014964), including both cross-population and population-specific markers with either functionality or strong evidence for association with certain populations (*Illumina Inc.*).

We designed the custom content for the NBA with four goals in mind: (1) identify rare coding variants associated with disease of interest to researchers in the neurology field, (2) improve imputation of known GWAS loci for neurodegenerative diseases across populations of diverse ancestries, (3) generally improve imputation quality across populations of diverse ancestries to facilitate risk locus discovery, and (4) facilitate fine mapping of risk loci.

To accomplish these aims, we aggregated custom content from a variety of neurodegenerative disease-relevant

TABLE 1 *Coverage differences between NeuroX, NeuroChip, and NeuroBooster arrays*

Variant Description	NeuroBooster	NeuroChip	Comparison with NeuroChip (+)	NeuroX	Comparison with NeuroX (+)
Total variants (pre–quality control)	2,010,207	486,137	1,524,070	267,607	1,742,600
Backbone	1,914,934	306,670	1,608,264	242,901	1,672,033
Custom-content variants	95,273	179,467	−84,194	24,706	70,567
Indels	49,554	16,259	33,295	200	49,354
Autosomal variants	1,873,290	473,442	1,399,848	261,477	1,611,813
Coding variants	1,223,002	88,560	1,134,442	226,104	996,898
Sex chromosomal variants	79,994	11,840	68,154	5906	74,088
Mitochondrial variants	1509	160	1349	219	1290
Variants with MAF < 0.05[a]	853,885	227,448	626,437	219,093	634,792
Variants with MAF < 0.001[a]	148,821	154,953	−6132	179,500	−30,679

[a]Based on annotations.

sources. In brief, building the array was carried out in five phases (described further later), consisting of sequential design components. The first design component was a systematic review and consultation with key opinion leaders in the field of neurology to identify putative rare variants associated with disease. The second design component consisted of annotating known GWAS risk loci and identifying multi-population tag SNPs to support improved imputation and locus tagging across ancestry groups in follow-up studies. Based on this information, the third component consisted of saturation of genotyping of risk loci from large GWAS regions. Next, we carried out probe design due diligence in collaboration with *Illumina Inc*. Finally, we used diverse samples and rare-variant-enriched cohorts to build a custom cluster file to improve genotype quality. The culmination of these phases of work allowed us to densely genotype and impute higher variant coverage into regions of interest related to neurological diseases.

Systematic Review of Publicly Annotated Databases and Key Opinion Leaders

Whole-exome sequencing and WGS variant prioritization analyses were conducted in families and cohort studies. This includes relevant datasets from diverse ancestries and population isolates, provided by International Parkinson's Disease Genomics Consortium (https://pdgenetics.org) members and collaborators, and used as the first layer of derived content nominated by key opinion leaders. We followed this up with systematic reviews of publicly annotated databases, including the Human Gene Mutation Database (HGMD) version 2019 (https://www.hgmd.cf.ac.uk/ac/index.php, RRID: SCR_001621)[12] and the Genomics England panel (accessed in May 2019) (https://panelapp.genomicsengland.co.uk/).

From HGMD, we extracted any nucleotide substitutions (missense/nonsense, regulatory, splicing) and also variants labeled as nucleotide deletions, insertions, and indels for any disease or phenotype combinations suggested by key opinion leaders as potentially being implicated in the etiology of neurological diseases or conditions presenting with neurological complications. Our exhaustive list of search terms is highlighted in Table S1. The database cross-referencing resulted in nearly 90,000 candidate variants initially. Due to array design and space concerns on the product, the variant list from this first portion of the design was reduced to 54,907 variants prioritizing pathogenic and possibly pathogenic annotations prior to Illumina due diligence (described later; Table S5).

We also included an allocation of 1809 polygenic risk score (PRS)-related variants for Parkinson's disease (PD) risk nominated by PRSice-2 Version 2.2.13[13] (https://choishingwan.github.io/PRSice/, DOI: 10.5281/zenodo.3703335, RRID: SCR_017057). This includes variants that did not reach genome-wide significance but positively impacted PRS modeling efforts in previous reports. For more details on PRS construction, refer to Nalls et al.[14]

Identification of Multi-ancestry Tag SNPs from GWAS Loci

Loci identified as genome-wide significant in large-scale GWAS meta-analyses of neurodegenerative diseases[5,14-20] were the next target of our design process. Genome-wide significant loci from eight neurodegenerative disease-related conditions, including Alzheimer's disease, PD, amyotrophic lateral sclerosis, multiple sclerosis, progressive supranuclear palsy, corticobasal degeneration, multiple system atrophy, frontotemporal dementia, and dementia with Lewy bodies were included with either multiple proxies for the top SNP at every locus or technical replicates when proxies were not available. GWAS tagging SNPs per locus were defined as any SNP reaching a genome-wide significant P value and correlated at $r^2 < 0.3$ with any other significant SNPs within 250 kb. This process was done for each disease of interest based on the combined European reference series from 1000 Genomes phase 1v5 data (https://www.internationalgenome.org/data-portal/data-collection/phase-1, RRID: SCR_006828) after filtering all variants for "PASS" annotation and minor allele count (MAC) of at least three. This led to the successful inclusion of ~183 GWAS-related variants (Table S2).

The 1000 Genomes sequence data were again categorized into ancestry reference "sub-populations" to build multipopulation references. These included conglomerate datasets from population labels as follows: MXL (Mexican Ancestry), CLM (Colombian), PEL (Peruvian), and PUR (Puerto Rican) were combined to form the AMR ancestry group; JPT (Japanese), CDX (Chinese Dai in Xishuangbanna), CHB (Han Chinese in Beijing), CHS (Han Chinese South), KHV (Kinh in Ho Chi Minh City, Vietnam), and CHD (Han Chinese in Beijing) were combined to form the EAS ancestry group; TSI (Toscani in Italy), IBS (Iberian Populations in Spain), GBR (British from England and Scotland), and CEU (Central European) were combined to form the EUR ancestry group; PJL (Punjabi in Lahore, Pakistan), ITU (Indian Telugu in the United Kingdom), STU (Sri Lankan Tamil in the United Kingdom), GIH (Gujarati Indians in Houston, TX, USA), and BEB (Bengali in Bangladesh) were combined to form the SAS ancestry group; GWD (Gambian in Western Division, Mandinka), MSL (Mende in Sierra Leone), ESN (Esan in Nigeria), YRI (Yoruba in Ibadan, Nigeria), LWK (Luhya in Webuye, Kenya), GWJ (Gambian Jola), GWF (Gambian Fula), and GWW (Gambian Wolof) were combined to form the AFR ancestry group; ASW (African Ancestry in SW

USA) and ACB (African Caribbean in Barbados) were combined to form the AAC ancestry group. Due to sample size issues for tagging analysis, a total of 115 samples from Finland were excluded from the references, leaving us with 3395 reference samples across all groups. Based on these reference samples and the 183 signals GWAS loci discussed earlier, the software TagIt version 1.0.8 (https://bioinformaticshome.com/tools/imputation/descriptions/TagIt.html)[21] was run using default settings per locus stratified by ancestry group to identify diverse tag SNP series ($r^2 > 0.3$).

Saturation Genotyping of Known Risk Loci

For each of the 183 GWAS variants of interest tagging unique loci, we identified upstream and downstream tag SNPs across all ancestry groups. For each region, all tag SNPs from diverse populations were included. To fill out content after the Illumina design triage described later in the due diligence content section, we added a surplus of EUR ancestry tag SNPs per locus. This phase of design included more than 10,000 SNPs tagging diverse ancestral linkage structure at all loci. Regions tagged by the multipopulation variants of interest covered more than 38 MB of the genome.

NBA Due Diligence Content

Illumina design scoring was undertaken at *Illumina Inc*. to screen potential array content. This included scoring of variants for quality and their ability to be included on arrays. Variants with problematic probes that could not reliably produce high genotyping scoring based on internal Illumina metrics for design quality were not preselected for the production phase. In addition, after clustering 2793 diverse samples from the GP2[22] (https://gp2.org/training/) across multiple populations (Table S3), we reviewed variants that were not accurately genotyped, with high rates of missingness across batches resulting in the exclusion of 49,590 variants (Table S4). After iterative design scoring at *Illumina Inc*., custom content for 95,273 variants (and the additional 1,914,934 standard content variants) was annotated using the Variant Effect Predictor Release 100 and made available as part of the standard series of GP2 releases (https://useast.ensembl.org/info/docs/tools/vep/index.html).

NBA Data Processing and Custom Clustering

Automated genotype data processing was conducted on GenoTools, a Python pipeline built for quality control and ancestry estimation of data. Additional details can be found online at: https://github.com/GP2code/GenoTools. Quality control was performed according to standard protocols. Samples with a call rate of less than 95%, sex mismatches, or high heterozygosity (estimated by an |F| statistics of >0.25) were excluded from analyses. Further quality-control measures included the removal of SNPs with missingness greater than 5%, variants with significant deviations from Hardy–Weinberg equilibrium ($P < 1E{-}4$), variants with nonrandom missingness, and variants with missing data patterns (haplotype at $P \leq 1E{-}4$ per ancestry). A total of 2793 diverse case samples across six ancestry groups were clustered using the Illumina's GenomeStudio Software v2.0.5 package (https://support.illumina.com/downloads/genomestudio-2-0.html, RRID: SCR_010973) (Fig. 1, Table S3).

The clustering file is available for download via the GP2 GitHub repository (https://github.com/GP2code; DOI: 10.5281/zenodo.7904832, release 5; https://gp2.org/). Clusters can be viewed using the GP2 data browser. Among the 2793 samples, a total of 2373 patients with PD were included in addition to 420 Gaucher disease cases to capture variants of interest in the *GBA1* gene.

Clustering was done following *Illumina Inc*. guidelines. Specifically, the robustness and reliability of genotype clustering was ensured by implementing a meticulous quality-control protocol on the genotyping data. The protocol started with an evaluation of BeadChip performance in GenomeStudio using the Controls Dashboard. Any BeadChips with profiles suggestive of technical issues were promptly excluded from the analysis to negate the influence of technical inaccuracies on the results. Next, an assessment of sample call rates was conducted using the Samples Table. Samples presenting call rates less than 0.98 were excluded from the subsequent analyses to uphold the standard of genotyping data.

Following the assessment of call rates, an in-depth review of the Copy Number Metrics report was carried out to detect samples with outlier values for 'LogRDev' or 'BAlleleDev.' Such samples, suggestive of potential genotyping anomalies, were excluded from further investigation to ensure a dataset free from major genotyping errors and anomalies. By employing this stringent sample inclusion/exclusion protocol, we assured that the subsequent genotype clustering was based on high-quality samples free from technical errors. This approach ensures that the genotyping data used for clustering are accurate, reliable, and most importantly, trustworthy.

Passing samples were then reclustered using the default *Illumina Inc*. cluster algorithm built into GenomeStudio, with autosomes clustered together and sex chromosomes done separately.

NBA Genotyping Protocol

An overview of the protocol can be visualized in Figure S1. In brief, for each sample a total of 200 ng of high-quality genomic DNA is amplified and

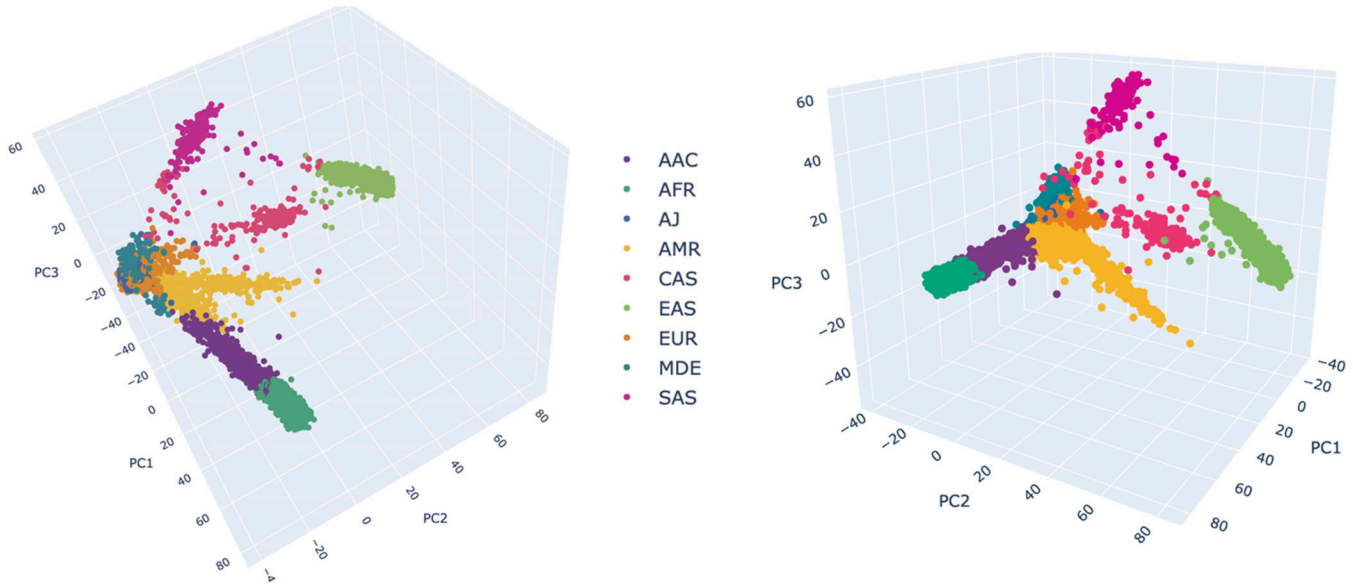

FIG. 1. Ancestry prediction clustering for samples genotyped on the NeuroBooster Array. Three-dimensional principal-component analysis to group individuals based on their genetic makeup. A total of 2793 samples from the Global Parkinson's Genetics Program were included, including 2373 Parkinson's disease cases and 420 Gaucher disease cases. Each point represents a sample, and the colors depict the ancestral background as shown in the color legend: Europe (EUR) = orange, East Asian (EAS) = lemon green, American Admixed (AMR) = yellow, Ashkenazi Jewish (AJ) = lapis blue, African (AFR) = teal blue, African admixed (AAC) = purple, South Asian (SAS) = magenta, Central Asian (CAS) = dark pink, and Middle East (MDE) = cerulean blue. [Color figure can be viewed at wileyonlinelibrary.com]

enzymatically fragmented. The resulting fragments are alcohol precipitated and resuspended. We used Illumina-provided RA1 resuspension buffer to resuspend the dried DNA pellets by alkaline lysis method. This buffer is often a basic (pH 8.0) Tris solution, which helps to denature the DNA. This is an ideal condition for subsequent lysis and ethylenediaminetetraacetic acid that binds divalent cations destabilizing the membrane and inhibiting DNase. Next, the fragmented DNA solution is hybridized to the NBA using a Tecan Freedom EVO liquid-handling robot (Tecan, Research Triangle Park, NC, USA). After hybridization, automated allele-specific, enzymatic base extension and fluorophore staining are performed. The stained genotyping arrays are washed, sealed, and vacuum dried prior to scanning them on the Illumina iScan system. Raw data files are imported into GenomeStudio (version 2.0, Genotyping Module; *Illumina Inc.*) using

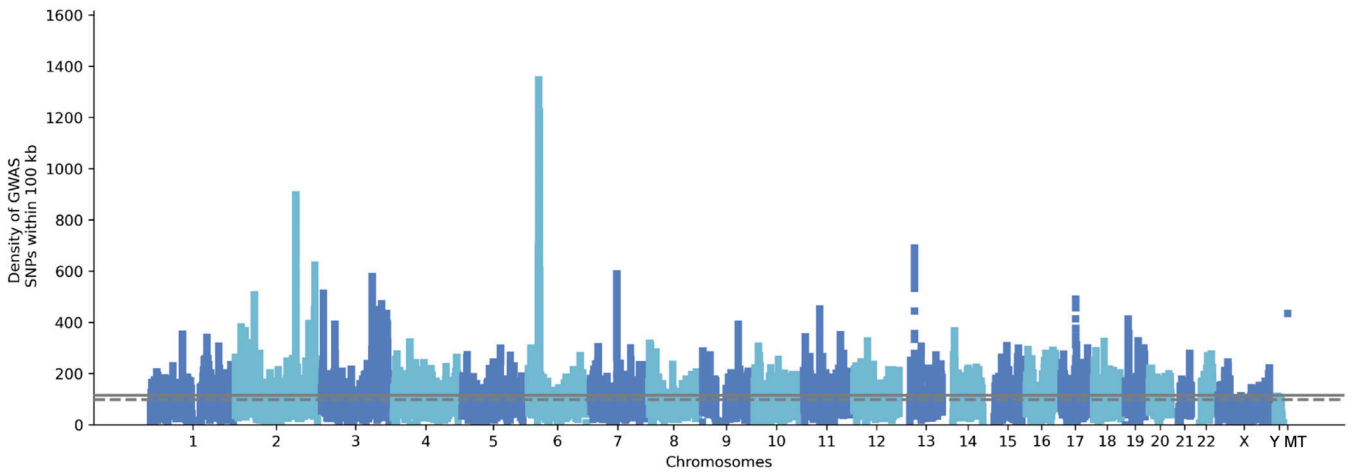

FIG. 2. Brisbane plot showing the genomic density of SNPs on Global Parkinson's Genetics Program raw genotyped data generated on the NeuroBooster Array. In a Brisbane plot, the x-axis typically represents the genomic position along the chromosomes, similar to a Manhattan plot. However, instead of displaying the degree of significance (−log10 *P* values), the *y*-axis represents the density of genetic variants within a specified genomic region (eg, within 100-Kb intervals). This distinction allows researchers to visualize the distribution of genetic variants across the genome and identify regions of high- or low-variant density tagged by the array. The gray solid line represents the mean coverage, and the gray dashed line represents the mean coverage per 100 Kb. GWAS, genome-wide association study. [Color figure can be viewed at wileyonlinelibrary.com]

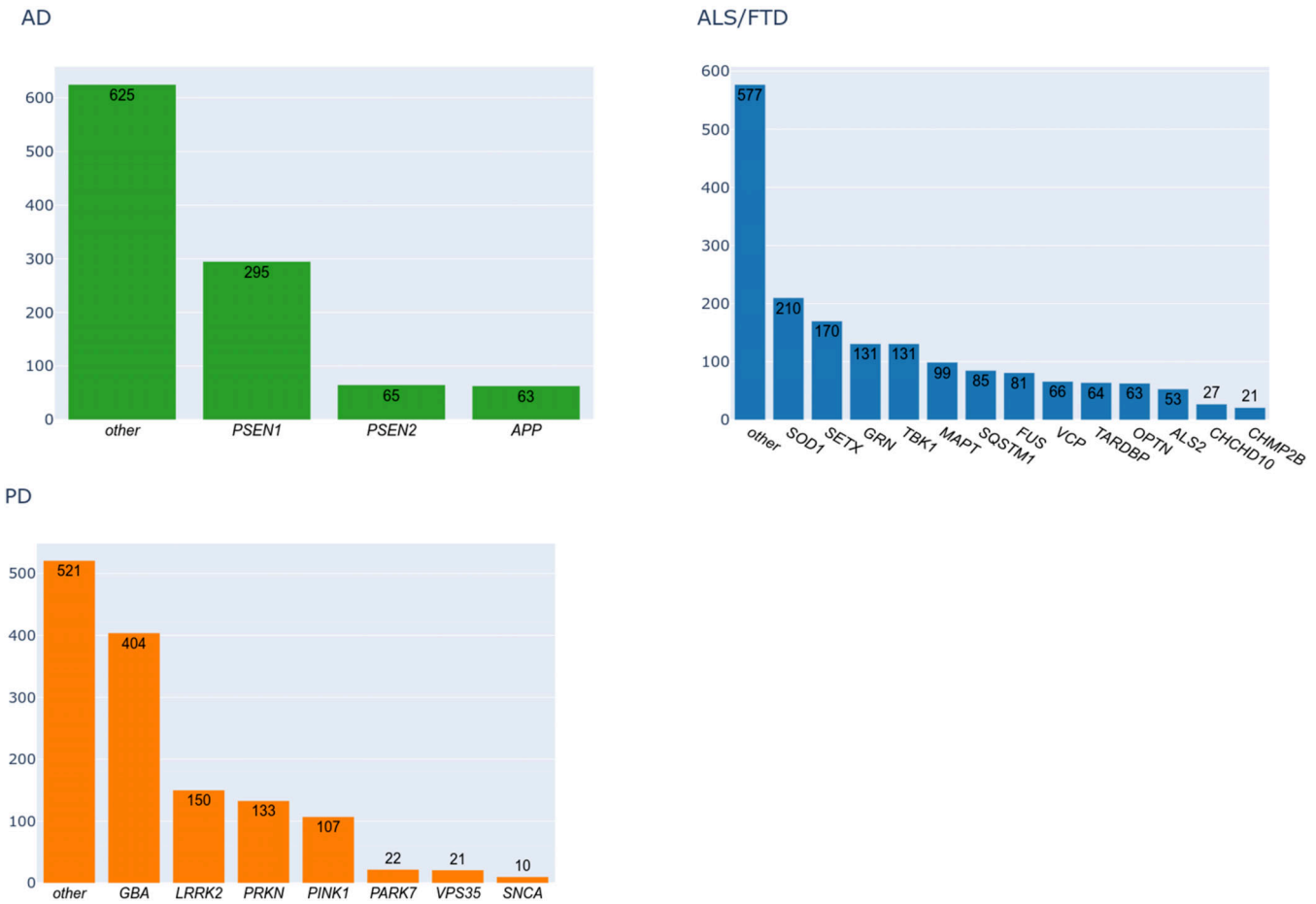

FIG. 3. Overview of the number of Human Gene Mutation Database disease-associated variants that are present on the NeuroBooster Array for the most prevalent neurodegenerative diseases. AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; FTD, frontotemporal dementia; PD, Parkinson's disease. [Color figure can be viewed at wileyonlinelibrary.com]

a custom-generated sample sheet, and genotypes are called using a GenCall threshold of 0.15.

Results

NeuroBooster Content Overview

In total, the NBA can detect 1,873,290 autosomal variants, 79,994 sex chromosomal variants (X and Y chromosomes), and 1509 mitochondrial variants (Fig. 2). The overlap between the NBA and NeuroChip is n = 126,220 variants. The NBA includes more than 10,000 multi-ancestry GWAS locus tagging variants to facilitate imputation and analyses of these neurodegenerative disease-related GWAS loci across diverse populations. Detailed pathogenic inferences and annotated NBA functionally variant content for the array in total are provided through GP2 (see https://github.com/GP2code/NeuroBoosterArray).[22] Figures 3 and 4 and Table S5 display the abundance of disease-associated variants per gene covered across the most prevalent neurodegenerative diseases based on initial systematic review and cross-referencing of the HGMD and Genomics England database content.

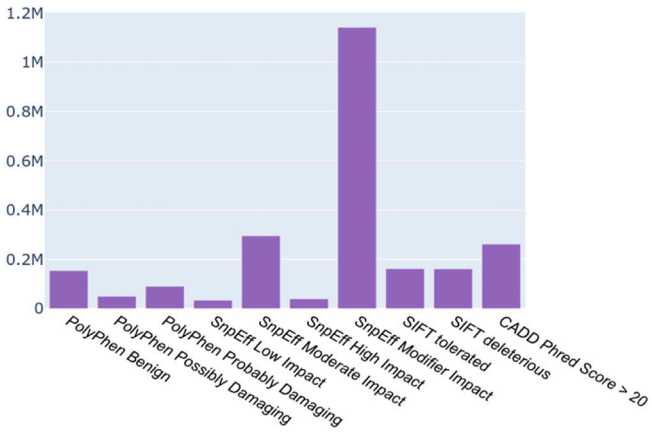

FIG. 4. Overview of NeuroBooster array content by pathogenicity predictors. [Color figure can be viewed at wileyonlinelibrary.com]

Genotyping Accuracy

GenTrain scores were calculated for all NeuroBooster variants using GenomeStudio version 2.0. The GenTrain score is a statistical metric based on the shapes of the different allelic clusters and their relative distance to each

other (*Illumina Inc*). Typically, GenTrain scores greater than 0.7 are considered high-quality genotypes. The mean (and SD) for the array is 0.83 (0.091) (Fig. S2). Due to improved technology from *Illumina Inc*. compared with NeuroChip or NeuroX arrays over the past decade, as well as the iterative design process the array underwent in collaboration with *Illumina Inc*., we are able to rescue many variants not previously able to be genotyped (see Fig. S3 for example comparison).

Genetic variants in *APOE*, *APP*, *GBA1*, *LRRK2*, *PARK2*, *DJ-1*, *PINK1*, *PRKN*, *PSEN1*, *SNCA*, *TREM2*, and *VPS35* contribute to Alzheimer's disease, Lewy body disease, and PD etiology among other neurodegenerative diseases. Genotyping of some of these genes remains complicated because of high GC content (*APOE*) or the presence of a pseudogene (*GBA1*). We comprehensively assessed variant content for tagging and imputed variants across multiple ancestries (additional details can be found in Tables S6 and S8).

Quality-Control Metrics for Imputation Accuracy and Allele Frequency Intervals Across Diverse Populations

Variants with a MAF < 0.005 and Hardy–Weinberg equilibrium $P < 1E{-}5$ were excluded before submission to the TOPMed Imputation server. The utilized TOPMed reference panel version, encompasses genetic information from 97,256 reference samples and more than 300 million genetic variants across the 22 autosomes and the X chromosome. As of October 2022, the TOPMed panel included approximately 180,000 participants, with 29% of AFR ancestry, 19% of Latino/admixed American ancestry, 8% of Asian ancestry, and 40% of EUR ancestry (https://topmed.nhlbi.nih.gov/). Further details about the TOPMed Study, Imputation Server, and Minimac Imputation can be accessed online (https://topmed.nhlbi.nih.gov/, https://imputation.biodatacatalyst.nhlbi.nih.gov/#! [RRID: SCR_009292], https://genome.sph.umich.edu/w/index.php?title=Minimac:_Tutorial&oldid=14562 [RRID: SCR_009292]). After imputation, the resulting files underwent pruning based on an MAC threshold of 10 and an imputation r^2 value of 0.3.

Imputation of autosomal variants was performed across 2793 individuals in 10 different ancestries (Table S7). In its application to recent GP2 dataset releases (https://gp2.org/), imputation yields were solid in comparison with previous genotyping arrays. After filtering for imputation quality ($r^2 > 0.3$ and MAC > 10) across ancestry groups, minimum variant yields using TOPMed imputation included 7.4 M to 19.6 M variants in total across ancestry groups. Lower-frequency variants at MAF < 5% were also well imputed in this data resource, with yields after filtering ranging from 0.8 M to 7.1 M across populations.

In addition to data imputation using TOPMed, we next assessed imputation accuracy performance by extracting the variant content available on the NBA from WGS data from six ancestral populations present on 1000 Genomes (AAC, AFR, AMR, EAS, EUR, and SAS). Optimal array design depends not only on tag SNP selection but also on empirical evaluation of imputation performance.[20] We compared imputation performance for neurodegenerative disease–related GWAS loci with and without the custom NBA content. The custom content was designed to enrich or "saturate" the NBA genotyping platform with variants representative of these disease-related loci. The goal of this comparison was to understand whether loci saturation in the platform would improve imputation for PD loci specifically. For each of the population-specific GWAS scaffolds, imputation accuracy was assessed by MAF bins versus various imputation reference panels, including the Haplotype Reference Consortium, CAAPA, and Genome Asia Pilot (Fig. S4A–H). We observed a slight improvement in imputation metrics for number of variants and imputation quality within PD known loci when including the NBA custom content in the imputation pipeline.

Discussion

Genetics research needs to expand from European to non-European populations. By filling this critical disparity, we will build a comprehensive picture of neurological diseases in the global population. The NBA represents a cost-effective and powerful tool for researchers to test and validate genetic associations in a multi-ancestry context, delivering insights for both discovery and screening applications. This effort demonstrates the power of modern genomic technologies to rapidly screen large sample collections and to unravel the genetic architecture underlying diverse neurodegenerative diseases in a multi-ancestry context.

The availability of data from diverse populations generated through NBA can directly lead not only to the identification and replication of identified GWAS loci but also help to narrow the search window within loci that had been previously discovered in the European population. The genome-wide assessment of human populations is useful in providing insights on the evolutionary history of the human genome, with a special emphasis on refining genetic associations and disease fine-mapping. Genotyping combined with imputation greatly enhances the ability to fine-map risk loci by increasing the density and coverage of genetic variants, thereby improving the precision and power of genetic studies aimed at identifying disease-associated variants. The NBA incorporates ~1.9 million genetic variants, significantly enhancing coverage compared with previous

platforms (see Table 1), with a specific focus on population-specific markers critical for transethnic fine-mapping analyses. Given the genetic diversity that exists across populations, an accurate screening array represents a valuable resource to provide insights into genetic determinants underlying disease risk and progression, modifiers of disease, potential differences in the heritable component of disease between groups, and the generation of population-specific genetic risk profiles. The goal is ultimately to drive transformational progress in our understanding of the genetic architecture of neurological diseases in a global context that can be of benefit to patients in all populations.

A crucial aspect is the use of NBA to generate predictive information that can play a significant role in improving the design of clinical trials. This may enable the support of trials even in preclinical participants, allowing mechanistic stratification of disease and adjustments in trial outcomes based on the personalized predictions of disease for each individual. Integrating NBA-generated data with clinical information enhances the pool of potential trial participants available for trial enrollment, particularly with regard to precision medicine efforts including trials for genetically derived interventions. By cross-referencing genetic profiles with clinical characteristics and disease histories, researchers can identify and recruit suitable candidates who meet the trial's genetic criteria. Educating potential participants about the importance of genetic research and clinical trial participation fosters informed decision-making. Clear communication about the benefits, risks, and implications of participating in genetics-based trials encourages engagement and enhances recruitment efforts. In this regard, ensuring compliance with ethical guidelines and regulatory requirements is crucial. Protecting participant privacy and adhering to data security protocols are essential for maintaining trust and confidentiality throughout the recruitment process. In addition, much has been discussed about PRS in clinical settings. PRSs aggregate the effects of multiple genetic variants across the genome to estimate an individual's genetic predisposition to certain diseases. Although PRSs can indicate an increased genetic risk for certain diseases, they do not provide a definitive prediction. The predictive power of PRS can vary widely depending on the disease and the ancestry reference population. Furthermore, interpreting these scores in a meaningful way for individual patients in a clinical setting remains challenging, because environmental and lifestyle factors also play significant roles in disease development.

It is essential to recognize the limitations inherent in this technology. Like all genotyping arrays, NBA does not detect DNA variants that were unknown prior to its design in 2019, as well as ultra-rare variants. It also cannot genotype variants in complex genomic regions (such as those involving pseudogenes) or identify repeat expansions or small structural variants because of the challenges in designing reliable probes. Our custom content is based on known disease-related genes, and we assume the limitation that novel genetic disease contributors may be missing in this first version of the array. In summary, the NBA is limited to detecting only the preselected and imputed variants, missing novel or rare variants that could be detected by sequencing technologies. Unlike WGS, the array does not provide information on the entire genome, which limits the ability to discover new genetic markers outside of the targeted regions. WGS captures variation more broadly, including both common and rare variants, structural variants (long-read WGS), and novel mutations; however, it is costly. It is recommended that researchers use the array for initial large-scale screening and then apply low-pass WGS for deeper exploration of novel variants. Although much of the genetic information contained on the array pertains to the autosomes, it is important to consider the unique genetic characteristics and implications associated with the sex chromosomes (X and Y). Imputing genetic data for the sex chromosomes relies heavily on reference panels that may be less comprehensive compared with those available for autosomes. This can affect the accuracy and reliability of imputed genotypes for sex chromosome variants. In addition, variants on the X and Y chromosomes can show greater variability across different populations. In general, genotyping arrays may not capture the full diversity of these variants, leading to reduced accuracy in certain ethnic or ancestral groups. Addressing these challenges often requires specialized methods and technologies tailored to the unique characteristics of the sex chromosomes.

In conclusion, we describe the design and implementation of the NBA, which offers more comprehensive and improved content compared with its predecessor platforms. The NBA serves as an invaluable asset in our quest to comprehend neurological diseases within a worldwide framework, particularly as we embark on the era of precision therapeutics. As we continue to advance in genomic research and discover new genetic loci associated with neurodegenerative diseases, we fully anticipate updating the NBA platform with new content. Our commitment to enhancing the array's utility includes regularly incorporating the latest genetic insights to ensure researchers have access to the most comprehensive and up-to-date tools for genetic analysis.

Author Contributions

Conception and design of the study: F.F., E.M., M.N., and J.J. Acquisition and analysis of data: S.B.-C., F.F., E.M., M.J.K., J.K., K.S.L., H.L., M.B.M., H.I., P.W.C., D.G.H., S.A., K.B., K.L., C.K., S.J.L., E.J., P.S.-A., D.N., A.R.-P., J.P.Q., C.S., H.R.M., B.J.T., S.W.S., H.H., J.H., S.D., E.R., C.B., A.S., M.N., J.J., and D.V. Drafting a significant portion of the manuscript or

figures: S.B.-C. and D.V. GP2 members, including their names and affiliations, are provided within Supporting Information Table S9.

Acknowledgments: This research was supported by the GP2 and the Intramural Research Program at the National Institute on Aging, National Institutes of Health, Department of Health and Human Services (projects ZO1 AG000535 and ZIA AG000949), as well as the National Institute of Neurological Disorders and Stroke (program ZIA NS003154) and the National Human Genome Research Institute. Data used in the preparation of this article were obtained from the GP2. GP2 was supported by the Aligning Science Across Parkinson's initiative and implemented by The Michael J. Fox Foundation for Parkinson's Research (for a complete list of GP2 members, see https://gp2.org). Additional funding was provided by The Michael J. Fox Foundation for Parkinson's Research through grant MJFF-009421/17483.

Competing Interests

DV, FF, HLL HI, KSL, and MAN declare that they are consultants employed by Data Tecnica International, whose participation in this is part of a consulting agreement between the US National Institutes of Health and said company. MAN also an advisor to Neuron23 Inc and Character Biosciences. SWS serves on the Scientific Advisory Council of the Lewy Body Dementia Association and the Multiple System Atrophy Coalition. S.W.S. and B.J.T. receive research support from Cerevel Therapeutics. HRM is employed by UCL. In the last 12 months he reports paid consultancy from Roche, Aprinoia, AI Therapeutics and Amylyx; lecture fees/honoraria - BMJ, Kyowa Kirin, Movement Disorders Society. Research Grants from Parkinson's UK, Cure Parkinson's Trust, PSP Association, Medical Research Council, Michael J Fox Foundation. Dr Morris is a co-applicant on a patent application related to C9ORF72 - Method for diagnosing a neurodegenerative disease (PCT/GB2012/052140). Dr. Christine Klein is a Medical Advisor to Centogene and Retromer Therapeutics and Speakers' honoraria from Desitin and Bial.

Data Availability Statement

The data that support the findings of this study are openly available in (DOI 10.5281/zenodo.7904832, release 5). Data used in the preparation of this article were obtained from the Global Parkinson's Genetics Program (GP2) and can be accessed at amp-pd.org.

References

1. Khani M, Cerquera-Cleves C, Kekenadze M, et al. Towards a global view of Parkinson's disease genetics. Ann Neurol 2024;95(5):831–842.
2. Fatumo S, Chikowore T, Choudhury A, et al. A roadmap to increase diversity in genomic studies. Nat Med 2022;28(2):243–250.
3. Vollstedt E-J, Schaake S, Lohmann K, et al. Embracing monogenic Parkinson's disease: the MJFF global genetic PD cohort. Mov Disord 2023;38(2):286–303.
4. Rizig M, Bandres-Ciga S, Makarious MB, et al. Identification of genetic risk loci and causal insights associated with Parkinson's disease in African and African admixed populations: a genome-wide association study [Internet]. Lancet Neurol 2023;22. https://doi.org/10.1016/S1474-4422(23)00283-1
5. Kunkle BW, Grenier-Boley B, Sims R, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nat Genet 2019;51(3):414–430.
6. Lake J, Warly Solsberg C, Kim JJ, et al. Multi-ancestry meta-analysis and fine-mapping in Alzheimer's disease [Internet]. Mol Psychiatry 2023;28. https://doi.org/10.1038/s41380-023-02089-w
7. Kim JJ, Vitale D, Otani DV, et al. Multi-ancestry genome-wide meta-analysis in Parkinson's disease [Internet]. medRxiv 2022;56. https://www.nature.com/articles/s41588-023-01584-8
8. Loesch DP, Horimoto ARVR, Heilbron K, et al. Characterizing the genetic architecture of Parkinson's disease in Latinos. Ann Neurol 2021;90(3):353–365.
9. Foo JN, Tan LC, Irwan ID, et al. Genome-wide association study of Parkinson's disease in east Asians. Hum Mol Genet 2017;26(1):226–232.
10. Blauwendraat C, Faghri F, Pihlstrom L, et al. NeuroChip, an updated version of the NeuroX genotyping platform to rapidly screen for variants associated with neurological diseases. Neurobiol Aging 2017;57:247.e9–247.e13.
11. Nalls MA, Bras J, Hernandez DG, et al. NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases. Neurobiol Aging 2015;36(3):1605.e7–1605.e12.
12. Stenson PD, Mort M, Ball EV, et al. The human gene mutation database (HGMD): optimizing its use in a clinical diagnostic or research setting. Hum Genet 2020;139(10):1197–1207.
13. Choi SW, O'Reilly PF. PRSice-2: polygenic risk score software for biobank-scale data [Internet]. Gigascience 2019;8(7). https://academic.oup.com/gigascience/article/8/7/giz082/5532407
14. Nalls MA, Blauwendraat C, Vallerga CL, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet Neurol 2019;18(12):1091–1102.
15. Iwaki H, Blauwendraat C, Leonard HL, et al. Genomewide association study of Parkinson's disease clinical biomarkers in 12 longitudinal patients' cohorts. Mov Disord 2019;34(12):1839–1850.
16. Nicolas A, Kenna KP, Renton AE, et al. Genome-wide analyses identify KIF5A as a novel ALS gene. Neuron 2018;97(6):1268–1283.e6.
17. Jansen IE, Savage JE, Watanabe K, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat Genet 2019;51(3):404–413.
18. Guerreiro R, Escott-Price V, Hernandez DG, et al. Heritability and genetic variance of dementia with Lewy bodies. Neurobiol Dis 2019;127:492–501.
19. Höglinger GU, Melhem NM, Dickson DW, et al. Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. Nat Genet 2011;43(7):699–705.
20. Ferrari R, Hernandez DG, Nalls MA, et al. Frontotemporal dementia and its subtypes: a genome-wide association study. Lancet Neurol 2014;13(7):686–699.
21. Wojcik GL, Fuchsberger C, Taliun D, et al. Imputation-aware tag SNP selection to improve power for large-scale, multi-ethnic association studies. G3 2018;8(10):3255–3267.
22. GP2. The global Parkinson's genetics program. Mov Disord 2021;36(4):842–851.

Supporting Data

Additional Supporting Information may be found in the online version of this article at the publisher's web-site.