Large-scale rare variant burden testing in Parkinson's disease

Mary B. Makarious,[1,2,3] Julie Lake,[1] Vanessa Pitz,[4] Allen Ye Fu,[1,5] Joseph L. Guidubaldi,[4,6] Caroline Warly Solsberg,[7,8] Sara Bandres-Ciga,[6] Hampton L. Leonard,[1,6,9] Jonggeol Jeffrey Kim,[4,10] Kimberley J. Billingsley,[1,6] Francis P. Grenn,[1] Pilar Alvarez Jerez,[1,6] Chelsea Alvarado,[6,9] Hirotaka Iwaki,[1,6,9] Michael Ta,[6,9] Dan Vitale,[6,9] Dena Hernandez,[1] Ali Torkamani,[11] Mina Ryten,[12,13] John Hardy,[14,15] UK Brain Expression Consortium (UKBEC), Sonja W. Scholz,[16,17] Bryan J. Traynor,[1,17] Clifton L. Dalgard,[18] Debra J. Ehrlich,[19] Toshiko Tanaka,[20] Luigi Ferrucci,[20] Thomas G. Beach,[21] Geidy E. Serrano,[21] Raquel Real,[2,3] Huw R. Morris,[2,3] Jinhui Ding,[1] J. Raphael Gibbs,[1] Andrew B. Singleton,[1,6] Mike A. Nalls,[1,6,9] Tushar Bhangale[22,†] and Cornelis Blauwendraat[1,4,6,†]

[†]**These authors contributed equally to this work.**

Abstract

Parkinson’s disease (PD) has a large heritable component and genome-wide association studies to date have identified over 90 loci with disease-associated common variants, providing deeper insights into the disease biology. However, there have not been large-scale rare variant analyses for PD. To address this gap, we investigated the rare genetic component of PD at minor allele frequencies <1%, using whole genome and whole exome sequencing data from 7,184 PD cases, 6,701 proxy-cases, and 51,650 healthy controls from the Accelerating Medicines Partnership Parkinson's disease (AMP-PD) initiative, the National Institutes of Health, the UK Biobank, and Genentech. We performed burden tests meta-analyses on small indels and single nucleotide protein-altering variants, prioritized based on their predicted functional impact. Our work identified several genes reaching exome-wide significance. Two of these genes, *GBA1* and *LRRK2*, have variants that have been previously implicated as risk factors for PD, with some variants in *LRRK2* resulting in monogenic forms of the disease. We identify potential novel risk associations for variants in *B3GNT3, AUNIP, ADH5, TUBA1B, OR1G1, CAPN10,* and *TREML1*, but were unable to replicate the observed associations in independent datasets. Of these, *B3GNT3* and *TREML1* could provide new evidence for the role of neuroinflammation in PD. To date, this is the largest analysis of rare genetic variants in PD.

© Published by Oxford University Press on behalf of the Guarantors of Brain 2023. This work is written by (a) US Government employee(s) and is in the public domain in the US.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

Author affiliations

1 Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD 20814, USA

2 Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, London, WC1N 3BG, UK

3 UCL Movement Disorders Centre, University College London, London, WC1N 3BG, UK

4 Integrative Neurogenomics Unit, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD 20814, USA

5 Department of Cell Biology and Neuroscience, Rutgers University, Piscataway, NJ 08854, USA

6 Center for Alzheimer's and Related Dementias (CARD), National Institute on Aging and National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20814, USA

7 Memory and Aging Center, Department of Neurology, University of California San Francisco, San Francisco, CA 94158, USA

8 Pharmaceutical Sciences and Pharmacogenomics, University of California San Francisco, San Francisco, CA 94143, USA

9 Data Tecnica International, Washington, DC 20812, USA

10 Preventive Neurology Unit, Wolfson Institute of Preventive Medicine, Queen Mary University of London, London, EC1M 6BQ, UK

11 The Scripps Research Institute, La Jolla, CA 92037, USA

12 NIHR Great Ormond Street Hospital Biomedical Research Centre, University College London, London, WC1N 1EH, UK

13 Department of Genetics and Genomic Medicine, Great Ormond Street Institute of Child Health, University College London, London, WC1N 1EH, UK

14 UK Dementia Research Institute and Department of Neurodegenerative Disease and Reta Lila Weston Institute, UCL Queen Square Institute of Neurology and UCL Movement Disorders

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

1 Centre, University College London, London, WC1N 3BG, UK
2 15 Institute for Advanced Study, The Hong Kong University of Science and Technology, Hong
3 Kong SAR, China
4 16 Neurodegenerative Diseases Research Unit, National Institute of Neurological Disorders and
5 Stroke, Bethesda, MD 20814, USA
6 17 Department of Neurology, Johns Hopkins University Medical Center, Baltimore, MD 21287,
7 USA
8 18 The American Genome Center, Uniformed Services University of the Health Sciences,
9 Bethesda, MD 20814, USA
10 19 Parkinson's Disease Clinic, Office of the Clinical Director, National Institute of Neurological
11 Disorders and Stroke, Bethesda, MD 20814, USA
12 20 Translational Gerontology Branch, National Institute on Aging, NIH, Baltimore, MD 21224,
13 USA
14 21 Civin Laboratory for Neuropathology, Banner Sun Health Research Institute, Sun City, AZ
15 85351, USA
16 22 Department of Human Genetics, Genentech, Inc., South San Francisco, CA 94080, USA
17
18 Correspondence to: Cornelis Blauwendraat
19 Laboratory of Neurogenetics
20 NIA, NIH, Building 35, 35 Convent Drive
21 Bethesda, MD 20892, USA
22 E-mail: cornelis.blauwendraat@nih.gov
23
24 **Running Title:** Large-scale rare variant burden testing in PD
25 **Keywords:** Parkinson's disease; burden; *GBA1*; *LRRK2*; genetics; rare variant
26 **Abbreviations:** PD = Parkinson's disease; PPMI = Parkinson's Progression Markers Initiative;

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

BioFIND = Fox Investigation for New Discovery of Biomarkers; PDBP = Parkinson's Disease Biomarker Program; NIH = National Institutes of Health; NINDS = National Institute of Neurological Disorders and Stroke; NIA = National Institute on Aging; AMP-PD = Accelerating Medicines Partnership in Parkinson's disease; LOF = Loss-of-Function; MAF = minor allele frequency; MAC = minor allele count; SKAT-O = Sequence Kernel Association Test – Optimal; CMC Wald = Combined and Multivariate Collapsing (CMC) Wald; LOFTEE = Loss-of-Function (LoF) Transcript Effect Estimator; VEP = Variant Effect Predictor; SNP = Single nucleotide polymorphism; INDEL = insertion/deletion; BWA = Burrows-Wheeler Aligner; GATK = Genome-analysis toolkit; CADD = Combined Annotation Dependent Depletion; VQSR = Variant Quality Score Recalibration; WGS = whole genome sequencing; LNG = Laboratory of Neurogenetics; UKBEC = United Kingdom Brain Expression Consortium; NABEC = North American Brain Expression Consortium; QC = quality control; GWAS = genome-wide association studies

Introduction

Parkinson's disease (PD) is a complex neurological disease likely caused by an interplay between aging, environmental factors and genetics. While the role of common genetic variants in PD has been extensively studied using large genome-wide association studies (GWAS), rare variants can also contribute to familial and sporadic disease. To date, 92 independent risk signals have been associated with PD including common variants in close proximity to *SNCA*, *TMEM175* and *MAPT* [1,2]. Most of the risk alleles found by array-based GWAS have frequencies over 5% in the population of interest, often reside in non-coding regions of the genome, and typically have small effect sizes. In contrast, rare damaging and pathogenic variants implicated in PD, such as coding variants in *SNCA* [3] and *PRKN* [4], have traditionally been identified using family-based approaches. One aspect of major interest in disease genetics is the large number of pleomorphic genes, where multiple variants of varying allele frequency present with a wide range of effect sizes [5]. For example, in PD, GWAS identified common variants with moderate effects near *GBA1*, *GCH1*, *LRRK2*, *SNCA* and *VPS13C* [1], while familial studies identified rare variants in the same genes resulting in more damaging effects (e.g., *GBA1* p.N370S, *LRRK2* p.G2019S, and *SNCA* p.A53T).[6–9]

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

1 In contrast to common variants, there have been no large-scale efforts investigating the role of 2 rare variants in PD on a genome-wide scale. Although rare variant associations for several PD 3 genes (such as *ARSA* and *ATP10B*) have been reported in candidate gene studies [10,11], these genes 4 remain controversial due to lack of replication in independent PD datasets [12–15]. One of the main 5 challenges that comes with analyzing rare variants is that the quality and reliability of imputation 6 procedures decreases with allele frequency. Since genome-wide genotyping methods are 7 currently much cheaper than sequencing, most large datasets used for GWAS rely on imputed 8 genotype data. A strength of the present study is that we focus on using whole genome (WGS) 9 and whole exome sequencing (WES) to facilitate the analysis of rare variants. We perform the 10 largest genome-wide analysis of rare variants in PD to date, investigating 7,184 PD cases, 6,701 11 proxy-cases (defined as having a parent or sibling with PD), and 51,650 neurologically healthy 12 controls of European ancestry from several large sequencing efforts. Using this data, we execute 13 gene-level burden testing in order to understand how moderate- to large-effect rare variants 14 contribute to the genetic etiology of PD.

15

16 Materials and methods

17 *AMP-PD and NIH Genome Sequencing Data*

18 Whole genome sequencing data was obtained from multiple datasets including the Parkinson's 19 Progression Markers Initiative (PPMI), the Parkinson's Disease Biomarkers Program (PDBP), 20 and the Harvard Biomarker Study (HBS), BioFIND, SURE-PD3, and STEADY-PD3 as part of 21 the Accelerating Medicines Partnership in Parkinson's Disease (AMP-PD) initiative. Several 22 other datasets were sequenced in parallel at the Laboratory of Neurogenetics (LNG) and the U.S. 23 Uniformed Services University (USHUS), including samples from the National Institutes of 24 Health (NIH) PD clinic, the United Kingdom Brain Expression Consortium (UKBEC) [16], the 25 North American Brain Expression Consortium (NABEC) [17], and Wellderly [18]. All cohorts from 26 AMP-PD (PPMI, PDBP, HBS, BioFIND, SURE-PD3, and STEADY-PD3) were processed using 27 the GATK Best Practices guidelines set by the Broad Institute's joint discovery pipeline and 28 elaborated on elsewhere [19]. All other cohorts were joint called separate from AMP-PD but in a 29 similar manner, also from the processed WGS data following the GATK Best Practices using the 30 Broad Institute's workflow for joint discovery and Variant Quality Score Recalibration

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

1 (VQSR)[20]. Data processing and quality control (QC) procedures have been described previously
2 [19,21]. Reported elsewhere, these sequencing metrics had a median/mean coverage between 33.3x
3 and 35.0x [19]. Additional quality control was performed to exclude closely related individuals
4 (PI_HAT >0.125) by selecting one sample at random using PLINK (v1.9; [22]). All individuals
5 were of European ancestry as confirmed by principal component analysis using HapMap3
6 European ancestry populations. Individuals recruited as part of a biased and/or genetic dataset,
7 such as *LRRK2* and *GBA1* rare variant carriers within a specific effort of PPMI, were excluded
8 from this analysis. Including all variants within the gene boundaries, a minimum allele count
9 (MAC) threshold of 1 was applied. Exonic regions were subset from the whole genome
10 sequencing data using the exome calling regions from gnomAD lifted over to hg38 [23].

11

12 *UK Biobank*

13 Exome sequencing data from a total of 200,643 individuals (OQFE dataset, field codes: 23151
14 and 23155) were downloaded from the UK Biobank in December of 2020 [24]. As previously
15 described elsewhere, the UK Biobank Exome Sequencing Consortium sequenced these exomes
16 with 95.8% of targeted bases covered at a depth of 20x or higher [25]. Standard quality control was
17 performed to exclude non-European outliers. Closely related individuals (PI_HAT >0.125) were
18 excluded by selecting one sample at random using PLINK (v1.9; [22]). Standard exome sequencing
19 data filtering was applied using suggested parameters as described in previous UK biobank
20 exome sequencing studies [25].

21

22 UK Biobank phenotype data were obtained from ICD10 codes (field code: 41270), PD (field
23 code: 131023), illnesses of father and mother (field codes: 20107 and 20110), parkinsonism
24 (field code: 42031) or dementia (field code: 42018), genetic ethnic grouping (field code: 22006),
25 year of birth (field code: 34) and age of recruitment (field code: 21022). Cases were defined as
26 any individual identified as having PD using the above field code. Proxy-cases were defined as
27 having a parent or sibling with PD as previously reported [1]. Controls were filtered to exclude any
28 individuals with an age of recruitment < 59 years, any reported nervous system disorders
29 (Category 2406), a parent with PD or dementia (field codes: 20107 and 20110) and any reported
30 neurological disorder (field codes: Dementia/42018, Vascular dementia/42022, FTD/42024,

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

1 ALS/42028, Parkinsonism/42030, PD/42032, PSP/42034, MSA/42036).

2

3 *Genentech*

4 Whole genome sequencing data from Genentech included a total of 2,710 PD cases and 8,994
5 individuals used as controls. PD cases included 2,318 individuals from 23andMe, a subset of
6 those included in the analysis by Chang and colleagues [26] who were contacted and provided
7 consent for this analysis. An additional 392 PD cases were obtained from the Roche clinical trial
8 TASMAR. Individuals included as controls were obtained from various Genentech clinical
9 trials/studies and included cases for four diseases that do not share notable heritability with PD:
10 age-related macular degeneration (AMD, n=1,735), asthma (n=3,398), idiopathic pulmonary
11 fibrosis (IPF, n=1,532), and rheumatoid arthritis (RA, n=2,329). Illumina HiSeq based 30x
12 genome sequencing was performed on all samples using 150bp paired-end reads. Genotypes with
13 a genotype quality (GQ) < 20 were labeled as missing. The reads were then mapped to the
14 GRCh38 reference genome with Burrows-Wheeler Aligner (BWA)[27], followed by application of
15 GATK [27,28] for base quality score recalibration, indel realignment, and duplicate removal. This
16 was followed by SNP and INDEL discovery and genotyping across all samples simultaneously
17 using variant quality score recalibration according to GATK Best Practices recommendations [29–
18 31]. The 11,704 samples included in these analyses passed the following QC steps: genotype
19 missing rate < 0.1, no sample pair had kinship coefficient (k0 i.e. probability of zero alleles
20 shared identical-by-descent; or the value Z0 reported by PLINK's –genome module) < 0.4; and
21 no sample was an outlier in five iterations of outlier removal using PCA [32].

22

23 *Variant Annotation*

24 Variants were annotated using the SnpEff and SnpSift annotation softwares (v4.3t; [33]) as well as
25 the Ensembl Variant Effect Predictor (VEP; v104; [34]) package. Both the Combined Annotation
26 Dependent Depletion (CADD; v1.4; [35]) and the Loss-of-Function (LoF) Transcript Effect
27 Estimator (LOFTEE; v1.02; [23]) VEP plugins were used. SnpEff is a toolbox based on 38,000
28 genomes that is designed to annotate genetic variants and predict their downstream functional
29 consequences. SnpSift leverages multiple databases to filter SnpEff outputs and prioritize

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

variants, and can predict amino acid changes as having “moderate” or “high” impact. The CADD plugin for VEP is a tool used to score the deleteriousness of single nucleotide variants, insertions, and deletions. A CADD PHRED score is a scaled measure of deleteriousness, with a score of 20 indicating that the variant is among the top 1% of deleterious variants in the genome [35]. The LOFTEE plugin for VEP is uniquely designed to assess stop-gain, frameshift, and splice-site disrupting variants and classify these as LoF with either low or high confidence. The following variant classes were used for gene burden analyses: 1) missense variants as defined by SnpEff, 2) moderate or high impact variants as defined by SnpEff/SnpSift, 3) high confidence LoF variants as defined by LOFTEE, and 4) variants with either a CADD PHRED score > 20 or high confidence LoF variants as defined by LOFTEE.

Gene Burden Analysis and Meta-Analysis

The AMP-PD and NIH datasets were merged prior to gene burden analysis, with 3,848 duplicates removed prior to analysis. Rare variant testing for this merged dataset, the UK Biobank case-control dataset, and the UK Biobank proxy-control datasets were performed using the Sequence Kernel Association Test – Optimal (SKAT-O) and the Combined and Multivariate Collapsing (CMC) Wald algorithms [36,37]. These algorithms were run using the RVtests package (v2.1.0; [38]). The CMC Wald test collapses and combines all rare variants and then performs a Wald test, where only an alternative model is fit and the effect size is estimated [39]. SKAT-O is an optimized sequencing kernel association test designed to combat limitations introduced by the SKAT and burden tests. SKAT-O aggregates the associations between variants and the phenotype of interest while allowing for SNP-SNP interactions, and has been proven to detect genes more reliably than a burden or SKAT test separately by adaptively selecting the best linear combination of both SKAT and burden tests to maximize test power [40]. All analyses were stratified by the four variant classes described above and by maximum minor allele frequencies (MAF) levels of 1% and 0.1%. For Genentech data, SKAT-O and CMC-Wald tests were performed using the R package SKAT [41].

The combined AMP-PD and NIH dataset was adjusted for sex, age, and the first five principal components. The UK Biobank datasets were adjusted for sex, Townsend scores, and the first five

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023
ACCEPTED MANUSCRIPT

principal components. For the UK Biobank analyses, only neurologically healthy controls 60 years and older were included in analyses, and therefore age was not included as a covariate. Meta-analyses of the resulting summary statistics per gene were performed using custom Python (v3.7) scripts, which we have made available on our GitHub (https://github.com/neurogenetics/PD-BURDEN). In summary, the two meta-analysis approaches used in this study were 1) a combined p-value approach using Fisher's test, and 2) a weighted Z-score approach. In previous studies, Fisher's method was reported to detect > 75% of causal effects (either deleterious or protective) that are in the same direction [42]. Unless otherwise stated, all results reported in this manuscript correspond to the SKAT-O rare variant test, and all meta-analyses were performed using the combined p-values reported following Fisher's test.

Rare variant analyses were performed on each dataset separately and all data is using genome build hg38. Two joint meta-analyses were performed as follows: 1) a case-control meta-analysis between the combined AMP-PD and NIH dataset, the Genentech dataset, and the UK Biobank case-control dataset, and 2) a meta-analysis of the case-control and proxy-control results from the combined AMP-PD and NIH dataset, the Genentech dataset, the UK Biobank PD case-control dataset, the UK Biobank sibling proxy-cases dataset, and the UK Biobank parent proxy-cases dataset. A summary of the analysis workflow is outlined in **Figure 1**.

Power Calculations

100 gene simulations were run using the power calculation function with default European haplotypes made available in the SKAT R package (v2.0.1; [40]). The total sample size was estimated at 65,535, with 7,184 PD cases, 6,701 proxy-cases down-weighted to ¼ of a PD case (corresponding to 1,675 cases), and 51,650 controls resulting in a case proportion of 13.5%. We estimated the disease prevalence of PD at 1% as previously described [43] and used an exome-wide significance threshold calculated by assuming close to 20,000 protein-coding genes, resulting in a Bonferroni correction of 2.50E-6. Since we used two different algorithms for burden testing, we set the final threshold of significance to 1E-6. Power calculations based on varying percentages of causality (10%, 5%, 3%, 1%, and 0.5%) and causal MAF (0.05%, 0.1%, 0.5%, 1%, 3%, and 5%) are reported in **Supplementary Table 5.** Assuming at least 3% of the rare

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

ACCEPTED MANUSCRIPT

alleles tested are causal, this analysis has ≥ 80% power to detect associations at the tested MAF cutoffs **(Supplementary Table 5)**.

Data and Code Availability

Accelerating Medicines Partnership in Parkinson's Disease (AMP PD data) and quality control notebooks are access-controlled [https://amp-pd.org/], and require individual sign-up to access the data. United Kingdom Biobank (UKBiobank) data are access-controlled and require an application for access [https://www.ukbiobank.ac.uk/]. The remaining cohorts were obtained through collaborations with the National Institutes of Health (NIH) and Genentech. Each contributing study abided by the ethics guidelines set out by their institutional review boards, and all participants gave informed consent for inclusion in both their initial cohorts and subsequent studies. Each contributing study abided by the ethics guidelines set out by their institutional review boards, and all participants gave informed consent for inclusion in both their initial cohorts and subsequent studies. The research using data from the NIH Parkinson's Disease clinic cohort was approved by the NIH Intramural IRB under protocol number 01-N-0206. The research with the remaining cohorts was deemed "not human subjects research" by the NIH Office of IRB Operations and stated that no IRB approval is required. The NIH Intramural IRB has waived ethical approval for the overall study (IRB #001161). All data produced in the present work are contained in the manuscript. All authors and the public can access the statistical programming code used in this project for the analyses and results generation on GitHub at https://github.com/neurogenetics/PD-BURDEN, as well as supplementary tables and full results. MBM and CB take final responsibility for the decision to submit the paper for publication. NABEC is available from NCBI dbGaP, study accession phs001300.v2.p1.

Results

Study overview

A total of 7,184 PD cases, 6,701 sibling/parent proxy-cases, and 51,650 controls with whole genome (AMP-PD, NIH and Genentech) or exome (UK Biobank) sequencing were included in this analysis **(Table 1)**. Rare variant gene-level burden tests were performed across all genes for

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

ACCEPTED MANUSCRIPT

1 four variant classes and two causal MAF cutoffs (**Figure 1**). As expected, we observed that more
2 deleterious variant classes resulted in fewer variants tested per gene. For a full overview of the
3 frequency and number of variants within each gene in cases and controls, stratified by variant
4 class and cohort (excluding Genentech), please see **Supplementary Tables 11, 12, 13, and 14.**
5

6 *Genetic burden testing in large PD case-control datasets*

7 Initial gene burden analyses per dataset (AMP-PD and NIH Genomes, Genentech, UK Biobank
8 cases, UK biobank sibling proxies, and UK Biobank parent proxies) resulted in several known
9 PD genes (e.g. *GBA1* and *LRRK2*) reaching significance exome-wide ($P < 1E-6$; **Tables 2 and 3;**
10 **Supplementary Tables 15 and 16**), confirming the validity of our approach. Lambda values per
11 dataset showed minimal genomic inflation when adjusted for the number of cases, proxy-cases,
12 and controls (λ_{1000}; **Supplementary Table 3**). As expected, datasets with smaller sample sizes,
13 such as the UK Biobank sibling proxy-control dataset, resulted in increased genomic deflation
14 when analyzed separately ($\lambda_{1000} < 0.9$).
15

16 Rare variant burden analysis of both *GBA1* and *LRRK2* reached significance exome-wide in the
17 initial analysis of missense, moderate/high impact, and LoF or highly deleterious (CADD
18 PHRED > 20) variants. In our analyses, we focused on LoF variants to limit the scope of burden
19 testing to rare variants that are the most likely to be highly deleterious. *GBA1* was significant for
20 these variant categories in both the Genentech (P=1.32E-08; P=5.70E-08; P=6.99E-08,
21 respectively) and UK Biobank parent proxies (P=2.15E-10; P=2.15E-10; 2.15E-10, respectively)
22 datasets. *LRRK2* was significant for these categories in the combined AMP-PD and NIH dataset
23 (P=1.96E-07; P=2.09E-07; P=2.23E-07, respectively). LoF variants in *B3GNT3* were significant
24 exome-wide in the Genentech dataset (P=4.40E-09) and replicated at nominal significance in the
25 UK Biobank parent proxies dataset (P=0.032; **Supplementary Figures 3 through 9** for
26 Genentech [hg38: chr19:17807816:T:G, chr19:17807816:T:G, chr19:17807816:T:G] and UK
27 Biobank [hg38: chr19:17807982:GC:G; chr19:17808033:C:T; chr19:17812105:C:CA]).
28 Moderate and high impact variants in *TUBA1B* were significant in the UK Biobank parent
29 proxies dataset (P=9.48E-07). LoF or highly deleterious variants in *ADH5* were significant in the

ACCEPTED MANUSCRIPT
Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

UK Biobank cases-control dataset (P=3.13E-07), and LoF or highly deleterious variants in *OR1G1* were significant in the UK Biobank sibling proxies dataset (P=6.58E-07; **Table 2; Supplementary Table 16**).

Ultra-rare variant (MAF < 0.1%) burden analysis of missense, moderate/high impact, and LoF or highly deleterious variants in *GBA1* were significant exome-wide in the UK biobank parent proxies dataset (P=6.88E-08; P=5.13E-10; P=7.89E-08, respectively). LoF or highly deleterious variants in *GBA1* were also significant in the UK Biobank case-control dataset (P=4.56E-07). LoF or highly deleterious variants in *LRRK2* were significant in the Genentech dataset (P=6.15E-07). Moderate/high impact variants in *AUNIP* were significant in the UK Biobank case-control dataset (P=3.04E-08), and *TUBA1B* in the UK Biobank parent proxies dataset (P=9.48E-07). LoF variants in *B3GNT3* were significant in the Genentech dataset (P=4.40E-09), and *AUNIP* in the UK Biobank case-control dataset (P=3.13E-08). LoF or highly deleterious variants in *AUNIP* were significant in the UK Biobank case-control dataset (P=3.15E-08), and LoF or highly deleterious variants in *OR1G1* were significant in the UK biobank sibling proxies dataset (P=6.58E-07). Ultra-rare variant burden analysis identified no significant genes exome-wide in any of the four variant classes within the AMP-PD and NIH genomes (P < 1E-6; **Table 3; Supplementary Table 15**).

Meta-analyses of large PD datasets

The first meta-analysis (herein called the case-control meta-analysis) excluded any UK Biobank proxy-cases. The second meta-analysis (herein called the case-control-proxies meta-analysis) included UK Biobank proxy-cases in addition to cases and controls. No significant divergence from expected lambda values (range: 0.97-1.00) were detected in any of the meta-analyses performed (**Supplementary Table 4**). Rare variant burden analysis of missense, moderate/high impact, and LoF or highly deleterious variants in *GBA1* were significant exome-wide across both meta-analyses ([case-control P=3.27E-14; P=9.10E-15; P=3.722E-14, respectively] and [case-control-proxies P=1.46E-21; P=1.32E-22; P=9.12E-22, respectively]). High confidence LoF variants in *CAPN10* (case-control P=3.60E-07; case-control-proxies P=7.84E-07) and *B3GNT3* (case-control P=4.40E-09; case-control-proxies P=3.36E-09) were also significant exome-wide

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

1 (**Table 2**).

2

3 Ultra-rare variant burden analysis of moderate/high impact variants and high confidence LoF
4 variants in *AUNIP* were significant exome-wide across both meta-analyses ([case-control
5 P=1.54E-08; P=1.64E-08, respectively] and [case-control-proxies P=2.70E-07; P=2.04E-07,
6 respectively]). Moderate/high impact variants in *TREML1* were significant with the inclusion of
7 proxy-cases. As in the rare variant burden analysis, ultra-rare LoF variants in *CAPN10* (case-
8 control P=3.60E-07; case-control-proxies P=7.84E-07) and *B3GNT3* (case-control P=4.40E-09;
9 case-control-proxies P=3.36E-09) were also significant. Notably, both rare (MAF < 1%) and
10 ultra-rare (MAF < 0.1%) *GBA1* variants showed significant associations with PD risk (**Tables 2**
11 **and 3**).

12

13 *B3GNT3* was identified in the high confidence LoF variant class group with p-values of 4.40E-
14 09 in the Genentech dataset and P=0.032 in the UK biobank parent proxies. However, no
15 variants meeting this criteria were present in the AMP-PD and NIH genomes, so the association
16 of rare LoF variants in *B3GNT3* could not be confirmed. The majority of novel candidate genes
17 identified in this study (*B3GNT3*, *AUNIP*, *ADH5*, *TUBA1B*, *OR1G1*, *CAPN10*, and *TREML1*)
18 only reached significance exome-wide using the SKAT-O test. (**Supplementary Table 7**). Full
19 results from the SKAT-O and CMC Wald burden tests performed for each variant class, MAF
20 cutoff, and meta-analysis group can be found on our GitHub repository
21 (https://github.com/neurogenetics/PD-BURDEN).

22

23 *Conditional LRRK2 analysis*

24 Since *LRRK2* p.G2019S is a relatively common risk factor for PD, we explored whether the rare
25 variant association at *LRRK2* is driven primarily by this variant. For these analyses, *LRRK2*
26 p.G2019S status per individual was coded as 0, 1, or 2 depending on the allelic dosage, allowing
27 us to condition on *LRRK2* p.G2019S status without removing carriers. Allelic status was then
28 included as a covariate for the burden analyses. The observed association at *LRRK2* was lost (P >
29 0.05) after conditioning on the allelic status of *LRRK2* p.G2019S for all of the tested variant

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

1 categories and MAF thresholds in the discovery datasets (excluding Genentech; **Supplementary**
2 **Table 8**). Besides *LRRK2* p.G2019S, no other substantial coding risk in *LRRK2* was detected.
3 However, it is important to note that other previously identified rare coding variants that have
4 been shown to increase risk to PD were not detected in this study, including *LRRK2* p.R1441H
5 [44].

6

7 *Assessment of previously reported PD causal or high risk genes and GWAS regions*

8 We next assessed a large number of genes that showed rare variant associations with PD in
9 previous studies (for a full list, please see **Supplementary Table 9;** for frequencies and number
10 of variants for each gene, variant class, and dataset; please see **Supplementary Tables 11, 12,**
11 **13, and 14**). Besides the previously discussed *GBA1* and *LRRK2*, none of these genes met
12 exome-wide significance (P > 1E-6) in our analysis. However, we did observe sub-significant
13 association signals for LoF or highly deleterious variants in *ARSA* (P=8.73E-05) and *DNAJC6*
14 (P=8.08E-04; **Supplementary Tables 9, 11, and 12**). Since we did not detect a P-value of
15 interest in *PRKN* (P=0.30), which has been robustly associated with predominantly early onset
16 PD in previous studies, we investigated the enrichment of possible homozygous and potentially
17 compound heterozygous *PRKN* mutations in PD. In the most stringent variant class (LoF or
18 highly deleterious variants), we found a frequency of 0.41% in cases and 0.07% in controls in the
19 combined AMP-PD and NIH dataset (**Supplementary Table 6**). We also did not detect P-values
20 of interest in *VPS35* (present only in Genentech dataset, $n_{variants}$=4; lowest meta $\text{P-value}_{\text{case-control meta-analysis}}$
21 = 0.235; MAF=0.001; high confidence LoF variants; **Supplementary Tables 11 and**
22 **13**), previously associated with causing PD in an autosomal dominant fashion, or *SNCA* (lowest
23 meta $\text{P-value}_{\text{case-proxy-control meta-analysis}}$ = 0.274; MAF=0.001; CADD>20 or LoF variants;
24 **Supplementary Tables 11 and 13**), with mutations in the gene previously associated with an
25 earlier onset (<50 years) and more severe form of PD.

26

27 We also attempted to determine whether known PD loci identified by GWAS present rare variant
28 associations, as has been shown previously near *SNCA, GBA1, GCH1, VPS13C,* and *LRRK2* [6–9].
29 We assessed a total of 82 PD GWAS regions, 78 of which were identified in the largest GWAS
30 of Europeans [1], two of which were identified in the largest PD GWAS of East Asians [2], and two

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

of which were identified in the largest PD GWAS investigating progression [45] (**Supplementary Table 10**). Looking broadly at each meta-analysis group, only two genes, *GBA1* and *LRRK2*, were significant after Bonferroni correction for 2,361 unique genes within 1 megabase of known PD loci, suggesting that coding variants do not play a large role in these GWAS regions, but rather that signals are driven by non-coding variants in these regions.

Discussion

We report the results of rare variant gene burden tests of PD using the largest sample size to date including 7,184 PD cases, 6,701 proxy-cases, and 51,650 healthy controls. A meta-analysis of gene burden results reaffirms that rare variants in *GBA1* and *LRRK2* are associated with PD risk in individuals with European ancestry. However, we also observed several novel PD-associated genes (*B3GNT3*, *AUNIP*, *ADH5*, *TUBA1B*, *OR1G1*, *CAPN10* and *TREML1*) that met exome-wide significance (P < 1E-6) in our analysis. Although these genes were not significant across all of the datasets tested (**Supplementary Table 7)** and we were unable to replicate the associations at exome-wide significance in independent datasets, this may be due to varied power in the different datasets due to sample size and/or geographical population differences between the datasets that influence the presence or absence of rare variants of interest. We observed the strongest evidence of a novel rare variant association at *B3GNT3,* where LoF variants showed a significant meta-analysis P-value (P=4.40E-09) primarily driven by the Genentech (P=4.40E-09) and UK Biobank (parent proxies P=0.032) datasets. Variants meeting this criteria were not present in the combined AMP-PD and NIH genomes, requiring additional data to confirm association with PD risk. Upon investigation, we found that three LoF variants in *B3GNT3* are associated with increased risk of PD. Out of the four individuals carrying these *B3GNT3* variants in the Genentech dataset, three were PD cases and one was a control. The three PD cases reported a family history of PD, which is not uncommon in this cohort and not necessarily indicative of familial PD (up to 30% self-report a family history of PD). While these three PD cases reported earlier age at onset than typical PD (manifestation in their 30s and below), no enrichment for tremors, gait disturbances, REM sleep disturbances, or anosmia were reported. Additionally, no evidence of excess identity-by-descent (IBD) between these three PD cases were found (average k_0=0.91). These variants in *B3GNT3* are rare, with three variants driving the association in both the Genentech and UK Biobank parent proxies datasets, and are therefore

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

1 likely to be absent in the remaining datasets analyzed.

2

3 Previously suggested PD GWAS loci also harbor rare variants of interest, such as *SYT11*,
4 *FGF20*, and *GCH1* [46]. We identified no significant p-values in these genes, consistent with a
5 similar, albeit smaller, analysis performed in the East Asian population [46]. Therefore, it is
6 tempting to speculate what the exact mechanism is that underlies these PD GWAS loci. While
7 likely that some risk variants will affect gene expression differences, it is, however, unclear if all
8 risk variants contribute to risk via this mechanism.

9 The vast majority of previously PD-associated genes were not nominated by our analysis,
10 including *PINK1* and *PRKN (PARK2)*, which are the most common genetic cause of early onset
11 PD [47]. This is somewhat expected since burden testing algorithms are most well-powered to
12 detect dominant and high-risk variants such as those in *GBA1* and *LRRK2*, and are less sensitive
13 to recessive and ultra rare mutations. It is also important to note that PD patients who carry
14 *PRKN*, *PINK1,* and *SNCA* mutations often have a slightly different PD phenotype (e.g. earlier
15 onset, varying progression rates, rapid dementia onset) compared to the general PD population [48].
16 Since most PD cases included in this analysis showed onset of symptoms in their sixties, it is less
17 likely that they will harbor pathogenic *PRKN* mutations than those with early onset PD (**Table**
18 **1**). Additionally, it is also worth noting that certain known disease causing variants are extremely
19 rare, for example *SNCA* pathogenic missense variants have so far been identified in ~25 reports
20 and therefore are likely too rare to be identified in the current dataset. It is therefore likely that
21 such mutation carriers are underrepresented in the datasets included in this study.

22

23 Immune involvement including adaptive T lymphocyte response in PD is well described and
24 reviewed elsewhere [49]. *B3GNT3* encodes an enzyme involved in the synthesis of L-selectin
25 required for lymphocyte homing, particularly for rolling of leukocytes on endothelial cells,
26 facilitating their migration into inflammatory sites. *TUBA1B* encodes the 1B chain of alpha-
27 tubulin, the main constituent of cytoskeleton. Growing evidence suggests the role of microtubule
28 defects in progressive neuronal loss in PD [50,51]. Alpha-tubulin has previously been shown to
29 aggregate as a result of mutations in genes encoding proteins well known to be implicated in PD,
30 including parkin [52] and alpha-synuclein [53]. *TREML1* is one of the TREM receptors that are

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

1 increasingly being implicated in neurodegenerative disorders like Alzheimer's disease, PD, and
2 multiple sclerosis [54–56]. *ADH5* encodes for one of the alcohol dehydrogenases, which have been
3 studied in the past for association with PD risk with conflicting results [57–59]. There is no clear,
4 discernible connection between known PD biology and the function of the remaining three
5 genes: *AUNIP*, *OR1G1*, and *CAPN10*. Further studies providing genetic support and functional
6 data for these and related genes will be necessary to uncover their potential role in PD.
7
8 There are several limitations of this study. First, our analysis was restricted to individuals of
9 European ancestry. It is important to expand rare variant analyses of PD to non-European
10 populations, as well as varying age-at-onset ranges, as more whole genome and whole exome
11 sequencing data becomes available. While our analysis was constrained to assessing four variant
12 classes, we acknowledge that by creating these variant classes we are, in turn, testing specific
13 types of mechanisms. For example, in the case of LoF variants, we are assessing mutations that
14 impair protein function and its impact on disease risk, which is a limitation if the disease
15 mechanism is gain-of-function. Although the sample size is large compared to previous rare
16 variant analyses of PD, we lack power to detect associations in genes where ≤ 3% of the
17 variants tested are putatively functional or causal, as some rare variant tests weigh rarer variants
18 with increased penetrance and effect size differently or not at all (**Supplementary Table 5**).
19 Since our literature search for previously reported rare variant associations was comprehensive
20 and not limited to late-onset PD, it is possible that failure to replicate these associations is due to
21 our analysis focusing on associations in late-onset PD compared to controls. Another limitation
22 is, since not all the datasets included in the meta-analysis were not jointly called from an
23 alignment of raw reads, it is possible that batch effects in sites, sequencing and data processing
24 may bias the results. The meta-analysis model of these analyses to leverage the power of the
25 datasets without combining them should however limit these biases. Further follow-up of
26 candidate genes via segregation in multiplex families or resequencing in large case-control
27 datasets, particularly those enriched for early onset and familial cases, is warranted. Additionally,
28 our analysis included parent and sibling proxy-cases from the UK Biobank to increase statistical
29 power. Although PD proxy-cases have shown to be valuable in large-scale studies investigating
30 common variants [1] and we have demonstrated their utility at detecting rare variant associations in

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

ACCEPTED MANUSCRIPT

known PD genes such as *GBA1* (**Supplementary Table 7**), we acknowledge that caution should be used when searching for recessive forms of disease. Finally, the vast majority of PD patients included in this study are from the "general" PD population, of which typically less than ~10% have a positive family history. Future rare variant studies will benefit from recruitment efforts that prioritize PD patients who are highly suspected to have a monogenic form of disease since these individuals are more likely to harbor highly pathogenic or causal mutations that have not previously been associated with PD. This strategy is being actively used for recruitment of PD patients by the Global Parkinson's Genetics Program [60].

Clinical heterogeneity within PD cases has been well documented, and further validation is needed to confirm the pathogenicity of rare or ultra-rare variants and their impact on disease [61–63]. Analysis of rare variants restricted to subtypes of PD may identify genes important in PD subtypes but not PD as a whole. Our analysis was also restricted to SNVs and small indels, as we did not look at copy number variants generated by short- or long-read sequencing since we did not have access to all raw data to perform such analyses. Future analyses will benefit from including copy number variants which have been shown to be important and causal for PD[4,64,65] and especially using of long-read sequencing, as long-read sequencing is able to identify more and more robustly copy number variants in comparison to short-read sequencing [66].

Overall, we performed the largest PD genetic burden test to date. We identified *GBA1* and *LRRK2* as two genes harboring rare variants associated with PD and nominated several other previously unidentified genes. While we have identified mutations in *B3GNT3* and *TREML1* potentially associated with increased risk of PD to be previously linked with neuroinflammation, further research into the biological mechanisms are critical to confirm the role of these genes in PD. Further replication in larger datasets that prioritize familial PD cases and individuals of non-European ancestry will provide greater insight into the nominated genes.

Acknowledgements

We would like to thank all of the subjects who donated their time and biological samples to be

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

part of this study. This study used the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health (http://hpc.nih.gov). Figure 1 was designed on Biorender.com. Data used in the preparation of this article were obtained from the AMP-PD Knowledge Platform. For up-to-date information on the study, visit https://www.amp-pd.org. 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, and Verily. We would like to thank AMP-PD for the publicly available whole-genome sequencing data, including cohorts from the Fox Investigation for New Discovery of Biomarkers (BioFIND), the Parkinson's Progression Markers Initiative (PPMI), and the Parkinson's Disease Biomarkers Program (PDBP). 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. Harvard Biomarker Study (HBS) is a collaboration of HBS investigators (full list of HBS investigators 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 manuscript. We also thank all of our Genentech colleagues involved in the Human Genetics Initiative involved in generating the sequence data including Natalie Bowers, Julie Hunkapiller, Jens Reeder, and Suresh Selvaraj. We are grateful to the Banner Sun Health Research Institute Brain and Body Donation Program of Sun City, Arizona, for the provision of human brain tissue and data. UKBEC: Consortium members include; Juan A. Botía, University of Murcia & UCL Great Ormond Street Institute of Child Health; Karishma D'Sa, Crick Institute; Paola Forabosco, Istituto di Ricerca Genetica e Biomedica, Italy; Sebastian Guelfi, Verge Genomics & UCL Great Ormond Street Institute of Child Health; Adaikalavan Ramasamy, Singapore Institute for Clinical Sciences; Regina H. Reynolds, UCL Great Ormond Street Institute of Child Health; Colin Smith, The University of Edinburgh; Daniah Trabzuni, UCL Queen Square Institute of Neurology; Robert Walker, The University of Edinburgh; Michael E. Weale, Genomics Plc, Oxford UK. This work was supported by the UK Dementia Research Institute which receives its funding from DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK.Medical Research Council (award number MR/N026004/1) and Medical Research Council (award number MR/N026004/1). The

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

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, 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 Research. We thank the NIH NeuroBioBank (https://neurobiobank.nih.gov) for providing human brain tissue samples and data. Wellderly: This work is supported by Scripps Research Translational Institute, an NIH-NCATS Clinical and Translational Science Award (CTSA; 5 UL1TR002550). LNG Path confirmed: We are grateful to the Banner Sun Health Research Institute Brain and Body Donation Program of Sun City, Arizona for the provision of human biological materials (or specific description, e.g. brain tissue, cerebrospinal fluid). The Brain and Body Donation Program has been 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, 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 Research. We thank the NIH NeuroBioBank for the provision of tissue samples. NABEC: We thank members of the North American Brain Expression Consortium (NABEC) for providing samples derived from brain tissue. Brain tissue for the NABEC cohort were obtained from the Baltimore Longitudinal Study on Aging at the Johns Hopkins School of Medicine, the NICHD Brain and Tissue Bank for Developmental Disorders at the University of Maryland, the Banner Sun Health Research Institute Brain and Body Donation Program, and from the University of Kentucky Alzheimer's Disease Center Brain Bank.

Funding

This research was supported in part by the Intramural Research Program of the National Institutes of Health (National Institute on Aging and National Institute of Neurological Disorders

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

and Stroke; project numbers: 1ZIA-NS003154, Z01-AG000949-02, Z01-ES101986, and UK ADC NIA P30 AG072946). This research has been conducted using the UK Biobank Resource under Application Number 33601.

Competing interests

HL, HI, MT, DV, 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 currently serves on the scientific advisory board for Clover Therapeutics and is an advisor to Neuron23 Inc. HRM is employed by UCL and in the last 24 months he reports paid consultancy from Biogen, Biohaven, Lundbeck; lecture fees/honoraria from Wellcome Trust, Movement Disorders Society. Research Grants from Parkinson's UK, Cure Parkinson's Trust, PSP Association, CBD Solutions, Drake Foundation, Medical Research Council, and Michael J Fox Foundation. HRM is also a co-applicant on a patent application related to C9ORF72 - Method for diagnosing a neurodegenerative disease (PCT/GB2012/052140). TB is employed by Genentech, Inc., a member of the Roche group. CB takes final responsibility for the decision to submit the paper for publication.

Supplementary material

Supplementary material is available at Brain online.

Appendix 1

List of UK Brain Expression Consortium members

John Hardy, Mike Weale, Daniah Trabzuni, Sebastian Guelfi, Juan Botia, Karishma D'Sa, Paola Forabosco, Colin Smith, Adaikalavan Ramasamy, Mina Ryten, Regina H. Reynolds, and Robert Walker. For a full list of UK Brain Expression Consortium (UKBEC) authors, see https://ukbec.wordpress.com/.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

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(12):1091-1102.

2. Foo JN, Chew EGY, Chung SJ, et al. Identification of Risk Loci for Parkinson Disease in Asians and Comparison of Risk Between Asians and Europeans: A Genome-Wide Association Study. *JAMA Neurol*. 2020;77(6):746-754.

3. Polymeropoulos MH, Lavedan C, Leroy E, et al. Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science*. 1997;276(5321):2045-2047.

4. Kitada T, Asakawa S, Hattori N, et al. Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. *Nature*. 1998;392(6676):605-608.

5. Singleton A, Hardy J. A generalizable hypothesis for the genetic architecture of disease: pleomorphic risk loci. *Hum Mol Genet*. 2011;20(R2):R158-R162.

6. Jansen IE, Gibbs JR, Nalls MA, et al. Establishing the role of rare coding variants in known Parkinson's disease risk loci. *Neurobiol Aging*. 2017;59:220.e11-e220.e18.

7. Gaare JJ, Nido G, Dölle C, et al. Meta-analysis of whole-exome sequencing data from two independent cohorts finds no evidence for rare variant enrichment in Parkinson disease associated loci. *PLoS One*. 2020;15(10):e0239824.

8. Rudakou U, Yu E, Krohn L, et al. Targeted sequencing of Parkinson's disease loci genes highlights SYT11, FGF20 and other associations. *Brain*. 2021;144(2):462-472.

9. Mencacci NE, Isaias IU, Reich MM, et al. Parkinson's disease in GTP cyclohydrolase 1 mutation carriers. *Brain*. 2014;137(Pt 9):2480-2492.

10. Lee JS, Kanai K, Suzuki M, et al. Arylsulfatase A, a genetic modifier of Parkinson's disease, is an α-synuclein chaperone. *Brain*. 2019;142(9):2845-2859.

11. Martin S, Smolders S, Van den Haute C, et al. Mutated ATP10B increases Parkinson's disease risk by compromising lysosomal glucosylceramide export. *Acta Neuropathol*. 2020;139(6):1001-1024.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

1 12. Makarious MB, Diez-Fairen M, Krohn L, et al. ARSA variants in α-synucleinopathies.
2 *Brain*. 2019;142(12):e70.

3 13. Fan Y, Mao CY, Dong YL, et al. ARSA gene variants and Parkinson's disease. *Brain*.
4 2020;143(6):e47.

5 14. Tesson C, Lohmann E, Devos D, Bertrand H, Lesage S, Brice A. Segregation of ATP10B
6 variants in families with autosomal recessive parkinsonism. *Acta Neuropathol*. 2020;140(5):783-
7 785.

8 15. Real R, Moore A, Blauwendraat C, Morris HR, Bandres-Ciga S, International
9 Parkinson's Disease Genomics Consortium (IPDGC). ATP10B and the risk for Parkinson's
10 disease. *Acta Neuropathol*. 2020;140(3):401-402.

11 16. Trabzuni D, United Kingdom Brain Expression Consortium (UKBEC), Thomson PC.
12 Analysis of gene expression data using a linear mixed model/finite mixture model approach:
13 application to regional differences in the human brain. *Bioinformatics*. 2014;30(11):1555-1561.

14 17. Gibbs JR, van der Brug MP, Hernandez DG, et al. Abundant quantitative trait loci exist
15 for DNA methylation and gene expression in human brain. *PLoS Genet*. 2010;6(5):e1000952.

16 18. Erikson GA, Bodian DL, Rueda M, et al. Whole-Genome Sequencing of a Healthy Aging
17 Cohort. *Cell*. 2016;165(4):1002-1011.

18 19. Iwaki H, Leonard HL, Makarious MB, et al. Accelerating Medicines Partnership:
19 Parkinson's Disease. Genetic Resource. *Mov Disord*. 2021;36(8):1795-1804.

20 20. Poplin R, Ruano-Rubio V, DePristo MA, et al. Scaling accurate genetic variant discovery
21 to tens of thousands of samples. *bioRxiv*. Published online November 14, 2017.
22 doi:10.1101/201178

23 21. Bandres-Ciga S, Saez-Atienzar S, Kim JJ, et al. Large-scale pathway specific polygenic
24 risk and transcriptomic community network analysis identifies novel functional pathways in
25 Parkinson disease. *Acta Neuropathol*. 2020;140(3):341-358.

26 22. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome
27 association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575.

28 23. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443.

24. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209.

25. Backman JD, Li AH, Marcketta A, et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature*. 2021;599(7886):628-634.

26. Chang D, Nalls MA, Hallgrímsdóttir IB, et al. A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat Genet*. 2017;49(10):1511-1516.

27. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.

28. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303.

29. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43(5):491-498.

30. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43:11.10.1-11.10.33.

31. Van der Auwera GA, O'Connor BD. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media; 2020.

32. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38(8):904-909.

33. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* . 2012;6(2):80-92.

34. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

1 35. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the
2 deleteriousness of variants throughout the human genome. *Nucleic Acids Res*.
3 2019;47(D1):D886-D894.

4 36. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association
5 studies. *Biostatistics*. 2012;13(4):762-775.

6 37. Lee S, Fuchsberger C, Kim S, Scott L. An efficient resampling method for calibrating
7 single and gene-based rare variant association analysis in case-control studies. *Biostatistics*.
8 2016;17(1):1-15.

9 38. Zeggini E, Morris A. *Assessing Rare Variation in Complex Traits: Design and Analysis*
10 *of Genetic Studies*. Springer; 2015.

11 39. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs
12 and statistical tests. *Am J Hum Genet*. 2014;95(1):5-23.

13 40. Lee S, Emond MJ, Bamshad MJ, et al. Optimal Unified Approach for Rare-Variant
14 Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing
15 Studies. *The American Journal of Human Genetics*. 2012;91(2):224-237.
16 doi:10.1016/j.ajhg.2012.06.007

17 41. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for
18 sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89(1):82-93.

19 42. Derkach A, Lawless JF, Sun L. Robust and powerful tests for rare variants using Fisher's
20 method to combine evidence of association from two or more complementary tests. *Genet*
21 *Epidemiol*. 2013;37(1):110-121.

22 43. Tysnes OB, Storstein A. Epidemiology of Parkinson's disease. *J Neural Transm*.
23 2017;124(8):901-905.

24 44. Liao J, Wu CX, Burlak C, et al. Parkinson disease-associated mutation R1441H in
25 LRRK2 prolongs the "active state" of its GTPase domain. *Proc Natl Acad Sci U S A*.
26 2014;111(11):4055-4060.

27 45. Tan MMX, Lawton MA, Jabbari E, et al. Genome-Wide Association Studies of Cognitive
28 and Motor Progression in Parkinson's Disease. *Mov Disord*. 2021;36(2):424-433.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

46. Pu JL, Lin ZH, Zheng R, et al. Association analysis of SYT11, FGF20, GCH1 rare variants in Parkinson's disease. *CNS Neurosci Ther*. 2022;28(1):175-177.

47. Kasten M, Hartmann C, Hampf J, et al. Genotype-Phenotype Relations for the Parkinson's Disease Genes Parkin, PINK1, DJ1: MDSGene Systematic Review. *Mov Disord*. 2018;33(5):730-741.

48. Klein C, Westenberger A. Genetics of Parkinson's disease. *Cold Spring Harb Perspect Med*. 2012;2(1):a008888.

49. Mosley RL, Hutter-Saunders JA, Stone DK, Gendelman HE. Inflammation and adaptive immunity in Parkinson's disease. *Cold Spring Harb Perspect Med*. 2012;2(1):a009381.

50. Calogero AM, Mazzetti S, Pezzoli G, Cappelletti G. Neuronal microtubules and proteins linked to Parkinson's disease: a relevant interaction? *Biol Chem*. 2019;400(9):1099-1112.

51. Pellegrini L, Wetzel A, Grannó S, Heaton G, Harvey K. Back to the tubule: microtubule dynamics in Parkinson's disease. *Cell Mol Life Sci*. 2017;74(3):409-434.

52. Ren Y, Zhao J, Feng J. Parkin binds to alpha/beta tubulin and increases their ubiquitination and degradation. *J Neurosci*. 2003;23(8):3316-3324.

53. Cartelli D, Aliverti A, Barbiroli A, et al. α-Synuclein is a Novel Microtubule Dynamase. *Sci Rep*. 2016;6:33289.

54. Dardiotis E, Siokas V, Pantazi E, et al. A novel mutation in TREM2 gene causing Nasu-Hakola disease and review of the literature. *Neurobiol Aging*. 2017;53:194.e13-e194.e22.

55. Feng CW, Chen NF, Sung CS, et al. Therapeutic Effect of Modulating TREM-1 via Anti-inflammation and Autophagy in Parkinson's Disease. *Front Neurosci*. 2019;13:769.

56. Piccio L, Buonsanti C, Cella M, et al. Identification of soluble TREM-2 in the cerebrospinal fluid and its association with multiple sclerosis and CNS inflammation. *Brain*. 2008;131(Pt 11):3081-3091.

57. Kim JJ, Bandres-Ciga S, Blauwendraat C, International Parkinson's Disease Genomics Consortium, Gan-Or Z. No genetic evidence for involvement of alcohol dehydrogenase genes in risk for Parkinson's disease. *Neurobiol Aging*. 2020;87:140.e19-e140.e22.

58. Buervenich S, Carmine A, Galter D, et al. A rare truncating mutation in ADH1C

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

(G78Stop) shows significant association with Parkinson disease in a large international sample. *Arch Neurol*. 2005;62(1):74-78.

59. García-Martín E, Diez-Fairen M, Pastor P, et al. Association between the missense alcohol dehydrogenase rs1229984T variant with the risk for Parkinson's disease in women. *J Neurol*. 2019;266(2):346-352.

60. Global Parkinson's Genetics Program. GP2: The Global Parkinson's Genetics Program. *Mov Disord*. 2021;36(4):842-851.

61. Campbell MC, Myers PS, Weigand AJ, et al. Parkinson disease clinical subtypes: key features & clinical milestones. *Ann Clin Transl Neurol*. 2020;7(8):1272-1283.

62. Mu J, Chaudhuri KR, Bielza C, de Pedro-Cuesta J, Larrañaga P, Martinez-Martin P. Parkinson's Disease Subtypes Identified from Cluster Analysis of Motor and Non-motor Symptoms. *Front Aging Neurosci*. 2017;9:301.

63. Sauerbier A, Jenner P, Todorova A, Chaudhuri KR. Non motor subtypes and Parkinson's disease. *Parkinsonism Relat Disord*. 2016;22 Suppl 1:S41-S46.

64. Singleton AB, Farrer M, Johnson J, et al. alpha-Synuclein locus triplication causes Parkinson's disease. *Science*. 2003;302(5646):841.

65. Scott AJ, Chiang C, Hall IM. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *Genome Res*. Published online September 20, 2021. doi:10.1101/gr.275488.121

66. Billingsley KJ, Ding J, Jerez PA, et al. Genome-Wide Analysis of Structural Variants in Parkinson Disease. *Ann Neurol*. Published online January 25, 2023. doi:10.1002/ana.26608

Figure Legends

Figure 1 Graphical representation of the Analysis Workflow. 1) Annotation was performed using VEP and four variant groups were selected: a) missense variants as defined by SnpEff, b) moderate or high impact variants as defined by SnpEff/SnpSift, c) high confidence LoF variants as defined by LOFTEE, and d) variants with either a CADD PHRED score > 20 or high confidence LoF variants as defined by LOFTEE. 2) Burden analysis was performed on each

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

dataset separately at rare (MAF<1%) and ultra-rare (MAF<0.1%) cut-offs. 3) Meta-analysis strategy 1 using only PD cases and controls, otherwise referred to as the "case-control" meta-analysis. 4) Meta-analysis strategy 2 using PD cases, PD proxy cases (siblings and parent), and controls, otherwise referred to as the "case-control-proxies" meta-analysis.

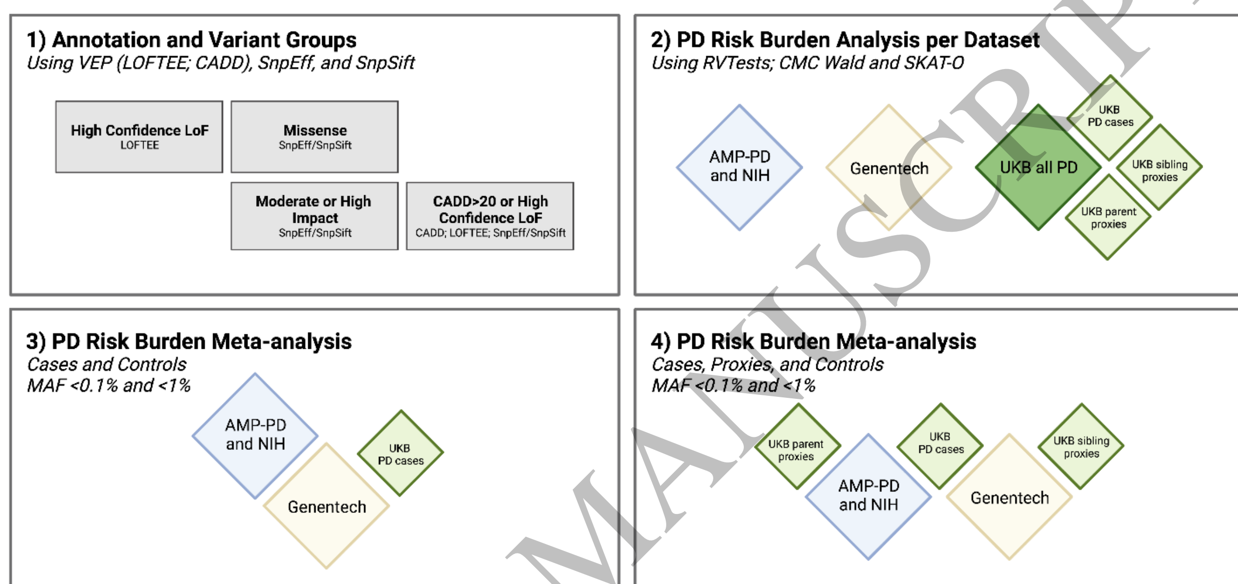

Figure 1
165x78 mm (x DPI)

Table 1 Datasets Overview after Quality Control

	Sample Size		Age[a] (Mean ± SD)		Sex (Male; %)	
Dataset	**Cases**	**Controls**	**Cases**	**Controls**	**Cases sex (Male; %)**	**Controls sex (Male; %)**
AMP-PD and NIH Genomes (Includes = PPMI, PDBP, HBS, BioFIND, NIH PD clinic, UKBEC, NABEC)	3369	4605	62.1 (11.8)	71.9 (16.2)	63.6	47.6
UKB case-control (WES)	1105	5643	62.9 (5.24)	64.1 (2.84)	62.4	47.6
UKB sibling proxy-control (WES)	668[b]	3463	62.2 (5.59)	64.1 (2.83)	45.5	49.5
UKB parent proxy-control (WES)	6033[b]	28 945	58.1 (7.23)	64.1 (2.82)	42.5	48.7
Genentech case-control (WGS)	2710	8994	64.7 (10.4)	59.2 (15.6)	59.2	40.7
Total	**7184 cases; 6701 proxies**	**51 650 controls**				

AMP-PD = Accelerating Medicines Partnership Parkinson's disease; NIH = National Institutes of Health; PPMI = Parkinson's Progression Markers Initiative; PDBP = Parkinson's disease Biomarkers Project; HBS = Harvard Biomarker Study; UKBEC = UK Brain Expression Consortium; NABEC = North American Brain Expression Consortium; UKB = UK Biobank; WES = Whole exome sequences; WGS = whole genome sequences.

[a]Age for AMP-PD and NIH datasets reported at recruitment or baseline, ages reported for UK Biobank datasets at recruitment, ages reported for Genentech at recruitment.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

1 [b]Indicates proxy cases.

2

3 **Table 2 Genes reaching exome-wide significance ($P < 1 \times 10^{-6}$) in MAF <1% in meta-analyses and individual datasets following**
4 **SKAT-O**
5

VARIANT CLASS (MAF <1%)	GENE	CASE ONLY META PVAL	CASE PROXIES META PVAL	AMP NIH PVAL	GNE PVAL	UKB CASE PVAL	UKB SIBLING PVAL	UKB PARENT PVAL
Missense	*GBA***	3.27×10^{-14}	1.46×10^{-21}	1.05×10^{-5}	1.32×10^{-8}	3.14×10^{-4}	0.247	2.15×10^{-10}
	*LRRK2**	7.15×10^{-7}	9.46×10^{-6}	1.96×10^{-7}	0.047	0.372	0.615	0.482
Moderate or High Impact	*GBA***	9.10×10^{-15}	1.32×10^{-22}	1.05×10^{-5}	5.70×10^{-8}	1.89×10^{-5}	0.073	2.15×10^{-10}
	*LRRK2**	7.23×10^{-7}	9.85×10^{-6}	2.09×10^{-7}	0.040	0.413	0.584	0.527
	TUBA1B	0.69	9.02×10^{-5}	NA	0.647	0.501	0.352	9.48×10^{-7}
LOF	*B3GNT3***	4.40×10^{-9}	3.36×10^{-9}	NA	4.40×10^{-9}	NA	NA	0.032
	*CAPN10***	3.60×10^{-7}	7.84×10^{-7}	NA	0.005	3.75×10^{-6}	0.053	0.394
CADD>20 or LOF	*GBA***	3.72×10^{-14}	9.12×10^{-22}	1.24×10^{-5}	6.99×10^{-8}	5.77×10^{-5}	0.130	2.15×10^{-10}
	*LRRK2**	2.49×10^{-7}	4.22×10^{-6}	2.23×10^{-7}	0.012	0.409	0.735	0.485
	ADH5	4.62×10^{-6}	6.15×10^{-5}	0.512	0.170	3.13×10^{-7}	0.491	0.768
	OR1G1	0.215	6.56×10^{-6}	0.848	0.029	0.620	6.58×10^{-7}	0.063

6 *Denotes genes that pass exome-wide significance ($P < 1 \times 10^{-6}$) in one meta-analysis
7 **Denotes genes that pass exome-wide significance ($P < 1 \times 10^{-6}$) in both meta-analyses
8

9 **Table 3 Genes reaching exome-wide significance ($P < 1 \times 10^{-6}$) in MAF < 0.1% in meta-analyses and individual datasets**
10 **following SKAT-O**

VARIANT CLASS (MAF <0.1%)	GENE	CASE ONLY META PVAL	CASE PROXIES META PVAL	AMP NIH PVAL	GNE PVAL	UKB CASE PVAL	UKB SIBLING PVAL	UKB PARENT PVAL
Missense	*GBA1**	1.86×10^{-5}	4.48×10^{-12}	0.022	2.30×10^{-2}	2.55×10^{-4}	5.41×10^{-3}	6.86×10^{-8}
Moderate or High Impact	*GBA1**	1.71×10^{-6}	4.87×10^{-16}	NA	0.088	1.13×10^{-6}	0.001	5.13×10^{-10}
	*AUNIP***	1.54×10^{-8}	2.70×10^{-7}	NA	0.023	3.04×10^{-8}	0.170	1
	TUBA1B	0.690	9.02×10^{-5}	NA	0.647	0.501	0.352	9.48×10^{-7}
	*TREML1**	0.048	3.58×10^{-7}	NA	0.010	0.858	0.001	1.41×10^{-5}
LOF	B3GNT3**	4.40×10^{-9}	3.36×10^{-9}	NA	4.40×10^{-9}	NA	NA	0.032
	*AUNIP***	1.64×10^{-8}	2.04×10^{-7}	NA	0.024	3.13×10^{-8}	0.116	1
	*CAPN10***	3.60×10^{-7}	7.84×10^{-7}	NA	0.005	3.75×10^{-6}	0.053	0.394
CADD>20 or LOF	*GBA1***	2.33×10^{-7}	1.20×10^{-14}	0.017	0.127	4.56×10^{-7}	8.93×10^{-4}	7.89×10^{-8}
	LRRK2	3.46×10^{-6}	2.65×10^{-6}	0.727	6.15×10^{-7}	0.044	0.771	0.014
	*AUNIP**	2.12×10^{-7}	1.53×10^{-6}	0.886	0.032	3.15×10^{-8}	0.125	1
	OR1G1	0.215	6.56×10^{-6}	0.848	0.029	0.620	6.58×10^{-7}	0.063

11 NA =
12 *Denotes genes that pass exome-wide significance ($P < 1 \times 10^{-6}$) in one meta-analysis.
13 **Denotes genes that pass exome-wide significance ($P < 1 \times 10^{-6}$) in both meta-analyses.

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awad214/7205395 by Catherine Sharp user on 04 July 2023

ACCEPTED MANUSCRIPT