RESEARCH ARTICLE

Accelerating Medicines Partnership: Parkinson's Disease. Genetic Resource

Hirotaka Iwaki, MD,[1,2,3] Hampton L. Leonard, MS,[1,2,3] Mary B. Makarious, BS,[3] Matt Bookman, MS,[4] Barry Landin, BS,[5] David Vismer, BS,[5] Bradford Casey, PhD,[6] J. Raphael Gibbs, PhD,[3] Dena G. Hernandez, PhD,[3] Cornelis Blauwendraat, PhD,[3] Daniel Vitale, MS,[1,2,3] Yeajin Song, MS,[1,2,3] Dinesh Kumar, PhD,[7] Clifton L. Dalgard, PhD,[8,9] Mahdiar Sadeghi, MS,[7,10] Xianjun Dong, PhD,[11] Leonie Misquitta, PhD,[12] Sonja W. Scholz, PhD,[13,14] Clemens R. Scherzer, MD,[11] Mike A. Nalls, PhD,[1,2,3] Shameek Biswas, PhD,[15] Andrew B. Singleton, PhD,[2,3*] Uniformed Services University of the Health Sciences Associates, AMP PD Whole Genome Sequencing Working Group, and AMP PD consortium

[1]*Data Tecnica International, Glen Echo, Maryland, USA*
[2]*Center for Alzheimer's and Related Dementias, National Institute on Aging, Bethesda, Maryland, USA*
[3]*Laboratory of Neurogenetics, National Institute on Aging, Bethesda, Maryland, USA*
[4]*Verily Life Sciences, San Jose, California, USA*
[5]*Technome, Herndon, Virginia, USA*
[6]*The Michael J. Fox Foundation for Parkinson's Research, New York, New York, USA*
[7]*Sanofi, Framingham, Massachusetts, USA*
[8]*Department of Anatomy, Physiology & Genetics, Uniformed Services University of the Health Sciences, Bethesda, Maryland, USA*
[9]*The American Genome Center, Uniformed Services University of the Health Sciences, Bethesda, Maryland, USA*
[10]*Northeastern University, Boston, Massachusetts, USA*
[11]*Harvard Medical School, Brigham and Women's Hospital, Boston, Massachusetts, USA*
[12]*Publicis Sapient, Bethesda, Maryland, USA*
[13]*National Institute of Neurological Disorders and Stroke, Bethesda, Maryland, USA*
[14]*Department of Neurology, Johns Hopkins University, Baltimore, Maryland, USA*
[15]*Bristol Myers Squibb, Seattle, Washington, USA*

ABSTRACT: Background: Whole-genome sequencing data are available from several large studies across a variety of diseases and traits. However, massive storage and computation resources are required to use these data, and to achieve sufficient power for discoveries, harmonization of multiple cohorts is critical.

Objectives: The Accelerating Medicines Partnership Parkinson's Disease program has developed a research platform for Parkinson's disease (PD) that integrates the storage and analysis of whole-genome sequencing data, RNA expression data, and clinical data, harmonized across multiple cohort studies.

Methods: The version 1 release contains whole-genome sequencing data derived from 3941 participants from 4 cohorts. Samples underwent joint genotyping by the TOPMed Freeze 9 Variant Calling Pipeline. We performed descriptive analyses of these whole-genome sequencing

***Correspondence to:** Andrew B. Singleton, Center for Alzheimer's and Related Dementias, Chief, Molecular Genetics Section, NIA IRP NIH, Building 35, Room 1A1014, 35 Convent Drive, Bethesda, MD 20892, USA; E-mail: andrewsingleton987@gmail.com

Relevant conflicts of interest/financial disclosures: Andrew Singleton received grants from the Michael J. Fox Foundation for Parkinson's Research and the Department of Defense NETPR Program (grant IAA-XAG16001-001-00000). Sonja W. Scholz received support from the Intramural Research Program of the National Institute of Health (National Institute on Aging, National Institute of Neurological Disorders and Stroke; project numbers 1ZIAAG000935, 1ZIANS003154).

Funding agencies: Support for this study came from Accelerating Medicines Partnership in Parkinson's Disease (AMP PD), a public-private partnership managed by the FNIH and funded by Celgene, GSK, the Michael J. Fox Foundation for Parkinson's Research, the National Institute of Neurological Disorders and Stroke, Pfizer, Sanofi, and Verily. This work was supported in part by the Intramural Research Program of the National Institute on Aging, National Institutes of Health, Department of Health and Human Services project ZO1 AG000949 and grants from the Department of Defense (IAA-XAG16001-001-00000) and the Michael J. Fox Foundation for Parkinson's Research. This research was supported in part by the Intramural Research Program of the National Institutes of Health (National Institute on Aging, National Institute of Neurological Disorders and Stroke project numbers 1ZIAAG000935 and 1ZIANS003154).

Received: 4 December 2020; **Revised:** 20 January 2021; **Accepted:** 11 February 2021

Published online 7 May 2021 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/mds.28549

data using the Accelerating Medicines Partnership Parkinson's Disease platform.

Results: The clinical diagnosis of participants in version 1 release includes 2005 idiopathic PD patients, 963 healthy controls, 64 prodromal subjects, 62 clinically diagnosed PD subjects without evidence of dopamine deficit, and 705 participants of genetically enriched cohorts carrying PD risk-associated *GBA* variants or *LRRK2* variants, of whom 304 were affected. We did not observe significant enrichment of pathogenic variants in the idiopathic PD group, but the polygenic risk score was higher in PD both in nongenetically enriched cohorts and genetically enriched cohorts. The population analysis showed a correlation between genetically enriched cohorts and Ashkenazi Jewish ancestry.

Conclusions: We describe the genetic component of the Accelerating Medicines Partnership Parkinson's Disease platform, a solution to democratize data access and analysis for the PD research community.

Key Words: Parkinson's disease; genetics; clinical; open science

The genetic investigation of Parkinson's disease (PD) has been a driving force in PD research over the last 20 years. Genetics serves to identify a starting point for the molecular and cellular processes that underlie disease. More recently, genetics has become part of an array of data types being used in an attempt to define disease at the individual level, with the aim of predicting who will get disease,[1] when they will get it,[2] and what their progression will look like.[3] Ultimately genetics is a foundational part of the science that promises to reveal rational and viable targets for therapeutic intervention and to highlight the patients most suitable for each interventional strategy.[4,5]

This work has been both enabled and accelerated by the rapid development and adoption of methods for the generation and analysis of massive-scale genetic data. The application of genome-wide genotyping and whole-genome sequencing (WGS) has significantly altered the speed and potential of genetic research in PD, resulting in the rapid identification of genetic variability linked to disease. A critical challenge to the effective use of these data centers on data scale, production, and sharing. Genetics is expensive, can be challenging to analyze in a uniform way and is often difficult to effectively share, because of both practical and regulatory reasons. Addressing these challenges promises to reduce duplicated effort, accelerate discovery, and democratize research.

A part of the Accelerating Medicines Partnership Parkinson's Disease (AMP PD) project is centered on the development and deployment of a knowledge platform. It will present varied data relevant to PD. A large component of these data comes in the form of WGS that has been or will be generated across the Parkinson's Progression Markers Initiative (PPMI), the Parkinson's Disease Biomarkers Project (PDBP), the Harvard Biomarkers Study (HBS), BioFIND, the Study of Urate Elevation in Parkinson's Disease trial (SURE-PD), and the Safety, Tolerability and Efficacy Assessment of Dynacirc CR in Parkinson Disease (STEADY-PD) trial. As of the time of writing, the AMP PD platform contains complete WGS data on 3941 individuals from these studies.

The knowledge platform provides users with access to the genetic data and a space in which to perform analyses in situ, without download. The flexible nature of the underlying Google Cloud Platform architecture affords users the ability to quickly deploy computing resources to analyze genome-scale data. This platform also enables the user to deploy workflows that incorporate other data modalities, including phenotypic and transcriptomic data sets.

Here, we describe the data generation, processing, and quality control of these genetic data. We also provide the user with access to the workflows and pipelines that were used to perform these analyses, which can be copied or modified by the user. We provide a summary of the genetic characterization of these samples and provide corresponding annotated code to execute such analyses.

Methods

Cohorts

Release 1 (AMP PD v1_release) included 4 multicenter observational studies: BioFIND (https://biofind.loni.usc.edu), the HBS[5,6] (https://www.bwhparkinsoncenter.org/biobank), PDBP (https://pdbp.ninds.nih.gov), and PPMI (https://www.ppmi-info.org). Participants' clinical information and genetic samples were obtained under appropriate written consent and with local institutional and ethical approvals. The details of these studies can be obtained from the AMP PD website (https://amp-pd.org) and each study website. The data from SURE-PD and STEADY-PD are being processed for the next release.

Data Flow Overview

The sample quality control steps and the released data are outlined in Figure 1. AMP PD requires quality control checks for all release-bound data at the sequencing facility first to ensure that minimum quality

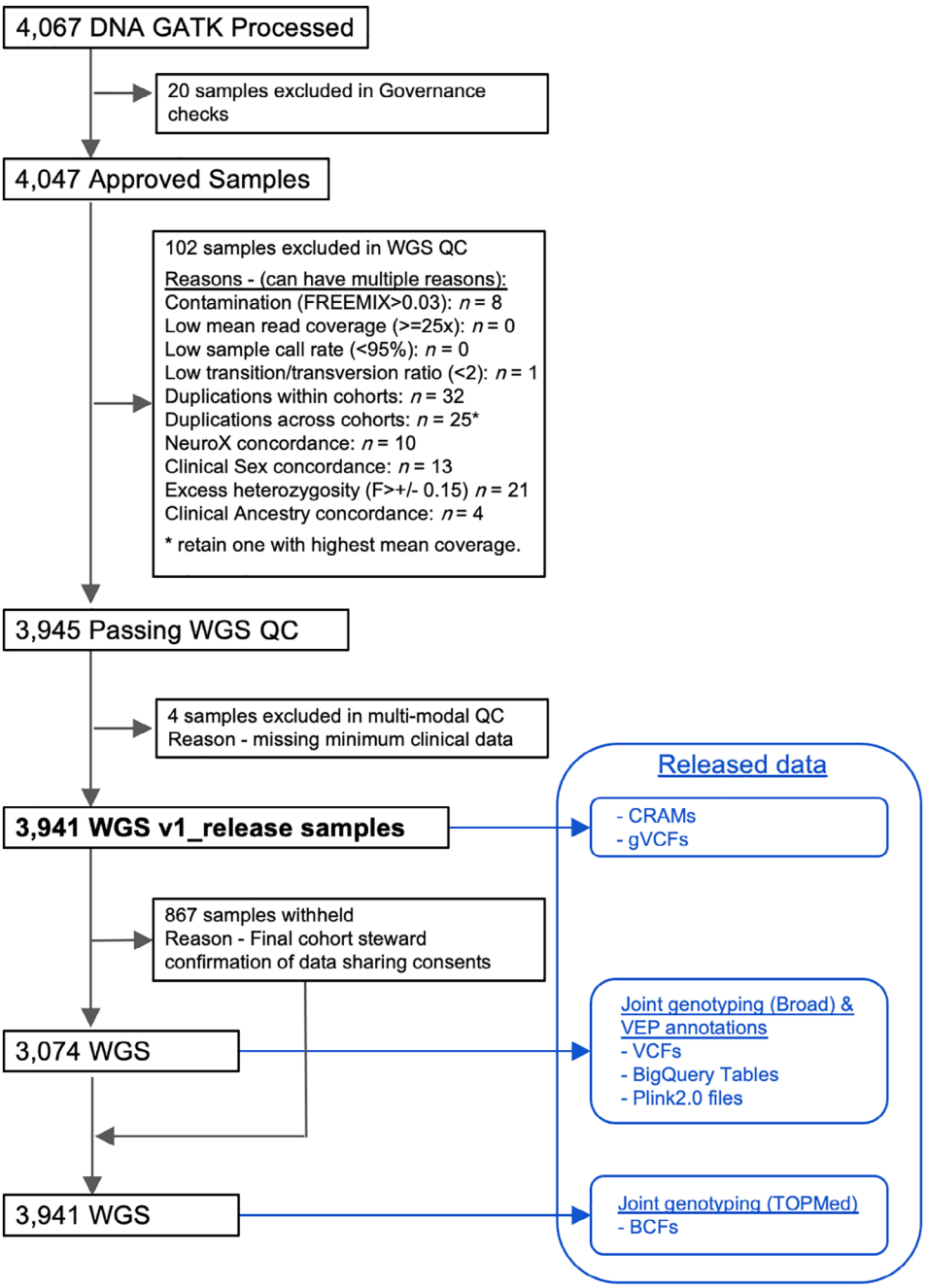

FIG. 1. WGS sample flowchart. WGS, whole-genome sequencing; QC, quality control. [Color figure can be viewed at wileyonlinelibrary.com]

controls are met for individual samples before being transferred to AMP PD. For the flagship AMP PD data release, version 1, all WGS samples (n = 4067) were processed from fastq and vetted through a series of independent genomic quality control (QC) checks and interdependent multimodal QC checks. Samples passing all QC checks were processed using the Broad joint discovery pipeline and annotated with Variant Effect Predictor (VEP; n = 3074)[6] or TOPMed variant calling pipeline (n = 3941; https://www.nhlbiwgs.org). For each type of QC test, a plan was created through a collaboration of the AMP PD WGS working group and contributors from the National Institutes of Health (NIH)/National Institute on Aging (NIA)/Laboratory of Neurogenetics (LNG). These QC tests are outlined below, and the scripts are shared with AMP PD users in the AMP PD workspace (https://app.terra.bio/#workspaces/amp-pd-release-v1/Getting%20Started%20Tier%202%20-%20Clinical%20and%20Omics%20Access). During QC test execution, failing samples were noted against each discrete quality control check, so that each test result could be evaluated independently. This approach enabled the Transcriptomics working group and Clinical Data Harmonization

working group to consider the impact of each WGS QC check on their constituent QC processes. The overall QC results for the AMP PD release could be interpreted without ambiguity about which QC check resulted in the exclusion of a participant sample, whether the exclusion arose from a QC test in the WGS, Transcriptomic, or Clinical Data Harmonization working group.

DNA Sequencing and Variants Calling

DNA samples were derived from the participants' whole-blood specimens and sequenced using Illumina HiSeq X Ten platform at Macrogen Corp. or the Uniformed Services University of Health Sciences. Paired-end 300–400 bp reads were processed in accordance with the functional equivalence pipeline[7] implemented at the Broad Institute. Alignment and variant calling were against the GRCh38DH reference genome.

WGS Quality Controls

AMP PD v1_release quality control tests included Governance checks that required contributing cohort stewards to identify participants that could be released by the AMP PD program in a formal artifact, the Subject Master List. Each cohort steward affirmed that the identifiers to be used as the basis for AMP PD participant identifiers were free of personal identifiable information (PII) and could not be deconstructed or reconstructed to reveal PII. Cohort stewards from contributing studies consented to modification of these identifiers to allow AMP PD to adapt them to a naming convention that was constructed and agreed to by participating members of the AMP PD working groups. We used 2 alphabets as a cohort identifier with a following 4-digit number to distinguish the participants. (eg, BF-0011). This uniform naming convention was then adopted by each AMP PD working group and threaded through all data types to achieve a uniform representation of the participant in all filenames and file contents, across WGS, Transcriptomics, and Clinical Data records. The Governance QC check enabled the release of consistently named data and confirmed consented cohort participants.

The WGS working group prepared a plan for testing WGS samples for contamination, quality, duplicates across studies, duplicates within studies, and concordance with clinical and with preexisting NeuroX genotyping array platform data.[8] These tests were broken into discrete QC checks that were defined in great detail, documented, and executed by contributing experts from the NIH/NIA/LNG. The complete analysis resulted in a recommendation to the WGS Working Group for each QC check about whether AMP PD should exclude a sample from the release, include the sample but withhold from joint genotyping results, or flag the sample as problematic for downstream consideration by the investigating end user. QC checks for duplicates resulted in exclusion of all but 1 sample from joint genotyping, whereby the sample of higher mean coverage was selected for inclusion, and all other samples were identified as duplicates in a release artifact and queryable table. Samples that passed WGS QC were further evaluated against other AMP PD data types. Although WGS QC test results primarily informed downstream Transcriptomics tests, clinical data QC tests were bidirectional. As the Clinical Data Harmonization group defined criteria for minimum clinical data, participant records were thus excluded during clinical QC, resulting in the exclusion of WGS samples (n = 4). The clinical data QC test for sample data asserts that no WGS or Transcriptomics data can be released without matching clinical participant data.

Joint Genotyping

The first set of Joint Genotyped variants consisted of 3074 samples and was published by AMP PD in November 2019. Joint Genotyping was run on Terra and used the Broad Institute's joint discovery pipeline (workflow and fixed inputs can be found on GitHub, https://github.com/amp-pd/amp-pd-workflows). The Joint Genotyped VCF files were then run through the VEP using the annotations feature of the Variant Transforms tool from Google Cloud (https://github.com/googlegenomics/gcp-variant-transforms). The VEP database used was version 91 of homoserines, GRCh38. The 3074 Joint Genotyped and annotated variants were made available in 4 forms: per-chromosome gzipped VCFs, Plink 1.9 files, Plink 2.0 files, and as a table in Google BigQuery. The VCFs were loaded to BigQuery using the vcf_to_bq command of Variant Transforms.

More recently, we published all 3941 samples in the release version 1 jointly genotyped by TOPMed Freeze 9 variant calling pipeline (webpage under preparation; previous versions were described at https://www.nhlbiwgs.org/data-sets). The AMP PD samples were combined with 143,415 samples sequenced in the NHLBI TOPMed program, 60,540 samples sequenced in the National Human Genomics Research Institute(NHGRI) Centers for Common Disease Genomics program, 15,042 sequenced samples from NIA- National Institute for Neurological Disorders and Stroke studies, and 2504 samples from the 1000 Genomes Project phase 3, deeply sequenced by the New York Genome Center. The genotypes for only the AMP PD samples were returned to AMP PD. Variant functional annotation is provided from snpEff 4.3 t (build 2017-11-24 10:18),[9] using the GRCh38.86 database. Statistically

phased haplotypes using Eagle 2.4 (December 13, 2017)[10] will be provided when they are ready.

Descriptive Analysis

We provide a descriptive analysis of baseline characteristics and of sequencing metrics. We summarized the carrier status of ClinVar "pathogenic" variants[11] for autosomal-dominant PD genes. To determine "pathogenic" variants, we applied 2 criteria. One derived a variant only annotated as "pathogenic," whereas the other included a wider set of variants that had at least 1 annotation such as "likely_pathogenic" or "pathogenic" among multiple annotations (pathogenic+). For autosomal-recessive genes, we also considered loss-of-function variants (LoF). The LoF variants were defined as having "HIGH" impact consequences determined by VEP annotation that includes transcript ablation, splice acceptor variant, splice donor variant, stop gained frameshift variant, stop lost, start lost, and transcript amplification.[6]

The population structure of the participants was analyzed using HapMap samples of European, Asian, and African continental ancestry.[12] We merged the study data with these referencing data and conducted a principal components analysis. Each continental-level ancestry was determined by mean ± 6 standard deviations from the reference panel. We also referenced genotyping array data from GSE23636 at Gene Expression Omnibus to identify the Ashkenazi Jewish population in the study.[13] For participants of European descent, we calculated the polygenic risk score (PRS) using the weights of 90 significant variants from the recent meta-analysis of PD GWAS[1] and conducted a descriptive analysis of PRS scores per study arm.

Data Availability

All data processing was conducted on the Google Cloud Platform. Processing/analysis scripts were provided at the related workspaces for reference (accessible for AMP PD users). The resulting CRAM files, VCFs, and jointly genotyped data (BCF, VCF, PLINK, and BigQuery format) are available through the AMP PD.

Results

Table 1 shows the baseline characteristics and the sample-level sequencing quality metrics. Among 3941 participants, there were 2005 participants with idiopathic PD and 963 controls from idiopathic case–control cohorts. Seven hundred five participants were from the genetically enriched cohorts (the genetic cohort or the genetic registry of PPMI), of whom 304 were affected, and the rest were unaffected. These PPMI genetically enriched cohorts are individuals who are specifically recruited for their genetic status and

TABLE 1. Whole-genome sequenced participants

	Overall	BioFIND	HBS	PDBP	PPMI
Total, n	3941	172	867	1469	1433
Sex and age					
Female, n (%)	1725 (43.8)	71 (41.3)	372 (42.9)	640 (43.6)	642 (44.8)
Age at baseline (years), mean (SD)	63.5 (10.7)	67.1 (6.9)	66.1 (10.1)	64.0 (10.0)	61.1 (11.7)
Self-reported race					
White, n (%)	3726 (94.6)	161 (93.6)	844 (97.3)	1397 (95.1)	1324 (92.5)
Mixed ancestry, n (%)	65 (1.6)	2 (1.2)	2 (0.2)	6 (0.4)	55 (3.8)
Black or African American, n (%)	63 (1.6)	3 (1.7)	10 (1.2)	32 (2.2)	18 (1.3)
Asian, n (%)	34 (0.9)	1 (0.6)	7 (0.8)	16 (1.1)	10 (0.7)
Study arms					
Parkinson's disease, n (%)	2005 (51.1)	99 (57.6)	640 (73.8)	858 (58.9)	408 (28.5)
Healthy control, n (%)	963 (24.5)	73 (42.4)	227 (26.2)	470 (32.3)	193 (13.5)
Genetic cohort PD, n (%)	179 (4.6)				179 (12.5)
Genetic cohort unaffected, n (%)	222 (5.7)				222 (15.5)
Genetic registry PD, n (%)	125 (3.2)				125 (8.7)
Genetic registry unaffected, n (%)	179 (4.6)				179 (12.5)
Prodromal, n (%)	64 (1.6)				64 (4.5)
SWEDD, n (%)	62 (1.6)				62 (4.3)
Disease control, n (%)	127 (3.2)			127 (8.7)	
Variant metrics					
MEAN_COVERAGE, median (Q1, Q3)	33.9 (31.2, 36.2)	35.0 (34.1, 35.7)	33.4 (30.7, 36.3)	33.3 (30.8, 36.3)	34.2 (31.8, 36.4)
MEDIAN_COVERAGE, median (Q1, Q3)	34.0 (32.0, 37.0)	35.0 (35.0, 36.0)	34.0 (31.0, 37.0)	34.0 (31.0, 37.0)	35.0 (32.0, 37.0)
READS/K, median (Q1, Q3)	3519 (3254, 3760)	3629 (3552, 3690)	3489 (3198, 3765)	3446 (3194, 3742)	3565 (3322, 3789)
AVG_DP, median (Q1, Q3)	35.2 (32.6, 37.6)	36.3 (35.5, 36.9)	34.9 (32.0, 37.7)	34.5 (31.9, 37.4)	35.7 (33.2, 37.9)
FREEMIX, median (Q1, Q3)	**0.0 (0.0, 0.0)**	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)

AVG_DP, average depth; PD, Parkinson's disease; SWEDD, scan without evidence of dopamine deficit.

TABLE 2. Pathogenic/LoF variant distribution of known PD genes

Genes	Category queried	Number of variants in the category	Among all participants, n = 3074 carriers (Homo, Multi)	Among cases in Nongenetically enriched cohorts, n = 1365 carriers (Homo, Multi)	Among controls in Nongenetically enriched cohorts n = 736 carriers (Homo, Multi)
ATP13A2	Pathogenic and LoF	4	4 (0, 0)	1 (0, 0)	3 (0, 0)
	Pathogenic+ and LoF	4	4 (0, 0)	1 (0, 0)	3 (0, 0)
GBA	Pathogenic	3	8 (0, 0)	6 (0, 0)	2 (0, 0)
	Pathogenic+	6	300 (8, 0)	30 (1, 0)	5 (0, 0)
LRRK2	Pathogenic	5	467 (7, 0)	30 (1, 0)	9 (0, 0)
	Pathogenic+	5	467 (7, 0)	30 (1, 0)	9 (0, 0)
PINK1	Pathogenic and LoF	2	2 (0, 0)	1 (0, 0)	1 (0, 0)
	Pathogenic+ and LoF	2	2 (0, 0)	1 (0, 0)	1 (0, 0)
PLA2G6	Pathogenic and LoF	5	9 (0, 0)	3 (0, 0)	3 (0, 0)
	Pathogenic+ and LoF	5	9 (0, 0)	3 (0, 0)	3 (0, 0)
PRKN	Pathogenic and LoF	7	204 (1, 2)	82 (1, 1)	56 (0, 0)
	Pathogenic+ and LoF	10	214 (1, 4)	88 (1, 1)	58 (0, 1)
SNCA	Pathogenic	1	27 (0, 0)	0 (0, 0)	0 (0, 0)
	Pathogenic+	1	27 (0, 0)	0 (0, 0)	0 (0, 0)
VPS35	Pathogenic	2	3 (0, 0)	2 (0, 0)	1 (0, 0)
	Pathogenic+	2	3 (0, 0)	2 (0, 0)	1 (0, 0)

Homo, homozygous; Multi, multiple variants carriers; pathogenic, "pathogenic" in ClinVar; pathogenic+, clinical significance containing "pathogenic" in ClinVar; LoF, IMPACT "HIGH" in VEP annotation.
Queried but no variants in the model categories for FBXO7 and PARK7 (DJ-1).

include carriers of *LRRK2* p.G2019S, p.R1441C/G, *GBA* p.N370S, p.L444P, 84GG, and *SNCA* p.A53T confirmed by Sanger Sequencing. Mutations status was concordant in the WGS data, with the exception of 4 GBA heterozygous mutation carriers (3 *GBA* p. L444P and 1 p.N370S carriers). Unfortunately, the *GBA* p.L444P variant was not reliably called because of failing the variant quality control process used in preparing the WGS data. Other study arms included participants with prodromal symptoms (n = 64), PD subjects without evidence of dopamine deficit (SWEDDs; n = 62) and disease controls (patients with other neurological diseases, n = 127). The sequencing metrics were compatible with recent genetic studies with the median/mean coverage between 33.3x and 35.0x.

TABLE 3. Cohort characteristics and polygenic risk score for European ancestry individuals

	Nongenetically enriched cohorts				Genetically enriched cohorts	
	HC	PD	Prodromal	SWEDD	Unaffected	Affected
n	905	1905	58	57	369	295
Female, n (%)	471 (52.0)	675 (35.4)	13 (22.4)	21 (36.8)	221 (59.9)	153 (51.9)
Inferred AJ, n (%)	73 (8.1)	144 (7.6)	2 (3.4)	4 (7.0)	263 (71.3)	203 (68.8)
Age at baseline, years	63.6 (10.8)	64.7 (9.5)	69.3 (5.9)	60.9 (10.4)	56.1 (12.7)	65.5 (10.9)
Education level						
Less than 12 years, n (%)	13 (1.4)	58 (3.0)	13 (22.4)	10 (17.5)	17 (4.6)	35 (11.9)
12–16 years, n (%)	659 (72.8)	1405 (73.8)	19 (32.8)	32 (56.1)	131 (35.5)	136 (46.1)
Greater than 16 years, n (%)	233 (25.7)	440 (23.1)	25 (43.1)	15 (26.3)	219 (59.3)	123 (41.7)
Latest case/control status						
Case, n (%)	3 (0.3)	1887 (99.1)	10 (17.2)	50 (87.7)	3 (0.8)	292 (99.0)
Control, n (%)	896 (99.0)		2 (3.4)	1 (1.8)	352 (95.4)	1 (0.3)
Other (including prodromal state), n (%)	6 (0.7)	18 (0.9)	46 (79.3)	6 (10.5)	14 (3.8)	2 (0.7)
Polygenic risk score						
90 Common risk SNPs from Nalls et al (2019)	0.0 (1.0)	0.6 (1.1)[c]	0.1 (0.9)	0.3 (0.9)[a]	2.7 (1.9)	3.2 (2.0)[b]
83 SNPs from Nalls et al (2019) - excluding 7 Variants in *GBA*, *LRRK2*, and *SNCA* regions	0.0 (1.0)	0.6 (1.0)[c]	0.1 (0.9)	0.3 (0.9)[a]	0.2 (1.0)	0.5 (1.0)[b]

Mean (SD) if not specified. Polygenic risk scores were standardized by the mean and the standard deviation of the scores in healthy controls in general cohorts.
P values for the score differences from the healthy controls (nongenetically enriched cohorts) or the unaffected (genetically enriched cohorts) are shown.
AJ, Ashkenazi Jewish; HC, Healthy controls or unaffected participants in genetic cohort/registry; PD, Parkinson's disease; SWEDD, scan without evidence of dopaminergic deficit.
[a] $P < 0.05$. (*t* test).
[b] $P < 0.01$.
[c] $P < 0.001$.

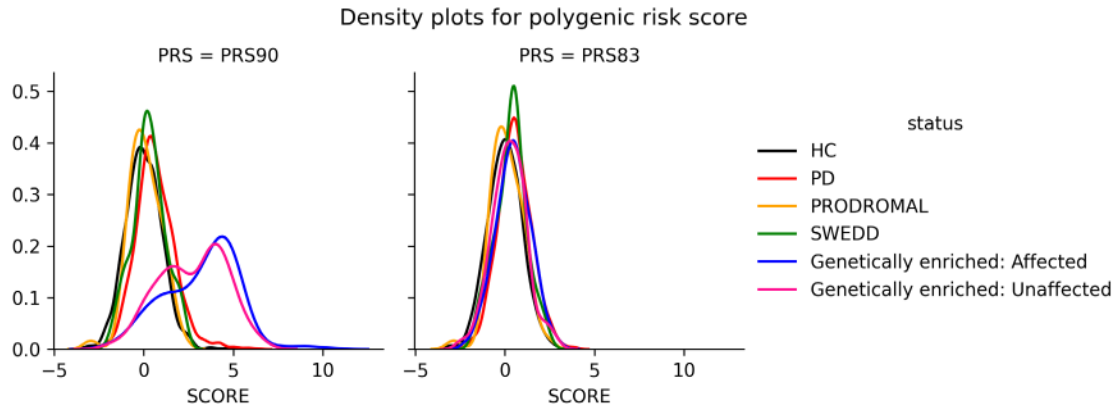

FIG 2. Density plots for polygenic risk score. HC, healthy volunteers; PD, participants with Parkinson's disease; SWEDD, scan without evidence of dopamine deficit. PRS90 is a weighted sum of the independent risk loci reported in Nalls et al (2019). PRS83 is the same but removing the 7 variants in *GBA*, *LRRK2*, and *SNCA* regions. [Color figure can be viewed at wileyonlinelibrary.com]

Carriers of pathogenic/LoF familial PD genes are summarized in Table 2. BigQuery enabled us to derive these variants of interest immediately. The carriers of these variants for *GBA* and *LRRK2* were relatively frequent because of the recruiting strategy for the targeted genetic recruitment. In the nongenetically enriched cohorts, the carrier frequencies between cases and controls were not statistically significant except for that of pathogenic+ variants of *GBA* (carriers/all were 30/1365 in cases and 5/736 in controls, $P = 0.0069$ with Fisher's exact test). We observed a relatively high number of *PRKN* pathogenic/LoF carriers compared with those of the other genes of interest. The majority of them (n = 125) were the carriers of a *PRKN* intron variant rs9364644 (G > A,C). Although the VEP annotated it as a high-impact variant (splice donor variant), the variant was not significantly enriched among cases in nongenetically enriched cohorts, and clinical significance was unclear. We also identified 2 pathogenic variants in the *VPS35* gene. In further evaluation, they seem to be pathogenic for another overlapping gene, *ORC6*, as being initiator codon variants and reported to cause Meier-Gorlin syndrome 3.[14] Twenty-seven *SNCA* pathogenic variant (p.A53T) carriers were all from the genetically enriched cohorts.

The population analysis identified 95.3% of the study participants (3755 of 3941) as of European descent (population plots in Supplemental Materials). Their PRS score distributions and other basic characteristics are summarized in Table 3. The mean PRS was significantly higher in PD cases ($P = 3.5 \times 10^{-47}$, *t* test) as well as SWEDDs ($P = 0.033$, *t* test) than controls in the nongenetically enriched cohorts. The mean PRS of the affected was also significantly higher than the unaffected in the genetically enriched cohorts ($P = 0.002$, *t* test). Participants in the genetically enriched cohorts had a higher PRS score than those in the nongenetically enriched cohorts ($P < 1.0 \times 10^{-300}$, *t* test). Indeed, the PRS scores showed distinguished distributions between the participants in the nongenetically enriched cohorts and the genetically enriched cohorts (Fig. 2). This is because of the results of the recruiting strategy of these cohorts. Most of them carried the high-risk variant of *GBA*, *LRRK2*, or *SNCA*, and when we recalculated the PRS excluding 7 risk variants in these gene regions (rs114138760, rs35749011, rs76763715,rs34637584, rs76904798, rs5019538, and rs13117519), the polygenic risk score (PRS83) distributions became similar (Fig. 2). However, the mean PRS83 was still significantly different between the unaffected in the enriched cohorts and the healthy volunteers in the nonenriched cohorts. ($P = 5.2 \times 10^{-5}$). When we calculated the effects of the risk variants on the PRS difference between the 2 arms, the variants with the largest 3 effect sizes were rs34637584 (*LRRK2* p.G2019S), rs76763715 (*GBA* p.N370S), and rs34311866 (*TMEM175* p.M393T). After adjusting for the 3 variants, PRS differences between the 2 arms were not significant anymore ($P = 0.40$, *t* test). These variants are known to be enriched in the Ashkenazi Jewish population (AJ).[15] We plotted the AJ reference with the study data sets, and it was indeed overlapped on a cluster of participants, especially those of genetically enriched cohorts (Supplemental Materials). When we applied the cutoff of minimum PC3 among the AJ reference population to infer the AJ ancestry (PC3 = 0.156), the majority of the participants in the genetically enriched cohorts were inferred as AJ (Supplemental Materials).

Discussion

Here we provide an overview of the DNA sequencing data that forms part of the first data release of the public-private partnership project AMP PD. Release

version 1 contains WGS data from 3941 participants. These data have undergone extensive quality control and standardized alignment and variant calling, including a single joint calling step. The data quality is high, enabling robust variant detection and calling across the full spectrum of variant frequencies.

We provide various formats of data: CRAMs, BCFs, VCFs, plink binary files, and BigQuery tables. As we demonstrated in the creation of Table 2, BigQuery allows rapid interrogation of the underlying data and retrieval of variants of interest. As we demonstrated in the creation of Table 2, BigQuery allows rapid interrogation of the underlying data and retrieval of variants of interest, and although automated annotation is generally correct, the status of individual pathogenic/nonpathogenic variants may not always be reliable and may require secondary evaluation. Tutorials for researchers not familiar with BigQuery are available on the AMP PD platform (https://amp-pd.org/amp-pd-webinar-videos).

Our characterization of the WGS data available on the AMP PD platform as part of release 1 centered on topics of interest to the users of these genetic data or on issues that genetics could inform. The resource predominantly contains subjects of European ancestry, and we believe that the genetically derived ancestry should be taken into account in many of the research questions that will be addressed with the AMP PD data set, even those outside genetics. A large number of subjects show Ashkenazi Jewish ancestry, driven by the preselection of genetic cases and the high number of *LRRK2* p.G2019S and *GBA* p.N370S carriers. The genetic characterization extends beyond the classification of these mutations to include a range of disease-linked mutations present in both cases and as yet asymptomatic individuals.

We also assessed the common genetic risk burden in these subjects. This calculation was based on the latest work identifying genetic risk loci in PD.[1] The cases, as well as SWEDDs, carry higher cumulative genetic burden of common PD risk variants compared with controls in the nongenetically enriched cohorts, as expected.[16] Affected individuals also carry higher burden of cumulative risk than unaffected individuals in the genetically enriched cohorts, concordant with previous work.[17,18] Importantly, the score distributions were substantially different across study arms, reflecting the different recruitment strategies of the study arms.

This project has a unique architecture. The AMP PD project provides an integrated analytical platform, and much of the typical quality control and data processing that would be performed in WGS data has already been done to industry standards. Thus, although the underlying data are large, the most often used results and data forms have already been derived and can be readily accessed. Researchers can concentrate on their own analyses without time-consuming logistics such as setting up and maintaining computational infrastructure. Transparency and extensibility are additional significant advantages of this project. Analyses using AMP PD data and the AMP PD platform are easily shared or copied and are inherently reproducible. The data processing and analysis scripts discussed in this article are shared in the AMP PD project, and a cornerstone of the AMP PD philosophy is that other researchers are encouraged to share their processes, code, and results on the AMP PD analytical platform. We believe open science is the driving force of new discovery, and the architecture of the project supports this approach.

A key aspect of the current AMP PD data is the harmonization of both a broad and deep range of data. A particular strength therefore of AMP PD will be the integrated analysis of these multimodal data, and most such analyses will include genetics. The available data types include transcriptomic data, biologic data from blood and cerebrospinal fluid, imaging summaries and detailed clinical phenotypes, and test results. In addition, many of the data are available longitudinally. Immediate opportunities arise in the analysis of these data alone and integrated together. In the context of genetics one can imagine myriad uses, from adjustment for population structure to grouped analyses of clinical and biologic measures across suitably powered mutation types (both in cases and in asymptomatic carriers) and, importantly, analysis based on the varied burden of PD genetic risk score. These data also provide opportunities for more detailed genetic analyses such as those examining structural/repeat variants. As part of future releases AMP PD aims to create reference calls for these variants; this work is ongoing.

There are successful open science examples in other fields. The Accelerating Medicines Partnership program had already been running in 3 disease areas when AMP PD started —Alzheimer's disease, type 2 diabetes, rheumatoid arthritis /lupus. These programs developed open-access resources and have made a substantial contribution to the research community.[19] (AD knowledge portal, https://adknowledgeportal.synapse.org/; T2D Knowledge Portal, https://t2d.hugeamp.org/; ImmPort, https://www.immport.org/). NHGRI's Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL, https://anvilproject.org/) is another example providing a cloud-based platform for genomics data and an analytical environment.

The primary limitation of the project from a purely genetic perspective is its size. Analyses on rare variants generally require a much larger sample size than that of common variants. A simulation reported that 5000 cases and 5000 controls are required to achieve power of 0.8 for a burden test under the prior condition of the

risk:nonrisk variants ratio of 1:20 with a somewhat large relative risk of 5.[20] Notably, current plans aim to substantially extend the number of genetically characterized subjects within AMP PD; thus, the potential of this platform to support pure genetic discovery will improve with time.

Another limitation relates to the use of short-read sequencing technology. This method is less accurate and less powerful in detecting structural variants and tandem repeat variations.[21] There are multiple tools proposed for calling structural variants, and multialgorithm consensus pipelines are proposed.[22] However, it is difficult to capture break points of structural variants containing repeats or embedded within repeats by aligning short-reads to a reference. Long-read sequencing technologies are expected to resolve these difficulties. Although it is still expensive and the error rate is high, it has been improving, and it may be a promising future direction.

We acknowledge that the current study population is biased to those with European ancestry. This is a community-wide issue, but there is an ongoing effort to overcome the racial disparity in genetic research.[23] We are optimistic that we will be able to share data from ancestrally diverse individuals in the near future.

Notably, the accessible nature of AMP PD and its suitability for iterative and crowd-sourced analytical approaches means that as additional samples are added and as novel analytical/processing strategies become available (for example, calling structural variation), these can be rapidly deployed in AMP PD, and this only needs to be done once to provide a standardized community resource.

In conclusion, we describe here the genetic arm of AMP PD, which includes a significant amount of raw and processed genetic data relevant to PD research and more broadly to neurodegenerative disease research. We believe this will be the foundation of a growing fund of genetic knowledge that will serve the PD research community. ●

Acknowledgments: AMP PD, a public-private partnership, is managed by the FNIH and funded by Celgene, GSK, the Michael J. Fox Foundation for Parkinson's Research, the National Institute of Neurological Disorders and Stroke, Pfizer, Sanofi, and Verily. Data used in the preparation of this article were obtained from the AMP PD Knowledge Platform. For up-to-date information on the study, go to https://www.amp-pd.org. Clinical data and biosamples used in preparation of this article were obtained from the Fox Investigation for New Discovery of Biomarkers (BioFIND), the Harvard Biomarker Study (HBS), the Parkinson's Progression Markers Initiative (PPMI), and the Parkinson's Disease Biomarkers Program (PDBP). BioFIND is sponsored by the Michael J. Fox Foundation for Parkinson's Research (MJFF) with support from the National Institute for Neurological Disorders and Stroke (NINDS). The BioFIND Investigators have not participated in reviewing the data analysis or content of the article. For up-to-date information on the study, visit michaeljfox.org/biofind. The Harvard Biomarkers Study (HBS) is a collaboration of HBS investigators (full list of HBS investigator found at https://www.bwhparkinsoncenter.org/biobank/) and funded through philanthropy and NIH and Non-NIH funding sources. The HBS Investigators have not participated in reviewing the data analysis or content of the article. PPMI, a public-private partnership, is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners (the full names of all the PPMI funding partners can be found at www.ppmi-info.org/fundingpartners. The PPMI Investigators have not participated in reviewing the data analysis or content of the article. For up-to-date information on the study, visit www.ppmi-info.org. The Parkinson's Disease Biomarker Program (PDBP) consortium is supported by the National Institute of Neurological Disorders and Stroke (NINDS) at the National Institutes of Health. A full list of PDBP investigators can be found at https://pdbp.ninds.nih.gov/policy. The PDBP Investigators have not participated in reviewing the data analysis or content of the article. Participation by individuals employed by Data Tecnica International LLC was supported in part by a consulting contract between the National Institutes of Health (NIA/NINDS) and the company. Individuals employed by Data Tecnica International LLC report no conflict of interest relating to the work carried out in this report.

References

1. 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;18:1091–1102.
2. Blauwendraat C, Heilbron K, Vallerga CL, et al. Parkinson's disease age at onset genome-wide association study: defining heritability, genetic loci, and α-synuclein mechanisms. Mov Disord 2019;34(6):866–875.
3. 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;34:1839–1850.
4. King EA, Davis JW, Degner JF. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. PLoS Genet 2019;15:e1008489.
5. Leonard H, Blauwendraat C, Krohn L, et al. Genetic variability and potential effects on clinical trial outcomes: perspectives in Parkinson's disease. J Med Genet 2020;57(5):331–338.
6. McLaren W, Gil L, Hunt SE, et al. The Ensembl variant effect predictor. Genome Biol 2016;17:122.
7. Regier AA, Farjoun Y, Larson DE, et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat Commun 2018;9:4038.
8. Nalls MA, Bras J, Hernandez DG, et al. NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases. Neurobiol Aging 2015;36:1605.e7–1605.e12.
9. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly 2012;6:80–92.
10. Loh P-R, Danecek P, Palamara PF, et al. Reference-based phasing using the haplotype reference consortium panel. Nat Genet 2016;48:1443–1448.
11. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 2018;46:D1062–D1067.
12. International HapMap 3 Consortium, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. Nature 2010;467:52–58.
13. Bray SM, Mulle JG, Dodd AF, Pulver AE, Wooding S, Warren ST. Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population. Proc Natl Acad Sci U S A 2010;107:16222–16227.
14. de Munnik SA, Bicknell LS, Aftimos S, et al. Meier–Gorlin syndrome genotype–phenotype studies: 35 individuals with pre-replication complex gene mutations and 10 without molecular diagnosis. Eur J Hum Genet 2012;20:598–606.
15. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020;581:434–443.
16. Nalls MA, Keller MF, Hernandez DG, Chen L, Stone DJ, Singleton AB. Baseline genetic associations in the Parkinson's progression markers initiative (PPMI). Mov Disord 2016;31:79–85.

17. Blauwendraat C, Reed X, Krohn L, et al. Genetic modifiers of risk and age at onset in GBA associated Parkinson's disease and Lewy body dementia. Brain 2020;143:234–248.

18. Iwaki H, Blauwendraat C, Makarious MB, et al. Penetrance of Parkinson's disease in LRRK2 p.G2019S carriers is modified by a polygenic risk score. Mov Disord 2020;35(5):774–780.

19. Dolgin E. Massive NIH–industry project opens portals to target validation. Nat Rev Drug Discov 2019;18:240–242. https://doi.org/10.1038/d41573-019-00033-8

20. Sanders SJ, Neale BM, Huang H, et al. Whole genome sequencing in psychiatric disorders: the WGSPD consortium. Nat Neurosci 2017;20:1661–1668.

21. Lappalainen T, Scott AJ, Brandt M, Hall IM. Genomic analysis in the age of human genome sequencing. Cell 2019;177:70–84.

22. Werling DM, Brand H, An J-Y, et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. Nat Genet 2018;50:727–736.

23. Devaney S. All of us. Nature 2019;576:S14–S17.

Supporting Data

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