

AWARD NUMBER: IAA-XAG16001-001-00000
LOG # BA150530

TITLE: Whole Genome Sequencing in Parkinson's disease

PRINCIPAL INVESTIGATOR: Dr. Andrew Singleton

CONTRACTING ORGANIZATION: National Institute of Aging, National Institutes of Health

REPORT DATE: JULY 2021

TYPE OF REPORT: FINAL

PREPARED FOR: U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE JULY 2021		2. REPORT TYPE FINAL		3. DATES COVERED 03-14-2016 to 03-13-2021	
4. TITLE AND SUBTITLE Whole Genome Sequencing in Parkinson's disease				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER XAG16001-001-00000	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Andrew Singleton E-Mail:singleta@mail.nih.gov				5d. PROJECT NUMBER BA150530	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) National Institute on Aging, Laboratory of Neurogenetics, 35 Lincoln Drive, Bethesda, MD 20892				8. PERFORMING ORGANIZATION REPORT	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Development Command Fort Detrick, Maryland 21702-5012				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S NUMBER(S)	
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT The NIA aimed to generate deep genome sequence of 3050 Parkinson's disease (PD) cases and 700 matched controls and to aggregate data for control genomes. All genomes were to be joint called to harmonize data. These data would be used within NIA to identify and fine map PD linked risk loci. The cumulative genetic data was to be used to identify putative risk variants which could then be tested for biological consequences. The data was to be distributed and made available through NIH databases. These goals have been achieved. More than 5,000 PD genomes were sequenced, and harmonized with genomic data from other series, from control cohorts, and from other diseases. These data have been made available through the LONI website (for PPMI samples) and through the Accelerating Medicines Partnership PD portal (now ~10,000 genomes). These data have been used as a part of an effort to more than double the known risk loci for PD (90 loci) and to attribute biological consequences to these loci. In addition, this work has contributed to the increased precision of disease predictive modeling using machine learning.					
15. SUBJECT TERMS Parkinson's disease, genetics, risk, prediction					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRDC
Unclassified	Unclassified	Unclassified	Unclassified	152	19b. TELEPHONE NUMBER (include area code)

TABLE OF CONTENTS

INTRODUCTION.....	3
KEYWORDS	4
ACCOMPLISHMENTS	4
IMPACT.....	5
CHANGES/PROBLEMS.....	5
PRODUCTS	5
MANUSCRIPTS	6
PREPRINTS.....	9
PARTICIPANTS & COLLABORATING ORGANIZATIONS.....	10
SPECIAL REPORTING REQUIREMENTS.....	10
APPENDICES.....	10

INTRODUCTION

The basis of this genetic work was to both directly provide insights into the genetic basis of Parkinson's disease and to provide a community resource for the PD field by making available whole genome sequence data in a responsible but accessible manner. The data were to be used to identify additional risk for disease, thereby providing insight into the mechanistic basis of PD, and opportunities for biological investigation of these mechanisms. The longitudinal nature of the patient derived data and the availability of ancillary data in subsets of patients also provided the opportunity to look at measures of progression, the

effect of genetics on biologics, and the use of genetics in prediction of disease. The underlying basis of all of this work then was to create a greater understanding of the genetic mechanisms influencing many facets of PD, in turn providing biological mechanistic insights.

KEYWORDS

Parkinson's disease, genome, sequencing, genetics, risk, prediction

ACCOMPLISHMENTS

The major goals of the project were to 1) Generate genome wide sequence data for PD; 2) Identify novel association signals; 3) Analyze the immediate biological effects of risk variability; 4) Build a resource for the research community.

These goals were each achieved. These genome sequence data were successfully generated, quality controlled, and grouped with other sequence data. These data have contributed to a doubling in the number of known genetic risk loci for PD during this period and these sequence data have contributed to the identification of new genetic causes of disease, and the assessment and clarification of risk/mutations reported by others. The sequence data has contributed to the ongoing fine mapping of 90 known risk loci and the consequent investigation of the biological effects of PD risk loci on proximal gene expression. In addition the data have been used to map and understand structural genomic variability including major rearrangements, deletions, duplications, and repeats. The data were used to build a community resource and the underlying sequence data is being used by vetted and approved researchers from around the world.

While this project was not initially intended for training and professional development this has become a major consequence of these data. We have generated a series of training

tools to allow others to use these genetic data in a cloud-based platform, Terra, as part of the AMP-PD initiative.

The results of this work have been disseminated through traditional publication (see below) and through the genomic databases listed above and below. The results of this work have also been presented at conferences and meetings.

IMPACT

A significant extension in the number of known risk loci involved in PD; contribution to the identification of disease-causing mutations; clarification of the role of putative disease-causing mutations; improvements in disease risk profiling; insight into the biological consequences of genetic risk for disease; and the development of a resource for the PD research community.

CHANGES/PROBLEMS

No significant problems were encountered; no significant changes were required

PRODUCTS

The availability of harmonized accessible human genetic data for PD patients and controls through the portals at LONI (for the Parkinson's Progression Marker Initiative; <https://www.ppmi-info.org/access-data-specimens/>) and at the Accelerating Medicines Partnership for PD (AMP-PD; <https://amp-pd.org/>). This is a community wide resource accessed by researchers from around the world.

Manuscripts:

1. Iwaki H, Leonard HL, Makarios MB, Bookman M, Landin B, Vismer D, Casey B, Gibbs JR, Hernandez DG, Blauwendraat C, Vitale D, Song Y, Kumar D, Dalgard CL, Sadeghi M, Dong X, Misquitta L, Scholz SW, Scherzer CR, Nalls MA, Biswas S, Singleton AB; Uniformed Services University of the Health Sciences Associates; AMP PD Whole Genome Sequencing Working Group; AMP PD consortium. **Accelerating Medicines Partnership: Parkinson's Disease. Genetic Resource.** *Mov Disord.* 2021 May 7. doi: 10.1002/mds.28549. Epub ahead of print. PMID: 33960523.
2. Bandres-Ciga S, Saez-Atienzar S, Kim JJ, Makarios MB, Faghri F, Diez-Fairen M, Iwaki H, Leonard H, Botia J, Ryten M, Hernandez D, Gibbs JR, Ding J, Gan-Or Z, Noyce A, Pihlstrom L, Torkamani A, Soltis AR, Dalgard CL; American Genome Center, Scholz SW, Traynor BJ, Ehrlich D, Scherzer CR, Bookman M, Cookson M, Blauwendraat C, Nalls MA, Singleton AB; International Parkinson Disease Genomics Consortium. **Large-scale pathway specific polygenic risk and transcriptomic community network analysis identifies novel functional pathways in Parkinson disease.** *Acta Neuropathol.* 2020 Sep;140(3):341-358. doi: 10.1007/s00401-020-02181-3. Epub 2020 Jun 29. Erratum in: *Acta Neuropathol.* 2021 May 4;: PMID: 32601912; PMCID: PMC8096770.
3. Chia R, Sabir MS, Bandres-Ciga S, Saez-Atienzar S, Reynolds RH, Gustavsson E, Walton RL, Ahmed S, Viollet C, Ding J, Makarios MB, Diez-Fairen M, Portley MK, Shah Z, Abramzon Y, Hernandez DG, Blauwendraat C, Stone DJ, Eicher J, Parkkinen L, Ansorge O, Clark L, Honig LS, Marder K, Lemstra A, St George-Hyslop P, Londos E, Morgan K, Lashley T, Warner TT, Jaunmuktane Z, Galasko D, Santana I, Tienari PJ, Myllykangas L, Oinas M, Cairns NJ, Morris JC, Halliday GM, Van Deerlin VM, Trojanowski JQ, Grassano M, Calvo A, Mora G, Canosa A, Floris G, Bohannon RC, Brett F, Gan-Or Z, Geiger JT, Moore A, May P, Krüger R, Goldstein DS, Lopez G, Tayebi N, Sidransky E; American Genome Center, Norcliffe-Kaufmann L, Palma JA, Kaufmann H, Shakkottai VG, Perkins M, Newell KL, Gasser T, Schulte C, Landi F, Salvi E, Cusi D, Masliah E, Kim RC, Caraway CA, Monuki ES, Brunetti M, Dawson TM, Rosenthal LS, Albert MS, Pletnikova O, Troncoso JC, Flanagan ME, Mao Q, Bigio EH, Rodríguez-Rodríguez E, Infante J, Lage C, González-Aramburu I, Sanchez-Juan P, Ghetti B, Keith J, Black SE, Masellis M, Rogaeva E, Duyckaerts C, Brice A, Lesage S,

Xiomerisiou G, Barrett MJ, Tilley BS, Gentleman S, Logroscino G, Serrano GE, Beach TG, McKeith IG, Thomas AJ, Attems J, Morris CM, Palmer L, Love S, Troakes C, Al-Sarraj S, Hodges AK, Aarsland D, Klein G, Kaiser SM, Woltjer R, Pastor P, Bekris LM, Leverenz JB, Besser LM, Kuzma A, Renton AE, Goate A, Bennett DA, Scherzer CR, Morris HR, Ferrari R, Albani D, Pickering-Brown S, Faber K, Kukull WA, Morenas-Rodriguez E, Lleó A, Fortea J, Alcolea D, Clarimon J, Nalls MA, Ferrucci L, Resnick SM, Tanaka T, Foroud TM, Graff-Radford NR, Wszolek ZK, Ferman T, Boeve BF, Hardy JA, Topol EJ, Torkamani A, Singleton AB, Ryten M, Dickson DW, Chiò A, Ross OA, Gibbs JR, Dalgard CL, Traynor BJ, Scholz SW. **Genome sequencing analysis identifies new loci associated with Lewy body dementia and provides insights into its genetic architecture.** Nat Genet. 2021 Mar;53(3):294-303. doi: 10.1038/s41588-021-00785-3. Epub 2021 Feb 15. PMID: 33589841; PMCID: PMC7946812.

4. Kim JJ, Makarious MB, Bandres-Ciga S, Gibbs JR, Ding J, Hernandez DG, Brooks J, Grenn FP, Iwaki H, Singleton AB, Nalls MA, Blauwendraat C; International Parkinson's Disease Genomics Consortium (IPDGC). **The Parkinson's Disease DNA Variant Browser.** Mov Disord. 2021 May;36(5):1250-1258. doi: 10.1002/mds.28488. Epub 2021 Jan 26. PMID: 33497488.
5. Blauwendraat C, Reed X, Krohn L, Heilbron K, Bandres-Ciga S, Tan M, Gibbs JR, Hernandez DG, Kumaran R, Langston R, Bonet-Ponce L, Alcalay RN, Hassin-Baer S, Greenbaum L, Iwaki H, Leonard HL, Grenn FP, Ruskey JA, Sabir M, Ahmed S, Makarious MB, Pihlstrøm L, Toft M, van Hilten JJ, Marinus J, Schulte C, Brockmann K, Sharma M, Siitonen A, Majamaa K, Eerola-Rautio J, Tienari PJ; 23andMe Research Team, Pantelyat A, Hillis AE, Dawson TM, Rosenthal LS, Albert MS, Resnick SM, Ferrucci L, Morris CM, Pletnikova O, Troncoso J, Grosset D, Lesage S, Corvol JC, Brice A, Noyce AJ, Masliah E, Wood N, Hardy J, Shulman LM, Jankovic J, Shulman JM, Heutink P, Gasser T, Cannon P, Scholz SW, Morris H, Cookson MR, Nalls MA, Gan-Or Z, Singleton AB. **Genetic modifiers of risk and age at onset in GBA associated Parkinson's disease and Lewy body dementia.** Brain. 2020 Jan 1;143(1):234-248. doi: 10.1093/brain/awz350. Erratum in: Brain. 2020 Apr 1;143(4):e33. Erratum in: Brain. 2020 Mar 1;143(3):e24. PMID: 31755958; PMCID: PMC6935749.
6. Bustos BI, Bandres-Ciga S, Gibbs JR, Krainc D, Mencacci NE, Gan-Or Z, Lubbe SJ; International Parkinson's Disease Genomics Consortium (IPDGC). **Replication**

- assessment of NUS1 variants in Parkinson's disease.** *Neurobiol Aging.* 2021 May;101:300.e1-300.e3. doi: 10.1016/j.neurobiolaging.2020.11.007. Epub 2020 Nov 13. PMID: 33309333.
7. Real R, Moore A, Blauwendraat C, Morris HR, Bandres-Ciga S; International Parkinson's Disease Genomics Consortium (IPDGC). **ATP10B and the risk for Parkinson's disease.** *Acta Neuropathol.* 2020 Sep;140(3):401-402. doi: 10.1007/s00401-020-02172-4. Epub 2020 Jun 15. PMID: 32556962; PMCID: PMC7540943.
 8. Diez-Fairen M, Makarious MB, Bandres-Ciga S, Blauwendraat C; International Parkinson's Disease Genomics Consortium (IPDGC). **Assessment of LIN28A variants in Parkinson's disease in large European cohorts.** *Neurobiol Aging.* 2021 Apr;100:118.e1-118.e3. doi: 10.1016/j.neurobiolaging.2020.12.002. Epub 2020 Dec 11. PMID: 33423828; PMCID: PMC7920911.
 9. Senkevich K, Bandres-Ciga S, Gan-Or Z, Krohn L; **International Parkinson's Disease Genomics Consortium (IPDGC). Lack of evidence for association of UQCRC1 with Parkinson's disease in Europeans.** *Neurobiol Aging.* 2021 May;101:297.e1-297.e4. doi: 10.1016/j.neurobiolaging.2020.10.030. Epub 2020 Nov 2. PMID: 33248804.
 10. Perifán MT, Gómez-Garre P, Blauwendraat C, Mir P, Bandres-Ciga S; **International Parkinson's Disease Genomics Consortium (IPDGC). The role of RHOT1 and RHOT2 genetic variation on Parkinson disease risk and onset.** *Neurobiol Aging.* 2021 Jan;97:144.e1-144.e3. doi: 10.1016/j.neurobiolaging.2020.07.003. Epub 2020 Jul 14. PMID: 32948353; PMCID: PMC7736199.
 11. Hall A, Quinn JP, Billingsley KJ. **Letter to the editor regarding "TGM6 variants in Parkinson's disease: clinical findings and functional evidence".** *J Integr Neurosci.* 2020 Dec 30;19(4):735-737. doi: 10.31083/j.jin.2020.04.261. PMID: 33378849.
 12. Grenn FP, Moore A, Bandres-Ciga S, Krohn L, Blauwendraat C; International Parkinson's Disease Genomics Consortium (IPDGC). **Assessment of ANG variants in Parkinson's disease.** *Neurobiol Aging.* 2021 Mar 24:S0197-4580(21)00097-X. doi: 10.1016/j.neurobiolaging.2021.03.006. Epub ahead of print. PMID: 33875291.
 13. Nalls MA, Blauwendraat C, Vallerga CL, Heilbron K, Bandres-Ciga S, Chang D, Tan M, Kia DA, Noyce AJ, Xue A, Bras J, Young E, von Coelln R, Simón-Sánchez J, Schulte C, Sharma M, Krohn L, Pihlstrøm L, Siitonen A, Iwaki H, Leonard H, Faghri F, Gibbs JR,

Hernandez DG, Scholz SW, Botia JA, Martinez M, Corvol JC, Lesage S, Jankovic J, Shulman LM, Sutherland M, Tienari P, Majamaa K, Toft M, Andreassen OA, Bangale T, Brice A, Yang J, Gan-Or Z, Gasser T, Heutink P, Shulman JM, Wood NW, Hinds DA, Hardy JA, Morris HR, Gratten J, Visscher PM, Graham RR, Singleton AB; 23andMe Research Team; System Genomics of Parkinson's Disease Consortium; International Parkinson's Disease Genomics Consortium. **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 Dec;18(12):1091-1102. doi: 10.1016/S1474-4422(19)30320-5. PMID: 31701892.

Preprints

1. Yuri L. Sosero, Sara Bandres-Ciga, Ziv Gan-Or, Lynne Krohn. **Analysis of PTRHD1 common and rare variants in European patients with Parkinson's disease.** medRxiv 2020.12.11.20243402; doi: <https://doi.org/10.1101/2020.12.11.20243402>
2. Prabhjyot Saini, Sara Bandres-Ciga, Jose Luis Alcantud, Clara Ruz, Ziv Gan-Or, for the International Parkinson's disease Genomics Consortium. **Common and rare variants in HFE are not associated with Parkinson's disease in Europeans.** medRxiv 2020.10.22.20217190; doi: <https://doi.org/10.1101/2020.10.22.20217190>
3. Mary B. Makarious, Hampton L. Leonard, Dan Vitale, Hirotaka Iwaki, Lana Sargent, Anant Dadu, Ivo Violich, Elizabeth Hutchins, David Saffo, Sara Bandres-Ciga, Jonggeol Jeff Kim, Yeajin Song, Matt Bookman, Willy Nojopranoto, Roy H. Campbell, Sayed Hadi Hashemi, Juan A. Botia, John F. Carter, Melina Maleknia, David W. Craig, Kendall Van Keuren-Jensen, Huw R. Morris, John A. Hardy, Cornelis Blauwendraat, Andrew B. Singleton, Faraz Faghri, Mike A. Nalls. **Multi-Modality Machine Learning Predicting Parkinson's Disease.** bioRxiv 2021.03.05.434104; doi: <https://doi.org/10.1101/2021.03.05.434104>

PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

Not applicable

SPECIAL REPORTING REQUIREMENTS

Not applicable

APPENDICES

Publically available manuscripts (13) and preprints (3) as listed above in the section on products.

RESEARCH ARTICLE

Accelerating Medicines Partnership: Parkinson's Disease. Genetic Resource

Hirota Iwaki, MD,^{1,2,3}  Hampton L. Leonard, MS,^{1,2,3} Mary B. Makarios, 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

¹Data Tecnica International, Glen Echo, Maryland, USA

²Center for Alzheimer's and Related Dementias, National Institute on Aging, Bethesda, Maryland, USA

³Laboratory of Neurogenetics, National Institute on Aging, Bethesda, Maryland, USA

⁴Verily Life Sciences, San Jose, California, USA

⁵Technome, Herndon, Virginia, USA

⁶The Michael J. Fox Foundation for Parkinson's Research, New York, New York, USA

⁷Sanofi, Framingham, Massachusetts, USA

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

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

¹⁰Northeastern University, Boston, Massachusetts, USA

¹¹Harvard Medical School, Brigham and Women's Hospital, Boston, Massachusetts, USA

¹²Publicis Sapient, Bethesda, Maryland, USA

¹³National Institute of Neurological Disorders and Stroke, Bethesda, Maryland, USA

¹⁴Department of Neurology, Johns Hopkins University, Baltimore, Maryland, USA

¹⁵Bristol 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

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

***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 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. © 2021 The Authors. *Movement Disorders* published by Wiley Periodicals LLC on behalf of International Parkinson and Movement Disorder Society. This article is a U.S. Government work and is in the public domain in the USA.

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

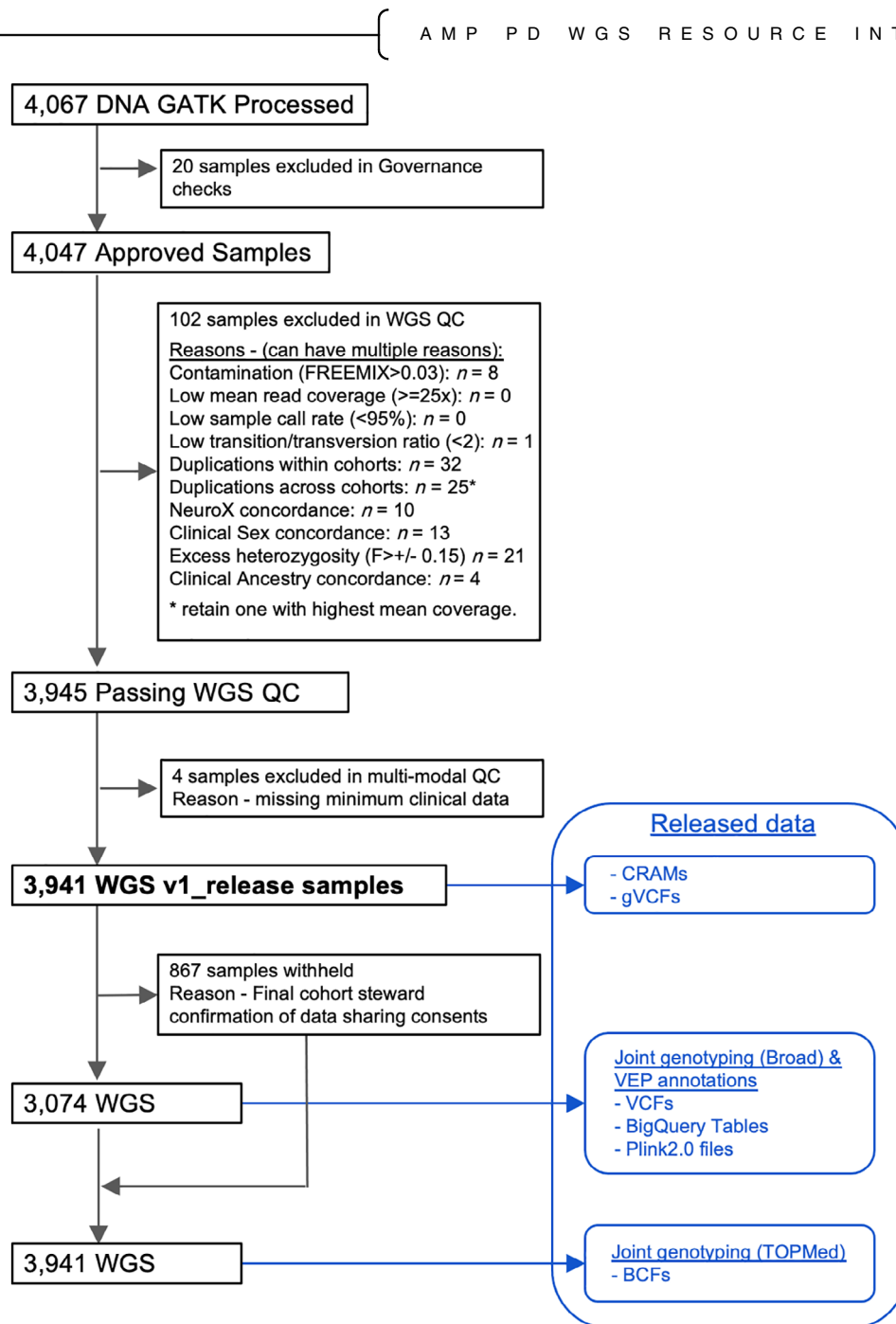


FIG 1. WGS sample flowchart. WGS, whole-genome sequencing; QC, quality control. [Color figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.com)]

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%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 pipeline⁷ 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” variants¹¹ 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 $\pm$ 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 GWAS¹ 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)
<i>ATP13A2</i>	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)
<i>GBA</i>	Pathogenic	3	8 (0, 0)	6 (0, 0)	2 (0, 0)
	Pathogenic+	6	300 (8, 0)	30 (1, 0)	5 (0, 0)
<i>LRRK2</i>	Pathogenic	5	467 (7, 0)	30 (1, 0)	9 (0, 0)
	Pathogenic+	5	467 (7, 0)	30 (1, 0)	9 (0, 0)
<i>PINK1</i>	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)
<i>PLA2G6</i>	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)
<i>PRKN</i>	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)
<i>SNCA</i>	Pathogenic	1	27 (0, 0)	0 (0, 0)	0 (0, 0)
	Pathogenic+	1	27 (0, 0)	0 (0, 0)	0 (0, 0)
<i>VPS35</i>	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 <i>GBA</i> , <i>LRRK2</i> , and <i>SNCA</i> 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.

^aP < 0.05. (t test).

^bP < 0.01.

^cP < 0.001.

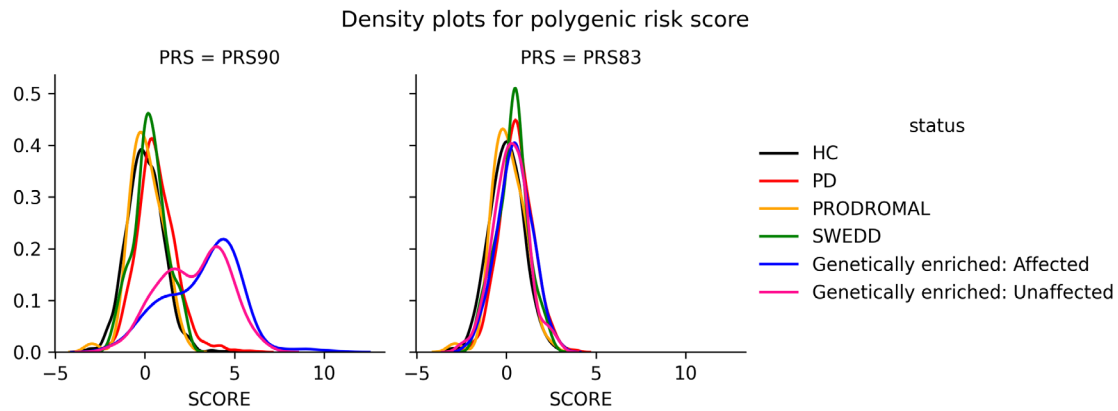


FIG 2. Density plots for polygenic risk score. HC, healthy volunteers; PD, participants with Parkinson's disease; SWEDD, scan without evidence of dopamine deficit. PRS90 is a weighted sum of the independent risk loci reported in Nalls et al (2019). PRS83 is the same but removing the 7 variants in *GBA*, *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/non-pathogenic 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](https://www.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, Valleria 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, Valleria 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.



Large-scale pathway specific polygenic risk and transcriptomic community network analysis identifies novel functional pathways in Parkinson disease

S. Bandres-Ciga¹ · S. Saez-Atienzar² · J. J. Kim¹ · M. B. Makarious¹ · F. Faghri¹ · M. Diez-Fairen³ · H. Iwaki¹ · H. Leonard¹ · J. Botia^{4,5} · M. Ryten⁶ · D. Hernandez¹ · J. R. Gibbs¹ · J. Ding¹ · Z. Gan-Or^{7,8,9} · A. Noyce¹⁰ · L. Pihlstrom¹¹ · A. Torkamani¹² · A. R. Soltis¹⁴ · C. L. Dalgard^{13,14} · The American Genome Center · S. W. Scholz^{15,16} · B. J. Traynor^{2,16} · D. Ehrlich¹⁷ · C. R. Scherzer¹⁸ · M. Bookman¹⁹ · M. Cookson²⁰ · C. Blauwendraat¹ · M. A. Nalls^{1,21} · A. B. Singleton¹ on behalf of the International Parkinson Disease Genomics Consortium

Received: 5 May 2020 / Revised: 7 June 2020 / Accepted: 14 June 2020
© The Author(s) 2020, corrected publication 2021

Abstract

Polygenic inheritance plays a central role in Parkinson disease (PD). A priority in elucidating PD etiology lies in defining the biological basis of genetic risk. Unraveling how risk leads to disruption will yield disease-modifying therapeutic targets that may be effective. Here, we utilized a high-throughput and hypothesis-free approach to determine biological processes underlying PD using the largest currently available cohorts of genetic and gene expression data from International Parkinson's Disease Genetics Consortium (IPDGC) and the Accelerating Medicines Partnership-Parkinson's disease initiative (AMP-PD), among other sources. We applied large-scale gene-set specific polygenic risk score (PRS) analyses to assess the role of common variation on PD risk focusing on publicly annotated gene sets representative of curated pathways. We nominated specific molecular sub-processes underlying protein misfolding and aggregation, post-translational protein modification, immune response, membrane and intracellular trafficking, lipid and vitamin metabolism, synaptic transmission, endosomal-lysosomal dysfunction, chromatin remodeling and apoptosis mediated by caspases among the main contributors to PD etiology. We assessed the impact of rare variation on PD risk in an independent cohort of whole-genome sequencing data and found evidence for a burden of rare damaging alleles in a range of processes, including neuronal transmission-related pathways and immune response. We explored enrichment linked to expression cell specificity patterns using single-cell gene expression data and demonstrated a significant risk pattern for dopaminergic neurons, serotonergic neurons, hypothalamic GABAergic neurons, and neural progenitors. Subsequently, we created a novel way of building de novo pathways by constructing a network expression community map using transcriptomic data derived from the blood of PD patients, which revealed functional enrichment in inflammatory signaling pathways, cell death machinery related processes, and dysregulation of mitochondrial homeostasis. Our analyses highlight several specific promising pathways and genes for functional prioritization and provide a cellular context in which such work should be done.

Keywords Parkinson disease · Polygenic risk · Transcriptome community maps · Mendelian randomization

S. Bandres-Ciga and S. Saez-Atienzar have contributed equally to the work.

We included the member of International Parkinson Disease Genomics Consortium in Acknowledgement section.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00401-020-02181-3>) contains supplementary material, which is available to authorized users.

Extended author information available on the last page of the article

Introduction

Although a great deal of progress in understanding the genetic underpinnings of familial and sporadic Parkinson disease (PD) has been made, the biological basis and cellular context of this risk remain unclear. We have learned that about 1–2% of PD is associated with a classical Mendelian inheritance pattern, while the majority of disease is driven by a complex set of factors in which polygenic risk seems to play a crucial role [3]. The fact that many of the genes that

contain disease-causing mutations also map within risk loci identified by genome-wide association studies (GWAS), supports the notion that common pathways are involved in both forms, and therefore, these pleomorphic genes might interact to regulate downstream common targets in both monogenic and non-monogenic PD [24].

Several common molecular processes have been suggested as critical in PD pathophysiology, including lysosome mediated autophagy, mitochondrial dysfunction, endosomal protein sorting and recycling, immune response, alpha-synuclein aggregation, lipid metabolism and synaptic transmission [2]. A goal in much of this work has been to unify the proteins encoded by PD-linked genes into common pathways. For instance, some success has been seen in this regard within the autosomal recessive genes *PINK1*, *PRKN*, and *DJ-1*, which share a common cellular mechanism: mitochondrial quality control and regulation [12, 18]. However, despite this success, the PD genetics field is still facing the challenge of understanding how genetic risk variants may disrupt biological processes and drive the underlying pathobiology of the disease. In the current era, using genetics to understand the disease process is a key milestone to facilitate the development of targeted therapies.

A priority in elucidating PD etiology lies in defining cumulative risk. GWAS continues to expand the number of genes and loci associated with disease [17], but the majority of these contributors individually exert small effects on PD risk. Current estimates of heritability explained by GWAS loci suggest that there is still an important component of risk yet to be discovered.

Here, we present a novel high-throughput and hypothesis-free approach to detect the existence of PD genetic risk linked to any particular biological pathway. We apply polygenic risk score (PRS) to a total of 2199 curated and well-defined gene sets representative of canonical pathways publicly available in the Molecular Signature Database v7.2 (MSigDB) [26] to define the cumulative effect of pathway-specific genetic variation on PD risk. To assess the impact of rare variation on PD risk explained by significant pathways, we perform gene-set burden analyses in an independent cohort of whole-genome sequencing (WGS) data, including 2101 cases and 2230 controls.

Additionally, we explore cell-type expression specificity enrichment linked to PD etiology by using single-cell RNA sequencing data from brain cells. Furthermore, we use graph-based analyses to generate de novo pathways that could be involved in disease etiology by constructing a transcriptome map of network communities based on RNA sequencing data derived from the blood of 1612 PD patients and 1042 healthy subjects.

Subsequently, we perform summary-data-based Mendelian randomization (SMR) analyses to prioritize genes from significant gene-sets by exploring possible genomic

associations with expression quantitative trait loci (eQTL) in public databases and nominate overlapping genes within our transcriptome communities for follow-up functional studies.

Finally, we present a user-friendly platform for the PD research community that enables easy and interactive access to these results (<https://pdgenetics.shinyapps.io/pathwaysbrowser/>).

Methods

Gene set selection representative of canonical pathways

The Molecular Signatures Database (MSigDB database v7.2) is a compilation of annotated gene sets from various sources such as online pathway databases, the biomedical literature, and manual curation by domain experts [15, 26]. We selected the collection “Canonical Pathways” composed of 2199 curated gene sets of pathways annotated from the following databases; Reactome (1499), KEGG (186), BIO-CARTA (289), Pathway Interaction Database (196), Matriosome project (10), Signaling Gateway (8), Sigma Aldrich (10), SuperArray SABiosciences (1) (<http://software.broadinstitute.org/gsea/msigdb>).

Genotyping data: cohort characteristics, quality control procedures, and study design

To assess PD risk, summary statistics from Chang et al. [7] PD GWAS meta-analysis involving 26,035 PD cases and 403,190 controls of European ancestry were used as the *reference dataset* for the primary analysis to define risk allele weights. In this study, there were 7,909,453 imputed SNPs tested for association with PD with a minor allele frequency (MAF) > 0.03. Recruitment and genotyping quality control procedures were described in the original report [7]. Individual-level genotyping data not included in Chang et al. [7] and from the last GWAS meta-analysis [17] was then randomly divided as the *training* and *testing datasets*. The *training dataset* used to construct the PRS consisted of 7218 PD cases and 9424 controls, while the *testing dataset* to validate the results consisted of 5429 PD cases and 5814 controls, all of European ancestry (see Fig. 1 for analysis workflow and rationale summary). Demographic and clinical characteristics of the cohorts under study are given in Supplementary Table 1, online resource.

Additional details of these cohorts, along with detailed quality control (QC) methods, can be found in Nalls et al. [17]. For sample QC, in short, individuals with low call rates (< 95%), discordance between genetic and reported sex, heterozygosity outliers (*F*-statistic cutoff of > -0.15 and < 0.15) and ancestry outliers (± 6 standard

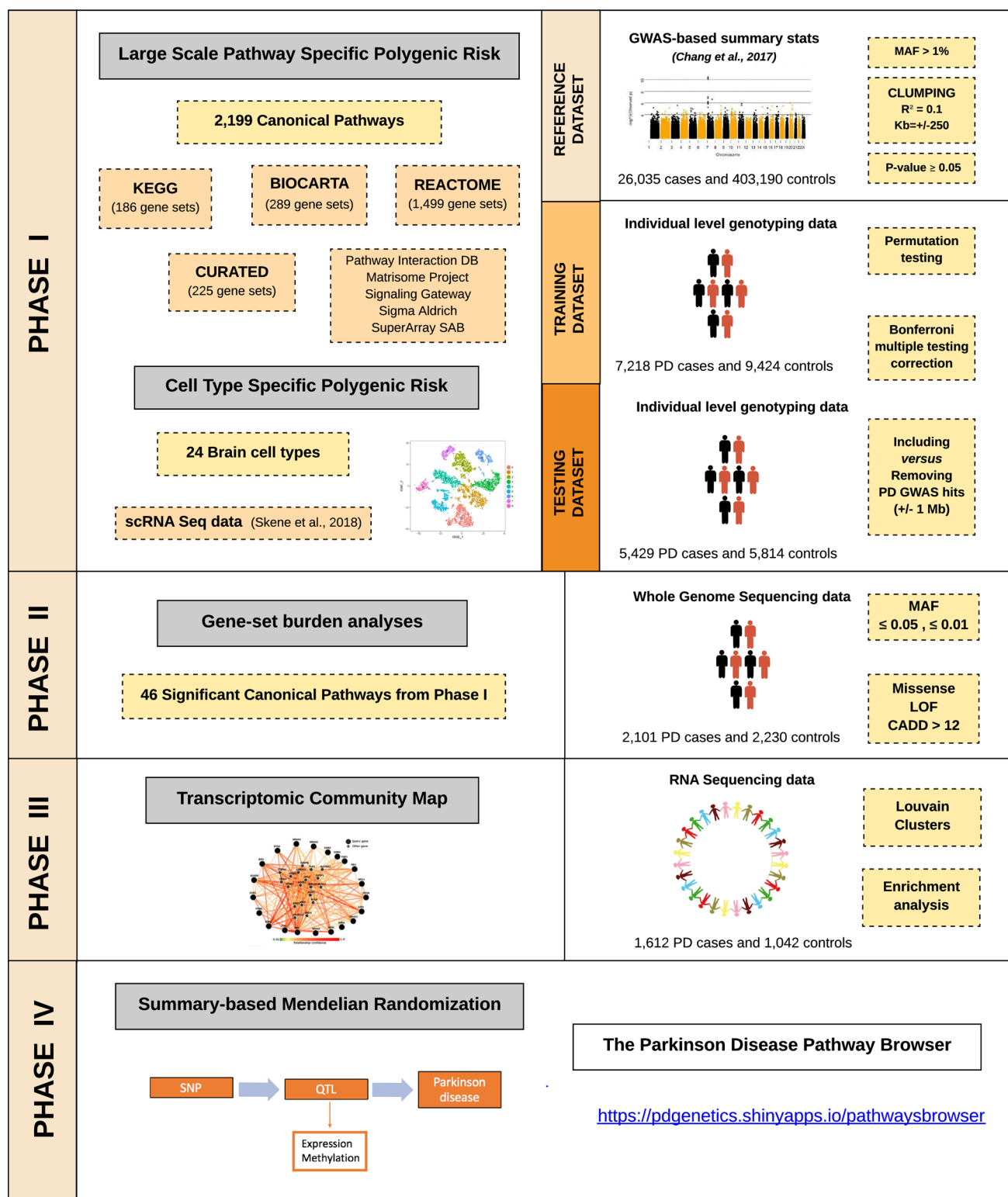


Fig. 1 Workflow and rationale summary

deviations from means of eigenvectors 1 and 2 of the 1000 Genomes phase 3 CEU and TSI populations from principal components) were excluded. Further, for genotype QC,

variants with a missingness rate of > 5%, minor allele frequency < 0.05, exhibiting Hardy–Weinberg Equilibrium (HWE) < $1E-5$ and palindromic SNPs were excluded.

Remaining samples were imputed using the Haplotype Reference Consortium (HRC) on the University of Michigan imputation server under default settings with Eagle v2.3 phasing based on *Haplotype Reference Consortium r1.1 2016* (<http://www.haplotype-reference-consortium.org>), and variants with an imputation quality ($R^2 > 0.3$) were included.

Polygenic effect scores for individual biological gene-sets versus PD risk

A polygenic effect score (PES) was generated to estimate polygenic risk for each of the 2199 gene sets representative of biological pathways and then tested for association with PD. PES was calculated based on the weighted allele dose as implemented in *PRSice2* (v2.1.1) (<https://github.com/choishingwan/PRSice>) [9]. Using the *reference dataset*, we selected variants with a summary statistic p value of association less than or equal to 0.05 and with $MAF > 1\%$. We extracted these variants from the *training dataset*, and linkage disequilibrium (LD) clumping was performed using the default $r^2 = 0.1$ and 250 Kb of distance. Then, 1000 permutations of sample labels were implemented to generate association p -value estimates for each gene-set. A p -value threshold = 0.05 was considered to prefilter the inclusion of variants in an effort to avoid overfitting when comparing across gene sets as well as improve computational efficiency. The permutation test in the *training dataset* provided a Nagelkerke's pseudo r^2 value after adjusting for an estimated prevalence of 0.005 (aged population estimate as per Gasser and colleagues), age at onset for cases and age at examination for controls, gender, and 20 PCs to account for population stratification. For those gene-sets surpassing Bonferroni multiple testing correction (p -value corrected = $0.05/2199$ gene-sets = $2.27E-5$), PES was then tested in an independent cohort (*testing dataset*) in a similar way, and overlapping gene-sets significantly associated with PD risk were reported. In an attempt to explore what biological processes were associated with PD risk after excluding known risk factors, the same analyses were performed after removing the 90 known PD GWAS hits [17] and additional SNPs located 1 Mb upstream and downstream from the signal. PES analyses considered that all the variants conferred risk under the additive model and did not cover regulatory regions adjacent to the up or downstream of the genes or intergenic variants.

Whole-genome sequencing data: cohort characteristics and quality control procedures

The following eight cohorts were utilized in this study; Biofind (<https://biofind.loni.usc.edu/>), NABEC [11], LNG Path confirmed, PDBP (<https://pdbp.ninds.nih.gov/>), NIH PD CLINIC, PPMI (<https://www.ppmi-info.org/>), WELLDERLY and UKBEC. Clinical and demographic

characteristics of the cohorts under study are summarised in Supplementary Table 2, online resource. Participants included sporadic PD cases clinically diagnosed by experienced neurologists. PD cases met criteria defined by the UK PD Society Brain Bank. This included 2101 cases and 2230 controls. All individuals were of European descent and were not age- or gender-matched.

DNA sequencing was performed using two vendors: Macrogen and USUHS. For samples sequenced at Macrogen, one microgram of each DNA sample was fragmented by the Covaris System and further prepared according to the Illumina TruSeq DNA Sample preparation guide to obtain a final library of 300–400 bp average insert size. Libraries were multiplexed and sequenced on the Illumina HiSeq X platform. For samples sequenced by USUHS, DNA samples were processed using the Illumina TruSeq DNA PCRFree Sample Preparation kit, starting with 500 ng input and resulting in an average insert size of 310 bp. USUHS processed single-libraries on single lanes on HiSeq X flow cells, and the Macrogen protocol used multiplexing. Paired-end read sequences were processed in accordance with the pipeline standard developed by the Centers for Common Disease Genomics [5]. The GRCh38DH reference genome was used for alignment as specified in the FE standardized pipeline [31]. The Broad Institute's implementation of this FE standardized pipeline, which incorporates the GATK [8] Best Practices is publicly available and used for WGS processing. Single-nucleotide (SNV) and InDel variants were called from the processed WGS data following the GATK [8] Best Practices [8] using the Broad Institute's workflow for joint discovery and variant quality score recalibration (VQSR). For quality control, each sample was checked using common methods for genotypes as well and sequence-related metrics. Using Plink v1.9 [6], each sample's genotype missingness rate ($< 95\%$), heterozygosity rate (exceeding ± 0.15 F -stat), and gender were checked. The King v2.1.3 kinship tool (8) was used to check for the presence of duplicate samples. Sequence and alignment related metrics generated by the Broad's implementation of the FE standardized pipeline were inspected for potential quality problems. This included the sample's mean sequence depth ($< 30\times$) and contamination rate ($> 2\%$), as reported by VerifyBamID (9), and single nucleotide variant count as reported by Picard's CollectVariantCallingMetrics (< 3 StDev) based on the sample's genomic vcf (gvcf). Principal components (PCs) were created for each dataset using PLINK. For the PC calculation, variants were filtered for minor allele frequency (> 0.01), genotype missingness (< 0.05), and HWE ($P \geq 1E-6$), and minor allele count < 3 . GCTA [33] was used to remove cryptically related at the level of first cousins or closer (sharing proportionally more than 12.5% of alleles).

Gene-set burden analyses

The sequence kernel association test-optimal (SKAT-O) [14] was implemented using default parameters in RVTESTS [35] to determine the difference in the aggregate burden of rare coding genetic variants (minor allele count ≥ 3) between PD cases and controls for the nominated gene-sets by PRS. SKAT-O was applied to aggregate genetic information across defined genomic regions to test for associations with gene-sets of interest under two frequency levels ($MAF \leq 0.03$ and $MAF \leq 0.01$) and three functional categories (missense, loss of function and Combined Annotation Dependent Depletion (CADD) score > 12 representing between 1 and 10% predicted most pathogenic variants in the genome). Covariates including gender, age at onset (cases), age at enrollment (controls), and 10 PCs were included to adjust the analyses. ANNOVAR was used for variant annotation [30].

Network expression community map in gene expression data

Baseline peri-diagnostic RNA sequencing data derived from the blood for 1612 PD patients and 1042 healthy subjects available from the Parkinson Progression Marker Initiative (PPMI) was used to construct a network of expression communities based on a graph model with Louvain clusters. This cleaned and normalized data was downloaded from the Accelerating Medicines Partnership for Parkinson's disease (AMP-PD) on March 1st, 2020. Library preparation, protocol, and transcriptomic quality control procedures can be found in detail in the original source <https://amp-pd.org/transcriptomics-data>. Prior to analyses, all data for the baseline visit were extracted. Data for each gene was then z -transformed to a mean of zero and a standard deviation of one. Scikit-learn's extraTreeClassifier option was used to extract coding gene features for inclusion in the network builds that are likely to contribute to classifying cases versus controls under default settings in the feature selection phase, leaving 8.3 k protein-coding genes for candidate networks [22]. Following this feature extraction phase, controls were excluded, and case-only correlations were calculated for all remaining gene features. Next, this correlation structure was converted to a graph object using NetworkX [28]. We filtered for network links at positive correlations (upregulated in cases together) between genes greater than or equal to 0.8. Subsequently, the Louvain algorithm was employed to build network communities within this graph object derived from the selected feature set [1].

Finally, pathway enrichment analysis within expression communities was performed to further dissect its biological function using the function `g:GOST` from `g:ProfileR` [19]. The significance of each pathway was tested by

hypergeometric tests with Bonferroni correction to calculate the error rate of each network.

Cell-type polygenic risk enrichment analysis

Single-cell RNA sequencing data [25] based on a total of 9970 cells obtained from several mouse brain regions (neocortex, hippocampus, hypothalamus, striatum, and midbrain) was used to explore cell types associated with PD risk. There are certainly differences between the mouse and the human brain. We used the package EWCE (v. 0.99.2) (<https://github.com/NathanSkene/EWCE>) to perform mouse to human homolog gene conversion. The package contains a dataset with the human orthologs of Mouse-Genomics-Informatics (MGI) mouse genes (mouse_to_human_homologs list). Out of the 14,579 mouse genes reported in Supplementary Table 4, Skene et al. [25], a total of 13,533 genes (92.82%) were converted to human HGNC symbols. Only genes with a high-confidence (1:1 mapping) were retained. As described in Skene et al. [25], a large fraction of non-matches is reasonable given evolutionary differences between humans and mice. The dataset described by Skene et al. [25], includes the specificity of expression for each gene within each cell type where values range from zero to one and represent the proportion of the total expression of a gene found in one cell type compared to all cell types. The closer the score is to 1, the more specific is the expression in that particular cell type. Taking this into account, PRS R^2 (variance) was calculated within each cell type using *PRSice2* (v2.1.1) as previously described in this manuscript. Cell type expression specificity levels ranging from 0 to 1 were then distributed in deciles. If a particular cell type is associated with PD risk, it is expected to observe a shift in the curve distribution with low PRS R^2 in non-specific gene sets (i.e., lower deciles) and a higher PRS R^2 in more specific gene sets (i.e., higher deciles). Linear regression adjusted by the number of SNPs included in the PRS was performed to assess the trend of increased PRS R^2 per decile of cell-type expression specificity.

Summary-data-based Mendelian randomization quantitative trait loci analyses

Two-sample SMR was applied to explore the enrichment of *cis* eQTLs within the 46 gene-sets nominated by our large-scale PRS analysis. The methodology can be interpreted as an analysis to test if the effect size of genetic variants influencing PD risk is mediated by gene expression or methylation to prioritize genes underlying these gene-sets for follow-up functional studies [37]. QTL association summary statistics from well-curated expression datasets were compared to Nalls et al. [17] summary statistics after extracting the gene-set-specific independent SNPs considered as the

instrumental variables. Expression datasets used for these analyses include estimates for cis-expression from the Genotype-Tissue Expression (GTEx) Consortium (v6; whole blood and 10 brain regions), the Common Mind Consortium (CMC; dorsolateral prefrontal cortex), the Religious Orders Study and Memory and Aging Project (ROSMAP), and the Brain eQTL Almanac project (Braineac; 10 brain regions). Additionally, we studied expression patterns in blood from the largest eQTL meta-analysis so far [29]. LD pruning and clumping were carried out using default SMR protocols (<http://cnsgenomics.com/software/smr>). Multi-SMR p-values (gene-level expression summaries for eQTLs) were adjusted by Bonferroni multiple test correction considering the number of genes tested per gene-set, and HEIDI was used to detect pleiotropic associations between the expression levels and PD risk that could be biasing the model at a p-value < 0.01 [32]. Effect estimates represent the change in PD odds ratio per one standard deviation increase in gene expression. Enrichment of cis expression was assessed per gene-set and per tissue. The number of genes tested per gene-set were Bonferroni corrected, and a Chisq test was applied to assess whether the proportion of QTLs per gene-set was significantly higher than expected by chance.

Results

Large-scale PES analysis nominates biological processes involved in PD risk

Out of the 2199 gene sets representative of biological processes included in this report, 279 gene-sets were significantly associated with PD risk in the *training phase* (Bonferroni threshold for significance $0.05/2,199 = 2.27E-5$) (Supplementary Table 3, online resource, <https://pdgenetics.shinyapps.io/pathwaysbrowser/>). Following the same analysis workflow, a total of 46 gene sets were replicated in the *testing phase* and nominated as potentially linked to PD risk through common genetic variation (Table 1, Fig. 2a, b).

Supplementary Table 4, (online resource) summarizes what SNPs within the 90 risk loci located up to 1 Mb upstream and downstream from the GWAS signal were included for each of the 2199 gene-sets as part of the large-scale polygenic risk score analyses for both the training and testing phases.

After excluding the 90 PD risk loci and SNPs located 1 Mb upstream and downstream from the GWAS hits, six gene sets including adaptive immune system, innate immune system, vesicle mediated transport, signaling by G protein-coupled receptors (GPCR) ligand binding, metabolism of lipids and neutrophil degranulation remained significant, suggesting as yet unidentified risk within these gene-sets (Bonferroni threshold for significance $2.27E-5$) (Table 2,

Fig. 2c, d). For an easy interpretation of these findings, significant gene-sets were clustered in hierarchies according to genetic redundancy, as highlighted in Supplementary Figs. 1, 2, online resource. Additionally, considering genetic pleiotropy across the 46 gene-sets, we prioritized the top 1% of genes involved in multiple pathways as a way of nominating promising PD candidate genes (Supplementary Table 5).

In an attempt to define etiological subtypes of PD, we performed Uniform Manifold Approximation and Projection for Dimension Reduction Analysis (UMAP) to explore the possibility of clustering different subgroups of patients that could be enriched for risk in certain molecular pathways. UMAP analysis showed two different clusters of patients according to the pathway-specific PES (subgroup 1 and subgroup 2; Supplementary Fig. 3a, online resource). Subgroup 1 was not enriched on any *LRRK2* G2019S carriers, while all patients from subgroup 2 ($N = 100$) were *LRRK2* G2019S carriers. When *LRRK2* gene boundaries were removed from the analysis and PES were calculated per individual, no subgroups were observed. We assume that since *LRRK2* G2019S is the main risk factor for PD, this variant overweights PES for those pathways in which *LRRK2* plays a role in (Supplementary Fig. 3b, online resource). This would suggest that pathway-specific PES by itself is not an accurate way to define etiological subgroups of the disease since association does not involve prediction. Future multimodality studies are necessary to increase discriminative accuracy given the heterogeneous nature of PD.

Gene-set-based burden analyses identifies gene-sets involved in PD risk through rare variation

To test whether the same biological processes are enriched by rare coding variants, we implemented gene-set based SKAT-O in a large WGS cohort composed of 2101 PD cases and 2230 controls. Out of the 46 gene-sets significantly associated with PD risk through common variation, 20 were linked through low-frequency genetic variation ($MAF \leq 3\%$) and 19 through rare variation ($MAF \leq 1\%$), at a p-value < 0.05 (Table 3). At a MAF threshold $\leq 3\%$, 12 gene-sets remained significantly associated with PD risk when focusing only on missense mutations, 4 when considering only loss of function variants and 6 when filtering by CADD score > 12 (~ among the 1–10% most pathogenic variants in the genome) (Table 3). At a MAF threshold $\leq 1\%$, 12 gene-sets remained significantly associated with PD risk when focusing only on missense mutations, four when considering only loss of function variants and five when filtering by CADD score > 12 (Table 3). Considering a more stringent p-value (Bonferroni threshold for significance $0.05/46$ gene-sets = 0.001), five gene sets including Alzheimer's disease, Parkinson's disease, Transmission across chemical synapses, Neuroactive ligand

Table 1 Canonical pathways significantly associated with PD risk in the *discovery* and *replication* phases through common variation

Gene set	N genes	Discovery			Replication						
		PRS R^2	Beta	SE	P	Num SNP	PRS R^2	Beta	SE	P	Num SNP
Activation of AMPK downstream of NMDARS (REACTOME)	29	0.0010	0.0936	0.0162	8.44E-09	18	0.0014	0.1088	0.0191	1.32E-08	9
Adaptive immune system (REACTOME)	811	0.0040	0.1867	0.0164	5.67E-30	455	0.0008	0.0815	0.0192	2.11E-05	169
Alpha synuclein pathway (PID)	32	0.0015	0.1157	0.0164	1.52E-12	34	0.0009	0.0902	0.0192	2.54E-06	17
Alzheimers disease (KEGG)	165	0.0023	0.1410	0.0163	5.91E-18	175	0.0017	0.1208	0.0192	3.34E-10	50
Amyloid fiber formation (REACTOME)	107	0.0031	0.1646	0.0164	7.60E-24	28	0.0019	0.1286	0.0193	2.40E-11	15
Apoptotic cleavage of cellular proteins (REACTOME)	38	0.0019	0.1289	0.0164	3.71E-15	41	0.0009	0.0890	0.0191	3.12E-06	23
Apoptotic execution phase (REACTOME)	52	0.0018	0.1243	0.0164	2.91E-14	48	0.0009	0.0867	0.0191	5.53E-06	26
Asparagine N linked glycosylation (REACTOME)	304	0.0020	0.1331	0.0163	3.21E-16	235	0.0008	0.0843	0.0191	1.01E-05	100
Caspase mediated cleavage of cytoskeletal proteins (REACTOME)	12	0.0021	0.1372	0.0164	5.91E-17	15	0.0009	0.0900	0.0191	2.47E-06	13
Chromatin organization (REACTOME)	272	0.0007	0.0773	0.0163	2.11E-06	201	0.0012	0.1004	0.0191	1.56E-07	79
Class B 2 secretin family receptors (REACTOME)	94	0.0015	0.1146	0.0163	1.98E-12	65	0.0011	0.0962	0.0191	4.84E-07	23
Clathrin mediated endocytosis (REACTOME)	145	0.0013	0.1079	0.0162	3.06E-11	151	0.0008	0.0810	0.0191	2.22E-05	81
COP1 dependent GOLGI to ER retrograde traffic (REACTOME)	99	0.0016	0.1187	0.0162	2.52E-13	76	0.0012	0.1008	0.0191	1.38E-07	30
COP1 mediated anterograde transport (REACTOME)	101	0.0008	0.0832	0.0162	3.00E-07	81	0.0015	0.1132	0.0191	3.23E-09	36
COP2 mediated vesicle transport (REACTOME)	68	0.0013	0.1070	0.0163	5.69E-11	46	0.0020	0.1325	0.0192	5.37E-12	20
ER to GOLGI anterograde transport (REACTOME)	154	0.0012	0.1008	0.0163	5.87E-10	114	0.0017	0.1203	0.0191	3.25E-10	48
Glutamate binding activation of AMPA receptors and synaptic plasticity (REACTOME)	31	0.0020	0.1308	0.0163	9.99E-16	48	0.0018	0.1252	0.0192	7.34E-11	12
GOLGI associated vesicle biogenesis (REACTOME)	56	0.0014	0.1118	0.0162	4.87E-12	61	0.0016	0.1151	0.0191	1.79E-09	32
GPCR ligand binding (REACTOME)	454	0.0017	0.1215	0.0163	9.49E-14	263	0.0010	0.0915	0.0191	1.65E-06	88
Innate immune system (REACTOME)	1104	0.0037	0.1790	0.0164	7.05E-28	677	0.0009	0.0870	0.0192	6.11E-06	281
Intra GOLGI and retrograde GOLGI to er traffic (REACTOME)	202	0.0019	0.1275	0.0163	4.52E-15	156	0.0011	0.0992	0.0191	2.19E-07	60
Intra GOLGI traffic (REACTOME)	44	0.0013	0.1049	0.0163	1.13E-10	36	0.0027	0.1526	0.0192	2.05E-15	12
LKB1 pathway (PID)	47	0.0011	0.0988	0.0162	1.12E-09	40	0.0012	0.1035	0.0191	6.25E-08	21
Long term depression (KEGG)	70	0.0017	0.1204	0.0163	1.35E-13	170	0.0010	0.0923	0.0192	1.47E-06	41
Lysosome (KEGG)	121	0.0009	0.0890	0.0164	5.51E-08	85	0.0010	0.0919	0.0192	1.67E-06	46
MAPK signaling pathway (KEGG)	267	0.0013	0.1058	0.0163	8.31E-11	322	0.0008	0.0851	0.0191	8.34E-06	84
Metabolism of lipids (REACTOME)	738	0.0029	0.1585	0.0163	2.99E-22	607	0.0009	0.0888	0.0192	3.58E-06	227
Metabolism of vitamins and cofactors (REACTOME)	189	0.0011	0.0965	0.0162	2.67E-09	201	0.0009	0.0891	0.0191	3.00E-06	86
Metabolism of water soluble vitamins and cofactors (REACTOME)	123	0.0009	0.0881	0.0162	5.47E-08	112	0.0008	0.0829	0.0191	1.38E-05	53
Neuroactive ligand receptor interaction (KEGG)	272	0.0020	0.1316	0.0163	6.58E-16	328	0.0011	0.0988	0.0191	2.34E-07	91
Neuronal system (REACTOME)	411	0.0044	0.1961	0.0164	7.03E-33	880	0.0012	0.0994	0.0191	1.92E-07	191
Neurotransmitter receptors and postsynaptic signal transmission (REACTOME)	204	0.0023	0.1430	0.0163	1.76E-18	390	0.0014	0.1079	0.0191	1.65E-08	98
Neutrophil degranulation (REACTOME)	478	0.0020	0.1316	0.0163	6.89E-16	294	0.0008	0.0847	0.0192	1.01E-05	128
P38 gamma delta pathway (PID)	11	0.0020	0.1334	0.0164	3.34E-16	13	0.0012	0.1035	0.0191	6.26E-08	11

Table 1 (continued)

Gene set	<i>N</i> genes	Discovery				Replication					
		PRS <i>R</i> ²	Beta	SE	<i>P</i>	Num SNP	PRS <i>R</i> ²	Beta	SE	<i>P</i>	Num SNP
Parkinsons disease (KEGG)	128	0.0029	0.1635	0.0168	2.09E−22	39	0.0023	0.1398	0.0192	3.01E−13	18
Post translational protein modification (REACTOME)	1429	0.0074	0.2568	0.0166	3.27E−54	1165	0.0011	0.0990	0.0191	2.20E−07	391
PTK6 promotes HIF1A stabilization (REACTOME)	6	0.0025	0.1580	0.0178	6.51E−19	18	0.0009	0.0897	0.0191	2.60E−06	13
Retrograde transport at the trans GOLGI network (reactome)	49	0.0008	0.0856	0.0163	1.52E−07	41	0.0008	0.0817	0.0192	1.98E−05	17
Signaling by GPCR (REACTOME)	1184	0.0027	0.1534	0.0164	7.91E−21	884	0.0010	0.0947	0.0191	7.58E−07	282
Snare interactions in vesicular transport (KEGG)	38	0.0008	0.0834	0.0162	2.64E−07	29	0.0009	0.0860	0.0191	6.67E−06	8
Trafficking of GLUR2 containing AMPA receptors (REACTOME)	17	0.0016	0.1184	0.0163	3.61E−13	37	0.0018	0.1249	0.0192	8.29E−11	10
Trans GOLGI network vesicle budding (REACTOME)	72	0.0015	0.1148	0.0162	1.31E−12	76	0.0016	0.1152	0.0191	1.79E−09	37
Transmission across chemical synapses (REACTOME)	269	0.0030	0.1634	0.0163	1.56E−23	517	0.0013	0.1055	0.0191	3.31E−08	124
Transport to the GOLGI and subsequent modification (REACTOME)	185	0.0018	0.1238	0.0163	3.14E−14	154	0.0016	0.1184	0.0191	6.02E−10	64
Vasopressin regulated water reabsorption (KEGG)	44	0.0012	0.1045	0.0163	1.43E−10	55	0.0016	0.1171	0.0192	1.03E−09	16
Vesicle mediated transport (REACTOME)	723	0.0044	0.1964	0.0164	5.90E−33	577	0.0018	0.1226	0.0192	1.57E−10	227

PRS R^2 variance explained by polygenic risk score, SE standard error, P P -value, num_SNP Number of SNPs included in each pathway analysis

receptor interaction and GPCR ligand binding remained significant at $MAF \leq 3\%$. When focusing on $MAF \leq 1\%$, the above mentioned gene sets in addition to Asparagine-N-glycosylation were significantly associated with PD risk.

After removing PD GWAS hits and SNPs located 1 Mb upstream and downstream, innate immune system and signaling by GPCR remained significantly associated with PD suggesting that rare variation within these gene-sets contributes to PD heritability (Supplementary Table 6, online resource).

In an effort to prioritize the top genes within significant gene-sets showing the highest cumulative effect on PD risk, individual gene-based SKAT-O analyses were performed considering a MAF threshold $\leq 3\%$ and three functional categories (missense, loss of function and CADD score > 12). Using this approach, gene-level prioritization is highlighted in Supplementary Table 7, online resource.

Transcriptome map reveals expression modules linked to PD etiology

Using Louvain community detection, we generated transcriptomic networks among PD cases. We identified 54 de novo expression communities (Supplementary Table 8, Supplementary Fig. 4, online resource). Overall, the communities generated were relatively robust, with a modularity score of 0.523 (modularity ranges from -1 to 1 , with closer to 1 suggesting stronger connectivity between network members). The 54 network communities were found to be enriched via hypergeometric tests after Bonferroni correction for processes relating to immune system response, ribosome RNA processing to the nucleus and cytosol, cell cycle, oxidative stress, and mitochondrial impairment (Fig. 3, Supplementary Table 9, online resource).

Dopaminergic neurons, serotonergic neurons and neural progenitors play a role on PD etiology

We used single-cell RNA sequencing data from 24 different brain cell types [25, 34]. For each of those cell types, genes were clustered into 10 gene sets according to the level of expression specificity, ranging from 0 to 1 (0 means that a gene is not expressed at all and 1 means the gene expression is highly specific for that cell type). Then, PRS was calculated per quintile of specificity within cells. Increased PRS R^2 , consistent with increased cell expression specificity, was observed for embryonic dopaminergic neurons, serotonergic neurons, hypothalamic GABAergic neurons, and neural progenitors at $P < 0.05$ in both the *training* and *replication* phases (Supplementary Table 10, online resource).

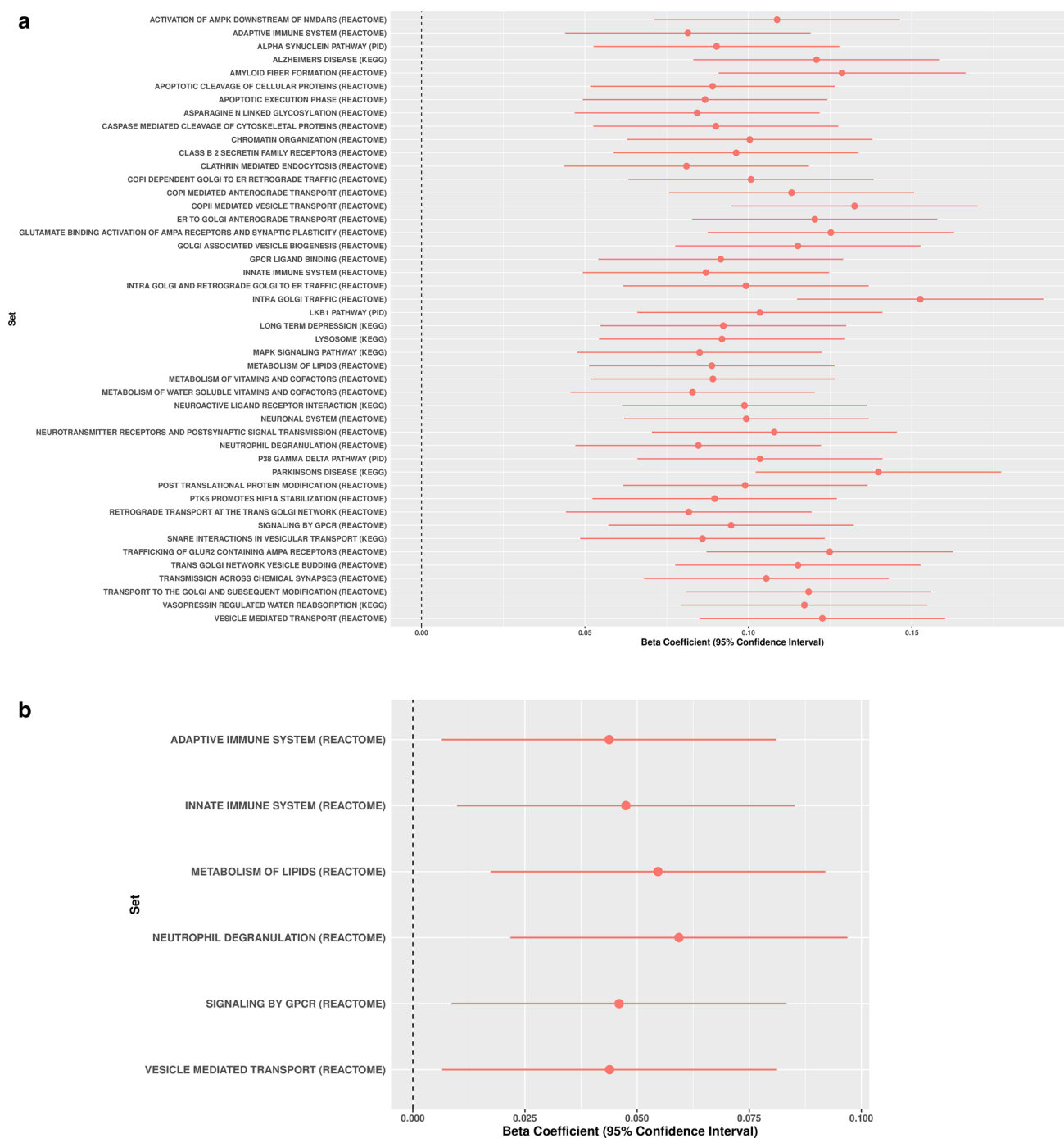


Fig. 2 Canonical pathways associated with Parkinson disease risk through common genetic variation based on PES analyses. Forest plots showing polygenic risk score estimates for the significant canonical pathways in the replication phase including (a) and removing (b) PD known risk loci $\pm$ 1Mb upstream and downstream. Estimates of variance explained by PRS for the significant canonical pathways including (c) or excluding (d) PD known risk loci $\pm$ 1Mb upstream and downstream

Mendelian randomization prioritizes pathways and genes based on their functional consequence

We aimed at nominating genes within significant gene-sets contributing to PD etiology by assessing changes in

expression across blood and brain. Out of the 46 gene-sets of interest, 7 showed a significant enrichment of QTLs more than expected by chance in the brain, 1 in substantia nigra and 11 in the blood (Supplementary Table 11, online resource).

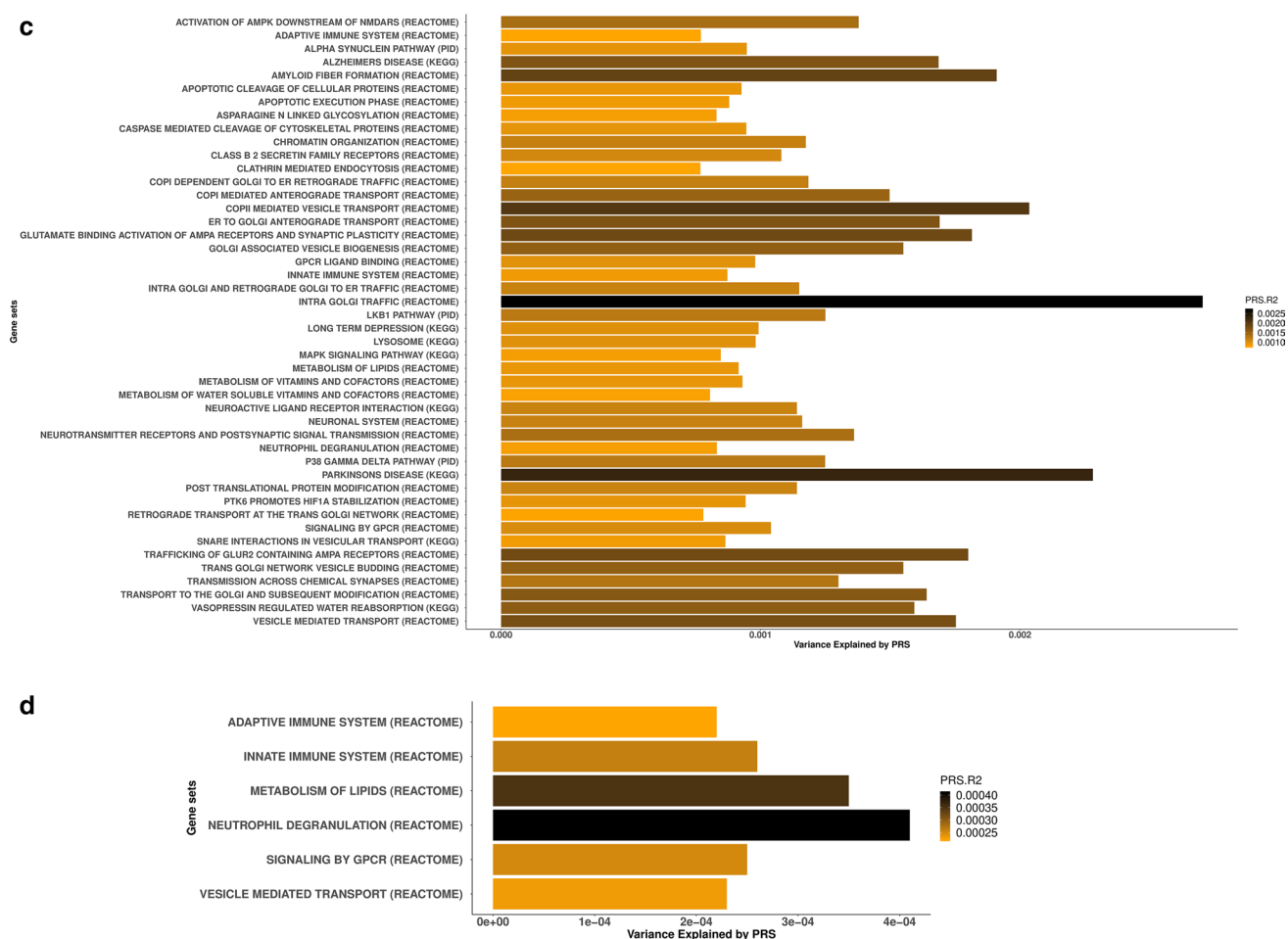


Fig. 2 (continued)

SMR revealed functional genomic associations with eQTLs in 201 genes (Supplementary Table 12, online resource) of which 88 were found to be part of the network communities significantly associated with PD in our transcriptome community map (Supplementary Table 13, online resource).

Discussion

Despite success at uncovering genetic risk factors associated with PD, our understanding of the molecular processes involved in disease is still limited. Using the largest genomic and transcriptomic PD cohorts currently available, our study sought to define both cumulative genetic risk and functional consequences linked to myriad biological pathways in an unbiased and data-directed manner. To our knowledge, there are no previous reports in the PD field where a similar approach has been implemented to explore the contribution of thousands of molecular processes on both the trigger

(risk) and the effect (expression changes) in a systematic manner.

Our large-scale PRS analysis identified multiple biological pathways associated with PD risk through common genetic variation. Overall, our results found that molecular processes underlying protein misfolding and aggregation, post-translational protein modification, immune response, membrane and intracellular trafficking, lipid metabolism, synaptic transmission, endosomal–lysosomal dysfunction and apoptosis mediated by initiator and executioner caspases are among the main contributors to PD etiology.

PD heritability remains incompletely deciphered by the genes and variants identified to date [17]. Here, we demonstrate that some of these significant gene-sets contribute to the heritability of PD outside of what is explained by current GWAS [17]. Notably, our genetic analyses provide definitive evidence for the role of several signal transduction mechanisms affecting adaptive and innate immune response, vesicular-mediated transport, and lipid metabolism on the risk for PD even after excluding PD known GWAS loci. The present study suggests that additional targets within these

Table 2 Canonical pathways significantly associated with PD risk through common variation in the *discovery* and *replication* phases after excluding PD known risk loci ± 1 Mb upstream and downstream

Gene set	Discovery			Replication						
	PRS R^2	Beta	SE	P	Num SNP	PRS R^2	Beta	SE	P	Num SNP
Adaptive immune system (REACTOME)	0.0028	0.1560	0.0163	1.42E-21	397	0.0002	0.0438	0.0190	2.15E-02	182
Innate immune system (REACTOME)	0.0026	0.1522	0.0163	1.12E-20	621	0.0003	0.0475	0.0192	1.34E-02	332
Vesicle mediated transport (REACTOME)	0.0022	0.1398	0.0163	1.03E-17	515	0.0002	0.0439	0.0190	2.12E-02	259
Signaling by GPCR (REACTOME)	0.0019	0.1299	0.0163	1.86E-15	816	0.0002	0.0460	0.0190	1.58E-02	329
Metabolism of lipids (REACTOME)	0.0018	0.1265	0.0163	8.08E-15	538	0.0004	0.0547	0.0190	4.08E-03	275
Neutrophil degranulation (REACTOME)	0.0012	0.1026	0.0163	2.73E-10	259	0.0004	0.0593	0.0192	1.96E-03	143

PRS R^2 variance explained by polygenic risk score, SE standard error, P P-value, num_SNP number of SNPs included in each pathway analysis

pathways are yet to be identified and prioritizes genes for follow-up functional studies.

A novel aspect of our study is that we nominate pathways whose implication on PD pathology has been poorly studied or debatable before. Our results support the hypothesis that chromatin remodeling and epigenetic mechanisms contribute to the development of PD [13]. An appropriate balance and distribution of active and repressed chromatin is required for proper transcriptional control, maintaining nuclear architecture and genomic stability, as well as regulation of the cell cycle [10]. Dysfunction in the epigenetic machinery has been shown to play a role in the etiology of a number of neurodegenerative and neurodevelopmental disorders either by genetic variation in an epigenetic gene or by changes in DNA methylation or histone modifications [13]. Similarly, our approach supports a role for vitamin metabolism on PD risk. Vitamins are crucial cofactors in the metabolism of carbohydrates, fat, and proteins, and vitamin deficiency has been widely proven to promote oxidative stress and neuro-inflammation [16].

Interestingly, some of the nominated gene-sets seem to span the etiological risk spectrum in which both common and rare variation contribute to PD susceptibility. In concordance with previous studies [21], our study identified an increased collective effect of rare lysosomal related variants in PD etiology. Additionally, we found evidence for a burden of rare damaging alleles in a range of specific processes, including neuronal transmission-related pathways and immune response.

The present study represents a significant step forward in our understanding of important connections between genetic factors, functional consequences and PD etiology. We constructed a transcriptome map by clustering de novo pathways relevant to disease pathology. Functional characterization analysis of these expression communities revealed that dysregulation of the immune system and inflammatory response including neutrophil degranulation, interferon alpha beta signaling, and other cytokine-related signaling pathways are key disease processes. Strikingly, when looking at molecular mechanisms significantly associated with PD risk, a cumulative effect of rare loss of function variants was found to be linked to disease through the adaptive immune system pathway. Both inflammation and autoimmune response have been widely studied with regard to PD etiology. Previous genetic studies have identified risk loci spanning key immune-associated genes such as *BST1* (*bone marrow stromal cell antigen 1*), a gene known to play role in neutrophil adhesion and migration, and *HLA* (human leukocyte antigen) [17, 23]. In support of this, it has been reported that α -synuclein-derived fragments act as antigenic epitopes displayed by HLA receptors, where both helper and cytotoxic T-cell responses are present in a high percentage of patients when tested [27].

Table 3 Association of canonical pathways and PD risk through rare variation

Gene set	Functional subcategory	MAF < 3		MAF < 1	
		Num SNP	SKAT-O P	Num SNP	SKAT-O P
Activation of AMPK downstream of NMDARS (REACTOME)	Missense	84	0.633	73	0.680
	Loss of function	31	0.015	28	0.892
	CADD > 12%	2	0.491	2	0.491
Adaptive immune system (REACTOME)	Missense	3369	0.061	3128	9.38E-02
	Loss of function	1724	0.003	1589	6.79E-03
	CADD > 12%	93	0.332	86	1
Alpha synuclein pathway (PID)	Missense	130	0.654	121	0.653
	Loss of function	82	0.023	77	1.32E-02
	CADD > 12%	3	0.006	3	5.95E-03
Alzheimers disease (KEGG)	Missense	594	2.15E-10	563	1.02E-06
	Loss of function	375	0.739	343	0.326
	CADD > 12%	13	0.075	13	0.075
Amyloid fiber formation (REACTOME)	Missense	308	0.026	289	0.204
	Loss of function	120	0.086	112	4.86E-02
	CADD > 12%	6	1	4	1
Apoptotic cleavage of cellular proteins (REACTOME)	Missense	356	0.073	332	0.686
	Loss of function	122	0.791	115	0.796
	CADD > 12%	5	0.616	5	0.616
Apoptotic execution phase (REACTOME)	Missense	397	0.111	372	0.631
	Loss of function	128	0.475	120	0.572
	CADD > 12%	7	0.794	7	0.794
Asparagine N linked glycosylation (REACTOME)	Missense	2956	0.021	2728	6.14E-04
	Loss of function	1226	0.233	1123	0.448
	CADD > 12%	69	0.544	68	0.427
Caspase mediated cleavage of cytoskeletal proteins (REACTOME)	Missense	159	0.313	151	0.516
	Loss of function	33	0.737	29	0.534
	CADD > 12%	1	0.691	NA	NA
Chromatin organization (REACTOME)	Missense	1213	0.284	1134	0.622
	Loss of function	680	0.813	628	0.777
	CADD > 12%	24	0.794	23	0.594
Class B 2 secretin family receptors (REACTOME)	Missense	327	0.452	298	0.080
	Loss of function	109	0.297	97	0.157
	CADD > 12%	15	0.525	13	0.592
Clathrin mediated endocytosis (REACTOME)	Missense	668	0.196	632	0.240
	Loss of function	410	0.256	363	0.892
	CADD > 12%	15	0.295	15	0.295
Copi dependent GOLGI to ER retrograde traffic (REACTOME)	Missense	576	0.474	525	0.237
	Loss of function	234	0.369	202	0.896
	CADD > 12%	8	0.081	8	0.081
COPI mediated anterograde transport (REACTOME)	Missense	595	0.711	547	0.568
	Loss of function	232	0.671	216	1
	CADD > 12%	13	1	13	1
COPII mediated vesicle transport (REACTOME)	Missense	327	0.742	300	3.13E-01
	Loss of function	149	0.729	140	0.417
	CADD > 12%	5	0.644	5	0.644
ER to Golgi anterograde transport (REACTOME)	Missense	900	0.307	826	3.42E-02
	Loss of function	378	0.426	356	0.808
	CADD > 12%	16	1	16	1

Table 3 (continued)

Gene set	Functional subcategory	MAF < 3		MAF < 1	
		Num SNP	SKAT-O P	Num SNP	SKAT-O P
Glutamate binding activation of AMPA receptors and synaptic plasticity (REACTOME)	Missense	83	0.222	75	0.399
	Loss of function	87	0.493	80	0.224
	CADD > 12%	5	0.236	5	0.236
GOLGI associated vesicle biogenesis (REACTOME)	Missense	277	0.224	260	0.428
	Loss of function	118	0.315	110	0.771
	CADD > 12%	8	1	8	1
GPCR ligand binding (REACTOME)	Missense	1769	2.12E-06	1632	2.16E-10
	Loss of function	438	0.185	395	0.445
	CADD > 12%	49	0.540	45	0.605
Innate immune system (REACTOME)	Missense	7162	0.009	6663	2.56E-03
	Loss of function	2965	0.122	2714	0.212
	CADD > 12%	178	0.819	169	0.521
Intra GOLGI and retrograde GOLGI to ER traffic (REACTOME)	Missense	1052	0.241	969	0.065
	Loss of function	441	0.853	389	0.405
	CADD > 12%	20	0.857	18	0.873
Intra GOLGI traffic (REACTOME)	Missense	213	0.284	199	0.077
	Loss of function	76	1	69	0.870
	CADD > 12%	5	0.801	5	0.801
Lkb1 pathway (PID)	Missense	233	0.114	224	0.440
	Loss of function	148	0.576	140	0.928
	CADD > 12%	6	0.270	4	1
Long-term depression (KEGG)	Missense	334	0.073	309	0.444
	Loss of function	195	0.164	166	0.818
	CADD > 12%	8	0.003	7	8.86E-03
Lysosome (KEGG)	Missense	673	0.034	628	0.528
	Loss of function	356	0.400	332	0.390
	CADD > 12%	29	0.272	28	0.837
Mapk signaling pathway (KEGG)	Missense	1156	0.091	1088	0.332
	Loss of function	669	0.691	616	1
	CADD > 12%	28	0.066	25	0.111
Metabolism of lipids (REACTOME)	Missense	3975	0.287	3701	0.204
	Loss of function	1881	0.713	1703	0.874
	CADD > 12%	149	0.782	138	0.640
Metabolism of vitamins and cofactors (REACTOME)	Missense	1252	0.210	1161	0.105
	Loss of function	593	0.765	527	0.872
	CADD > 12%	45	0.151	40	0.196
Metabolism of water soluble vitamins and cofactors (REACTOME)	Missense	740	0.487	688	0.187
	Loss of function	405	0.367	366	0.396
	CADD > 12%	29	0.287	26	0.114
Neuroactive ligand receptor interaction (KEGG)	Missense	1270	1.33E-09	1172	2.15E-10
	Loss of function	378	0.489	344	0.594
	CADD > 12%	37	0.260	33	0.814
Neuronal system (REACTOME)	Missense	1848	0.374	1704	0.606
	Loss of function	872	0.222	787	0.858
	CADD > 12%	47	0.011	40	9.62E-03
Neurotransmitter receptors and postsynaptic signal transmission (REACTOME)	Missense	753	0.931	696	0.870
	Loss of function	396	0.115	364	1

Table 3 (continued)

Gene set	Functional subcategory	MAF < 3		MAF < 1	
		Num SNP	SKAT-O P	Num SNP	SKAT-O P
Neutrophil degranulation (REACTOME)	CADD > 12%	16	0.006	15	7.76E-03
	Missense	2715	0.353	2525	0.111
	Loss of function	1207	0.766	1088	0.942
P38 gamma delta pathway (PID)	CADD > 12%	98	0.572	91	0.613
	Missense	62	0.386	56	0.391
	Loss of function	40	0.813	36	0.680
Parkinsons disease (KEGG)	CADD > 12%	2	0.664	2	0.664
	Missense	301	2.15E-10	276	2.72E-09
	Loss of function	185	0.024	175	9.14E-03
Post-translational protein modification (REACTOME)	CADD > 12%	11	0.289	11	0.289
	Missense	9055	0.011	8385	1.54E-03
	Loss of function	3643	0.675	3315	0.638
PTK6 promotes HIF1A stabilization (REACTOME)	CADD > 12%	213	0.471	199	0.467
	Missense	42	0.047	39	5.87E-03
	Loss of function	16	0.309	14	0.256
Retrograde transport at the trans GOLGI network (REACTOME)	CADD > 12%	2	0.170	2	0.170
	Missense	241	0.044	230	1.70E-02
	Loss of function	94	0.902	80	0.854
Signaling by GPCR (REACTOME)	CADD > 12%	3	0.499	2	0.792
	Missense	5789	0.002	5312	1.39E-05
	Loss of function	1623	0.388	1464	0.688
Snare interactions in vesicular transport (KEGG)	CADD > 12%	138	0.445	128	0.306
	Missense	108	0.316	101	0.087
	Loss of function	40	0.531	35	0.590
Trafficking of GLUR2 containing AMPA receptors (REACTOME)	CADD > 12%	3	0.034	2	0.072
	Missense	48	1	45	1
	Loss of function	49	0.095	43	0.920
Trans GOLGI network vesicle budding (REACTOME)	CADD > 12%			NA	NA
	Missense	330	0.118	313	0.229
	Loss of function	160	0.197	148	0.728
Transmission across chemical synapses (REACTOME)	CADD > 12%	11	1	11	1
	Missense	1117	0.441	1028	0.731
	Loss of function	572	0.134	518	0.771
Transport to the GOLGI and subsequent modification (REACTOME)	CADD > 12%	23	9.99E-04	21	7.88E-03
	Missense	1018	0.180	933	2.75E-02
	Loss of function	435	0.505	404	0.865
Vasopressin-regulated water reabsorption (KEGG)	CADD > 12%	19	0.672	19	0.672
	Missense	196	0.318	185	0.199
	Loss of function	84	0.650	78	0.439
Vesicle-mediated transport (REACTOME)	CADD > 12%	2	0.239	2	0.239
	Missense	3655	1	3395	0.484
	Loss of function	1757	1	1584	0.165
	CADD > 12%	87	0.818	80	1

MAF minor allele frequency, SKAT SNP-set Sequence Kernel Association Test, P P-value, CADD combined annotation dependent depletion, NA non applicable

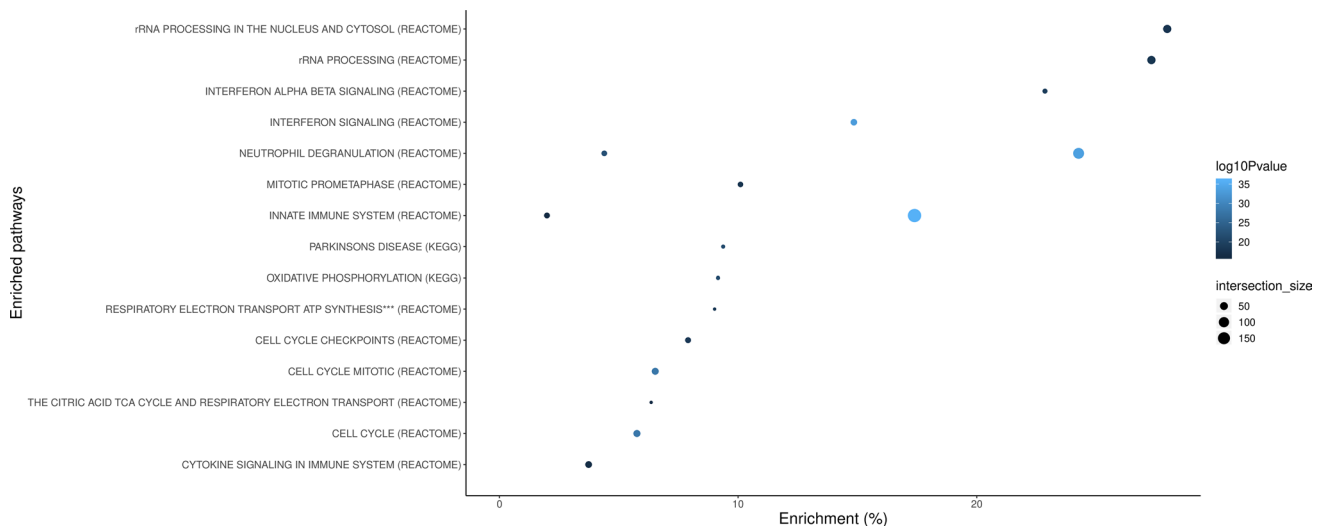


Fig. 3 Functional enrichment analyses of transcriptomic community maps. The x-axis represents the gene set enrichment (%) based on the community map gene lists. Intersection size denotes the number of

input genes within an enrichment category. Blue color indicates the adjusted association p -values on a $-\log_{10}$ scale. ***By chemiosmotic coupling and heat production by uncoupling proteins

Our analysis provides compelling evidence that dysregulation in genes that play a pivotal role in mitochondrial homeostasis exists in genetically complex PD. Despite not identifying these pathways as part of the stringent large-scale PRS analysis, our transcriptome community map showed an enrichment for the respiratory electron transport ATP synthesis by chemiosmotic coupling process and mitochondrial oxidative phosphorylation, in concordance with other reports [36]. Among the expression networks to highlight, it should be pointed out an enrichment in cell cycle and cell death machinery related processes and ribosome RNA processing to the nucleus and cytosol.

Our study aimed at pinpointing the specific drivers underlying these significant networks. Focusing on gene-sets linked to PD risk, SMR was applied to prioritize genes whose variation was found to be associated with expression changes linked with PD risk. Interestingly, we managed to replicate 88 of these genes after validating the functional consequence within our transcriptome community map.

Despite genetic efforts, it remains a matter of study in what cell types risk variants are active, which is essential for understanding etiology and experimental modeling. By integrating genetics and single-cell expression data, we found that PD risk is linked to expression specificity patterns in dopaminergic neurons, serotonergic neurons, hypothalamic GABAergic neurons, and neural progenitors, suggesting that these cell types disrupt biological networks that impact PD risk. Although our study failed at replicating specific enrichment patterns for oligodendrocytes and microglia as previously reported using other approaches [20], our results are in concordance with previous literature that applies various methodologies to gain similar conclusions [4].

The strengths of this study include an unbiased effort to link risk variants to biological pathways and characterize the functional consequence. While this study marks major progress in integrating human genetic and functional evidence, much remains to be established. A caveat of this study is that our approach was limited by the canonical gene sets publicly defined that were used for pathway analysis, and the relatively few brain regions studied for cell type analysis, which was based on mice data. We are aware that additional molecular networks and cell types from unsampled regions could contribute to PD. In addition, PRS analyses considered that all the variants conferred risk under the additive model and did not cover regulatory regions adjacent to the up or downstream of the genes or intergenic variants, which may be crucial for the disease. A further limitation of our study is that although we used state-of-the-art methodologies such as SMR to nominate candidate pathways and genes related to PD etiology, QTL datasets and associations are affected by both small sample size and low cis-SNP density. In addition, trans-QTL could not be assessed. Furthermore, our study focused on individuals of European ancestry, given that large sample sizes were required to create this resource. Replication in ancestrally diverse populations would be necessary for future studies. We also assume the limitation that gene redundancy might exist across the tested gene-sets and therefore overrepresentation of certain genes might lead to missing important gene-sets that in turn are associated with PD etiology. We anticipate that substantial collaborative efforts will lead to an improvement in statistical power and accuracy to define gene-sets linked to PD.

In conclusion, our high-throughput and hypothesis-free approach exemplifies a powerful strategy to provide valuable

mechanistic insights into PD etiology and pathogenesis. We highlight several promising pathways, cell types, and genes for further functional prioritization, aware that further in-depth investigation will be required to prove a definite link. As part of this study, we created a foundational resource for the PD community that can be applied to other neurodegenerative diseases with complex genetic etiologies (<https://pdgenetics.shinyapps.io/pathwaysbrowser/>). In future studies, linking specific phenotypic aspects of PD to pathways will constitute a critical effort using large longitudinal cohorts of well clinically characterized PD patients, with the hope of yielding disease-modifying therapeutic targets that are effective across PD subtypes.

Acknowledgements We would like to thank all of the subjects who donated their time and biological samples to be a part of this study. This work was supported in part by the Intramural Research Programs of the National Institute of Neurological Disorders and Stroke (NINDS), the National Institute on Aging (NIA), and the National Institute of Environmental Health Sciences both part of the National Institutes of Health, Department of Health and Human Services; project numbers Z01-AG000949-02 and Z01-ES101986. In addition this work was supported by the Department of Defense (award W81XWH-09-2-0128), and The Michael J Fox Foundation for Parkinson's Research. We would also like to thank all members of the International Parkinson Disease Genomics Consortium (IPDGC). For a complete overview of members, acknowledgements and funding, please see <http://pdgenetics.org/partners>. We would like to thank the Accelerating Medicines Partnership initiative (AMP-PD) for the publicly available whole genome sequencing data. This work utilized the computational resources of the NIH HPC Biowulf cluster. (<http://hpc.nih.gov>).

International Parkinson Disease Genomics Consortium We thank contributors who collected samples used in this initiative, as well as patients and families, whose help and participation made this work possible. Part of the data used for the analyses described in this manuscript was obtained from dbGaP at <http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap> through dbGaP accession phs000200, we thank members of the North American Brain Expression Consortium (NABEC) for providing DNA samples derived from brain tissue. Brain tissue for the NABEC cohort was obtained from the Baltimore Longitudinal Study on Aging at the Johns Hopkins School of Medicine, and from the NICHHD Brain and Tissue Bank for Developmental Disorders at the University of Maryland, Baltimore, MD, USA. We would like to thank the United Kingdom Brain Expression Consortium (UKBEC) for providing DNA samples. This study acknowledges the National Institute of Neurological Disorders and Stroke (NINDS) supported Parkinson's Disease Biomarkers Program Investigators (https://pdbp.ninds.nih.gov/sites/default/files/assets/PDBP_investigator_list.pdf). A full list of PDBP investigators can be found at <https://pdbp.ninds.nih.gov/policy>. Data and biospecimens used in the preparation of this manuscript were obtained from the Parkinson's Disease Biomarkers Program (PDBP) Consortium, part of the National Institute of Neurological Disorders and Stroke at the National Institutes of Health. Investigators include: Roger Albin, Roy Alcalay, Alberto Ascherio, Thomas Beach, Sarah Berman, Bradley Boeve, F. DuBois Bowman, Shu Chen, Alice Chen-Plotkin, William Dauer, Ted Dawson, Paula Desplats, Richard Dewey, Ray Dorsey, Jori Fleisher, Kirk Frey, Douglas Galasko, James Galvin, Dwight German, Lawrence Honig, Xuemei Huang, David Irwin, Kejal Kantarci, Anumantha Kanthasamy, Daniel Kaufer, James Leverenz, Carol Lippa, Irene Litvan, Oscar Lopez, Jian Ma, Lara Mangravite, Karen Marder, Laurie Ozelius, Vladislav

Petyuk, Judith Potashkin, Liana Rosenthal, Rachel Saunders-Pullman, Clemens Scherzer, Michael Schwarzschild, Tanya Simuni, Andrew Singleton, David Standaert, Debby Tsuang, David Vaillancourt, David Walt, Andrew West, Cyrus Zabetian, Jing Zhang, and Wenquan Zou. The PDBP Investigators have not participated in reviewing the data analysis or content of the manuscript. This work was supported by Scripps Research Translational Institute, an NIH-NCATS Clinical and Translational Science Award (CTSA; 5 UL1 RR025774). We are grateful to the NIH NeuroBioBank for providing brain tissue samples for Parkinson's disease cases. 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 (PI: Thomas G. Beach, MD). The Brain and Body Donation Program is supported by the National Institute of Neurological Disorders and Stroke (U24 NS072026 National Brain and Tissue Resource for Parkinson's Disease and Related Disorders), the National Institute on Aging (P30 AG19610 Arizona Alzheimer's Disease Core Center), the Arizona Department of Health Services (contract 211002, 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.

Author contributions (1) Research project: A. Conception (SBC, SSA, MN, CD, AS), B. Organization (SBC, SSA, MN), C. Execution (SBC, SSA, JK, MM, MN). Data generation: A. Experimental (RG, JD, CB, MN). (3) Statistical Analysis: A. Design (SBC, SSA, MN, FF), B. Execution (SBC, SSA, MM, MN, FF, JJ). (4) Manuscript Preparation: A. Writing of the first draft (SBC), B. Review and Critique (all authors).

Funding 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: project numbers 1ZIA-NS003154, Z01-AG000949-02 and Z01-ES101986). In addition, this work was supported by the Department of Defense (award W81XWH-09-2-0128), and The Michael J Fox Foundation for Parkinson's Research. C.R.S. was supported in part by NIH grants U01NS095736, U01NS100603, R01AG057331, and R01NS115144 and by the MJFF. Mike A. Nalls' participation is supported by a consulting contract between Data Tecnica International and the National Institute on Aging, NIH, Bethesda, MD, USA, as a possible conflict of interest, Dr. Nalls also consults for Neuron 23 s Inc, Lysosomal Therapeutics Inc, and Illumina Inc among others. C.R.S. is named as co-inventor on a US patent application on sphingolipids biomarkers that is jointly held by Brigham & Women's Hospital and Sanofi. C.R.S. has consulted for Sanofi Inc.; has collaborated with Pfizer, Opko, and Proteome Sciences, and Genzyme Inc.

Compliance with ethical standards

Conflict of interest No other disclosures were reported.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Combe D, Largeron C, Géry M, Egyed-Zsigmond E (2015) I-Louvain: an attributed graph clustering method. In: Fromont E, De Bie T, van Leeuwen M (eds) *Advances in intelligent data analysis. Lecture notes in computer science*, vol 9385. Springer, Cham
- Bandres-Ciga S, Diez-Fairen M, Kim JJ, Singleton AB (2020) Genetics of Parkinson's disease: an introspection of its journey towards precision medicine. *Neurobiol Dis* 137:104782
- Blauwendraat C, Nalls MA, Singleton AB (2020) The genetic architecture of Parkinson's disease. *Lancet Neurol* 19:170–178
- Bryoïs J, Skene NG, Hansen TF, Kogelman LJA, Watson HJ, Liu Z et al (2020) Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson's disease. *Nat Genet*. <https://doi.org/10.1038/s41588-020-0610-9>
- Centers for Common Disease Genomics. In: *Genome.gov*. <https://www.genome.gov/Funded-Programs-Projects/NHGRI-Genome-Sequencing-Program/Centers-for-Common-Disease-Genomics>. Accessed 12 Nov 2019
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4:7
- Chang D, Nalls MA, Hallgrímsdóttir IB, Hunkapiller J, van der Brug M, Cai F et al (2017) A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat Genet* 49(10):1511–1516
- Auwerda GAV der, Van der Auwerda GA, Carneiro MO, Hartl C, Poplin R, del Angel G et al (2013) From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinform* 11.10.1–11.10.33
- Euesden J, Lewis CM, O'Reilly PF (2015) PRSice: polygenic risk score software. *Bioinformatics* 31:1466–1468
- Frost B, Hemberg M, Lewis J, Feany MB (2014) Tau promotes neurodegeneration through global chromatin relaxation. *Nat Neurosci* 17:357–366
- Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai S-L et al (2010) Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet* 6:e1000952
- Hao L-Y, Giasson BI, Bonini NM (2010) DJ-1 is critical for mitochondrial function and rescues PINK1 loss of function. *Proc Natl Acad Sci USA* 107:9747–9752
- Hwang J-Y, Aromolaran KA, Zukin RS (2017) The emerging field of epigenetics in neurodegeneration and neuroprotection. *Nat Rev Neurosci* 18:347–361
- Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA et al (2012) Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 91:224–237
- Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27:1739–1740
- Lin MT, Beal MF (2006) Mitochondrial dysfunction and oxidative stress in neurodegenerative diseases. *Nature* 443:787–795
- Nalls MA, Blauwendraat C, Vallerga CL, Heilbron K, Bandres-Ciga S, Chang D et al (2019) Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol* 18:1091–1102
- Narendra DP, Jin SM, Tanaka A, Suen D-F, Gautier CA, Shen J et al (2010) PINK1 is selectively stabilized on impaired mitochondria to activate Parkin. *PLoS Biol* 8:e1000298
- Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H et al (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* 47:W191–W198
- Reynolds RH, Botía J, Nalls MA, Hardy J, International Parkinson's Disease Genomics Consortium (IPDGC), System Genomics of Parkinson's Disease (SGPD) et al (2019) Moving beyond neurons: the role of cell type-specific gene regulation in Parkinson's disease heritability. *NPJ Parkinsons Dis* 5:6
- Robak LA, Jansen IE, van Rooij J, Uitterlinden AG, Kraaij R, Jankovic J et al (2017) Excessive burden of lysosomal storage disorder gene variants in Parkinson's disease. *Brain* 140:3191–3203
- Saleh H (2018) *Machine learning fundamentals: use python and SCIKIT-learn to get up and running with the hottest developments in machine learning*. Packt Publishing Ltd, Birmingham
- Satake W, Nakabayashi Y, Mizuta I, Hirota Y, Ito C, Kubo M et al (2009) Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nat Genet* 41:1303–1307
- Singleton A, Hardy J (2011) A generalizable hypothesis for the genetic architecture of disease: pleomorphic risk loci. *Hum Mol Genet* 20:R158–R162
- Skene NG, Bryoïs J, Bakken TE, Breen G, Crowley JJ, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium et al (2018) Genetic identification of brain cell types underlying schizophrenia. *Nat Genet* 50:825–833
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102:15545–15550
- Sulzer D, Alcalay RN, Garretti F, Cote L, Kanter E, Agin-Liebes J et al (2017) T cells from patients with Parkinson's disease recognize α -synuclein peptides. *Nature* 546:656–661
- Torrents J, Ferraro F (2015) Structural cohesion: visualization and heuristics for fast computation with NetworkX and matplotlib. In: *Proceedings of the 14th python in science conference*
- Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B et al (2018) Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *Genomics* 228
- Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38:e164
- Website. CCDG. CCDG/Pipeline-Standardization. GitHub. <https://github.com/CCDG/Pipeline-Standardization>. (Accessed: 12th November 2019). Accessed 20 Nov 2019
- Wu Y, Zeng J, Zhang F, Zhu Z, Qi T, Zheng Z et al (2018) Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat Commun* 9:918
- Yang J, Hong Lee S, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 88:76–82
- Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, van der Zwan J et al (2018) Molecular architecture of the mouse nervous system. *Cell* 174:999.e22–1014.e22
- Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ (2016) RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* 32:1423–1426
- Zheng B, Liao Z, Locascio JJ, Lesniak KA, Roderick SS, Watt ML et al (2010) PGC-1 α , a potential therapeutic target for early intervention in Parkinson's disease. *Sci Transl Med* 2:52ra73
- Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE et al (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* 48:481–487

Affiliations

S. Bandres-Ciga¹  · S. Saez-Atienzar² · J. J. Kim¹ · M. B. Makarious¹ · F. Faghri¹ · M. Diez-Fairen³ · H. Iwaki¹ · H. Leonard¹ · J. Botia^{4,5} · M. Ryten⁶ · D. Hernandez¹ · J. R. Gibbs¹ · J. Ding¹ · Z. Gan-Or^{7,8,9} · A. Noyce¹⁰ · L. Pihlstrom¹¹ · A. Torkamani¹² · A. R. Soltis¹⁴ · C. L. Dalgard^{13,14} · The American Genome Center · S. W. Scholz^{15,16} · B. J. Traynor^{2,16} · D. Ehrlich¹⁷ · C. R. Scherzer¹⁸ · M. Bookman¹⁹ · M. Cookson²⁰ · C. Blauwendraat¹ · M. A. Nalls^{1,21} · A. B. Singleton¹ on behalf of the International Parkinson Disease Genomics Consortium

✉ A. B. Singleton
singleta@mail.nih.gov

¹ Molecular Genetics Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD 20892, USA

² Neuromuscular Diseases Research Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD 20892, USA

³ Fundació Docència i Recerca Mútua Terrassa and Movement Disorders Unit, Department of Neurology, University Hospital Mútua Terrassa, Terrassa, 08221 Barcelona, Spain

⁴ Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, Murcia, Spain

⁵ Department of Molecular Neuroscience, UCL, Institute of Neurology, London, UK

⁶ Department of Neurodegenerative Disease, University College London (UCL) Institute of Neurology, London, UK

⁷ Department of Neurology and Neurosurgery, McGill University, Montréal, QC, Canada

⁸ Montreal Neurological Institute, McGill University, Montréal, QC, Canada

⁹ Department of Human Genetics, McGill University, Montréal, QC, Canada

¹⁰ Preventive Neurology Unit, Wolfson Institute of Preventive Medicine, Queen Mary University of London and Department of Neurology, Royal London Hospital, London, UK

¹¹ Department of Neurology, Oslo University Hospital, Oslo, Norway

¹² The Scripps Research Institute, La Jolla, CA 92037, USA

¹³ Department of Anatomy, Physiology & Genetics, Uniformed Services University of the Health Sciences, Bethesda, MA, USA

¹⁴ The American Genome Center, Collaborative Health Initiative Research Program, Uniformed Services University of the Health Sciences, Bethesda, MA, USA

¹⁵ Neurodegenerative Diseases Research Unit, National Institute of Neurological Disorders and Stroke, Bethesda, MD 20892, USA

¹⁶ Department of Neurology, Johns Hopkins University Medical Center, Baltimore, MD 21287, USA

¹⁷ Parkinson's Disease Clinic, Office of the Clinical Director, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892, USA

¹⁸ Center for Advanced Parkinson Research, Harvard Medical School, Brigham and Women's Hospital, Boston, MA 0115, USA

¹⁹ Verily Life Sciences, South San Francisco, CA, USA

²⁰ Cell Biology and Gene Expression Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MA, USA

²¹ Data Tecnica International, Glen Echo, MD 20812, USA

COVID-19 Information[Public health information \(CDC\)](#)[Research information \(NIH\)](#)[SARS-CoV-2 data \(NCBI\)](#)[Prevention and treatment information \(HHS\)](#)[Español](#)

FULL TEXT LINKS

Multicenter Study [Nat Genet.](#) 2021 Mar;53(3):294-303. doi: 10.1038/s41588-021-00785-3.

Epub 2021 Feb 15.

Genome sequencing analysis identifies new loci associated with Lewy body dementia and provides insights into its genetic architecture

[FOLLOW NCBI](#)[Follow NLM](#)

National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894

[Copyright](#)[FOIA](#)[Privacy](#)[Help](#)[Accessibility](#)[Careers](#)[NLM](#) [NIH](#) [HHS](#) [USA.gov](#)

Ruth Chia ^{# 1}, Marya S Sabir ^{# 2}, Sara Bandres-Ciga ³, Sara Saez-Atienzar ¹,
Regina H Reynolds ^{4 5 6}, Emil Gustavsson ^{5 6}, Ronald L Walton ⁷, Sarah Ahmed ²,
Coralie Viollet ^{8 9}, Jinhui Ding ¹⁰, Mary B Makarios ², Monica Diez-Fairen ¹¹, Makayla K Portley ²,
Zalak Shah ², Yevgeniya Abramzon ^{1 12}, Dena G Hernandez ³, Cornelis Blauwendraat ³,
David J Stone ¹³, John Eicher ¹⁴, Laura Parkkinen ¹⁵, Olaf Ansorge ¹⁵, Lorraine Clark ¹⁶,
Lawrence S Honig ¹⁷, Karen Marder ¹⁷, Afina Lemstra ¹⁸, Peter St George-Hyslop ^{19 20},
Elisabet Londos ²¹, Kevin Morgan ²², Tammam Lashley ^{4 23}, Thomas T Warner ^{12 23},
Zane Jaunmuktane ²³, Douglas Galasko ^{24 25}, Isabel Santana ^{26 27 28 29}, Pentti J Tienari ^{30 31},
Liisa Myllykangas ^{32 33}, Minna Oinas ³⁴, Nigel J Cairns ³⁵, John C Morris ³⁵,
Glenda M Halliday ^{36 37 38}, Vivianna M Van Deerlin ³⁹, John Q Trojanowski ³⁹,
Maurizio Grassano ^{1 40}, Andrea Calvo ^{40 41}, Gabriele Mora ⁴², Antonio Canosa ^{40 41},
Gianluca Floris ⁴³, Ryan C Bohannon ⁴⁴, Francesca Brett ⁴⁵, Ziv Gan-Or ⁴⁶, Joshua T Geiger ²,
Anni Moore ¹⁰, Patrick May ⁴⁷, Rejko Krüger ^{47 48 49}, David S Goldstein ⁵⁰, Grisel Lopez ⁵¹,
Nahid Tayebi ⁵¹, Ellen Sidransky ⁵¹, American Genome Center; Lucy Norcliffe-Kaufmann ⁵²,
Jose-Alberto Palma ⁵², Horacio Kaufmann ⁵², Vikram G Shakkottai ⁵³, Matthew Perkins ⁵⁴,
Kathy L Newell ⁵⁵, Thomas Gasser ⁵⁶, Claudia Schulte ⁵⁶, Francesco Landi ⁵⁷, Erika Salvi ⁵⁸,
Daniele Cusi ⁵⁹, Eliezer Masliah ⁶⁰, Ronald C Kim ⁶¹, Chad A Caraway ⁶², Edwin S Monuki ⁶³,
Maura Brunetti ⁴⁰, Ted M Dawson ^{64 65 66 67}, Liana S Rosenthal ⁶⁴, Marilyn S Albert ⁶⁴,
Olga Pletnikova ^{68 69}, Juan C Troncoso ⁶⁸, Margaret E Flanagan ^{70 71}, Qinwen Mao ^{70 71},
Eileen H Bigio ^{70 71}, Eloy Rodríguez-Rodríguez ⁷², Jon Infante ⁷², Carmen Lage ⁷²,
Isabel González-Aramburu ⁷², Pascual Sanchez-Juan ⁷², Bernardino Ghetti ⁵⁵, Julia Keith ⁷³,
Sandra E Black ^{74 75 76 77 78}, Mario Masellis ^{77 78 79 80}, Ekaterina Rogaeva ⁸¹,
Charles Duyckaerts ^{82 83}, Alexis Brice ⁸³, Suzanne Lesage ⁸³, Georgia Xiromerisiou ⁸⁴,
Matthew J Barrett ⁸⁵, Bension S Tilley ⁸⁶, Steve Gentleman ⁸⁶, Giancarlo Logroscino ^{87 88},
Geidy E Serrano ⁸⁹, Thomas G Beach ⁸⁹, Ian G McKeith ⁹⁰, Alan J Thomas ⁹⁰, Johannes Attems ⁹⁰,
Christopher M Morris ⁹⁰, Laura Palmer ⁹¹, Seth Love ⁹², Claire Troakes ⁹³, Safa Al-Sarraj ⁹⁴,
Angela K Hodges ⁹³, Dag Aarsland ^{93 95}, Gregory Klein ⁹⁶, Scott M Kaiser ⁹⁷, Randy Woltjer ⁹⁸,
Pau Pastor ¹¹, Lynn M Bekris ⁹⁹, James B Leverenz ¹⁰⁰, Lilah M Besser ¹⁰¹, Amanda Kuzma ¹⁰²,
Alan E Renton ¹⁰³, Alison Goate ¹⁰⁴, David A Bennett ⁹⁶, Clemens R Scherzer ¹⁰⁵, Huw R Morris ¹⁰⁶,
Raffaele Ferrari ⁴, Diego Albani ¹⁰⁷, Stuart Pickering-Brown ¹⁰⁸, Kelley Faber ¹⁰⁹,
Walter A Kukull ¹¹⁰, Estrella Morenas-Rodriguez ^{111 112 113}, Alberto Lleó ^{112 113},

Juan Fortea^{112 113}, Daniel Alcolea^{112 113}, Jordi Clarimon^{112 113}, Mike A Nalls^{114 115 116},
Luigi Ferrucci¹¹⁷, Susan M Resnick¹¹⁸, Toshiko Tanaka¹¹⁷, Tatiana M Foroud¹⁰⁹,
Neill R Graff-Radford¹¹⁹, Zbigniew K Wszolek¹¹⁹, Tanis Ferman¹²⁰, Bradley F Boeve¹²¹,
John A Hardy^{4 12 122 123 124}, Eric J Topol¹²⁵, Ali Torkamani¹²⁵, Andrew B Singleton^{3 116},
Mina Ryten^{5 6}, Dennis W Dickson⁷, Adriano Chiò^{40 41 126}, Owen A Ross^{7 127},
J Raphael Gibbs¹⁰, Clifton L Dalgard^{128 129}, Bryan J Traynor^{1 12 64}, Sonja W Scholz^{130 131}

Collaborators, Affiliations

PMID: 33589841 PMCID: PMC7946812 (available on 2021-08-15)

DOI: [10.1038/s41588-021-00785-3](https://doi.org/10.1038/s41588-021-00785-3)

Abstract

The genetic basis of Lewy body dementia (LBD) is not well understood. Here, we performed whole-genome sequencing in large cohorts of LBD cases and neurologically healthy controls to study the genetic architecture of this understudied form of dementia, and to generate a resource for the scientific community. Genome-wide association analysis identified five independent risk loci, whereas genome-wide gene-aggregation tests implicated mutations in the gene GBA. Genetic risk scores demonstrate that LBD shares risk profiles and pathways with Alzheimer's disease and Parkinson's disease, providing a deeper molecular understanding of the complex genetic architecture of this age-related neurodegenerative condition.

Related information

[MedGen](#)

LinkOut – more resources

Full Text Sources

[Nature Publishing Group](#)

Other Literature Sources

[scite Smart Citations](#)

Medical

[Genetic Alliance](#)

[MedlinePlus Health Information](#)

Miscellaneous

[NCI CPTAC Assay Portal](#)

9. Romano R, Bertolino A, Gigante A, Martino D, Livrea P, Defazio G. Impaired cognitive functions in adult-onset primary cranial cervical dystonia. *Park Relat Disord* 2014;20:162–165.
10. Prell T, Peschel T, Köhler B, et al. Structural brain abnormalities in cervical dystonia. *BMC Neurosci* 2013;14:123. <https://doi.org/10.1186/1471-2202-14-123>
11. Filip P, Gallea C, Lehericy S, et al. Disruption in cerebellar and basal ganglia networks during a visuospatial task in cervical dystonia. *Mov Disord* 2017;32:757–768.
12. Munoz DP, Everling S. Look away: the anti-saccade task and the voluntary control of eye movement. *Nat Rev Neurosci* 2004;5: 218–228.
13. Terao Y, Fukuda H, Hikosaka O. What do eye movements tell us about patients with neurological disorders? -an introduction to saccade recording in the clinical setting. *Proc Japan Acad Ser B Phys Biol Sci* 2017;93:772–801.
14. Beck RB, Kneafsey SL, Narasimham S, et al. Reduced frequency of ipsilateral express saccades in cervical dystonia: probing the nigro-tectal pathway. *Tremor Other Hyperkin Mov (NY)* 2018;8:4–8.
15. Stell R, Bronstein AM, Gresty M, Buckwell D, Marsden CD. Saccadic function in spasmodic torticollis. *J Neurol Neurosurg Psychiatry* 1990;53:496–501.
16. Brodoehl S, Wagner F, Prell T, Klingner C, Witte OW, Günther A. Cause or effect: altered brain and network activity in cervical dystonia is partially normalized by botulinum toxin treatment. *NeuroImage Clin* 2019;22:101792. <https://doi.org/10.1016/j.nicl.2019.101792>
17. Boyce MJ, Canning CG, Mahant N, Morris J, Latimer J, Fung VSC. The Toronto Western spasmodic torticollis rating scale: reliability in neurologists and physiotherapists. *Park Relat Disord* 2012;18: 635–637.
18. Tsui JKC, Jon Stoessl A, Eisen A, Calne S, Calne DB. Double-blind study of Botulinum toxin in spasmodic torticollis. *Lancet* 1986;328: 245–247.
19. Jost WH, Heffer H, Stenner A, Reichel G. Rating scales for cervical dystonia: a critical evaluation of tools for outcome assessment of botulinum toxin therapy. *J Neural Transm* 2013;120:487–496.
20. Hallett PE. Primary and secondary saccades to goals defined by instructions. *Vision Res* 1978;18:1279–1296.
21. Antoniadis C, Ettinger U, Gaymard B, et al. An internationally standardised antisaccade protocol. *Vision Res* 2013;84:1–5.
22. Fischer B, Ramsperger E. Human express saccades: extremely short reaction times of goal directed eye movements. *Exp Brain Res* 1984; 57:191–195.
23. Smyrnis N, Karantinos T, Malogiannis I, et al. Larger variability of saccadic reaction times in schizophrenia patients. *Psychiatry Res* 2009;168:129–136.
24. IBM Corp. Released 2016. IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp.
25. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand* 1983;67:361–370.
26. Hallett M. Neurophysiology of dystonia: the role of inhibition. *Neurobiol Dis* 2011;42:177–184.
27. Coe BC, Munoz DP. Mechanisms of saccade suppression revealed in the anti-saccade task. *Philos Trans R Soc B Biol Sci* 2017;372 (1718):20160192. <https://doi.org/10.1098/rstb.2016.0192>
28. Xu KZ, Anderson BA, Emeric EE, et al. Neural basis of cognitive control over movement inhibition: human fMRI and primate electrophysiology evidence. *Neuron* 2017;96:1447–1458.e6.
29. Stinear CM, Byblow WD. Impaired inhibition of a pre-planned response in focal hand dystonia. *Exp Brain Res* 2004;158:207–212.
30. Everling S, Dorris MC, Klein RM, Munoz DP. Role of primate superior colliculus in preparation and execution of anti-saccades and pro-saccades. *J Neurosci* 1999;19:2740–2754.
31. Everling S, Munoz DP. Neuronal correlates for preparatory set associated with pro-saccades and anti-saccades in the primate frontal eye field. *J Neurosci* 2000;20:387–400.
32. Mahajan A, Zillgitt A, Alshamma A, et al. Cervical dystonia and executive function: a pilot magnetoencephalography study. *Brain Sci* 2018;8(9):159. <https://doi.org/10.3390/brainsci8090159>
33. Ospina-García N, Escobar-Barrios M, Rodríguez-Violante M, Benítez-Valenzuela J, Cervantes-Arriaga A. Neuropsychiatric profile of patients with craniocervical dystonia: a case-control study. *Clin Neurol Neurosurg* 2020;193:105794. <https://doi.org/10.1016/j.clineuro.2020.105794>
34. Avanzino L, Tinazzi M, Ionta S, Fiorio M. Sensory-motor integration in focal dystonia. *Neuropsychologia* 2015;79:288–300.
35. Desrochers P, Brunfeldt A, Sidiropoulos C, Kagerer F. Sensorimotor control in dystonia. *Brain Sci* 2019;9:1–18.
36. Zee DS, Lasker AG. Antisaccades: probing cognitive flexibility with eye movements. *Neurology* 2004;63:1554.
37. Barbosa P, Kaski D, Castro P, Lees AJ, Warner TT, Djamshidian A. Saccadic direction errors are associated with impulsive compulsive behaviours in Parkinson's disease patients. *J Parkinsons Dis* 2019;9: 625–630.
38. Reuter B, Rakusan L, Kathmanna N. Poor antisaccade performance in schizophrenia: an inhibition deficit? *Psychiatry Res* 2005; 135:1–10.
39. Kahana Levy N, Lavidor M, Prosaccade VE. Antisaccade paradigms in persons with Alzheimer's disease: a meta-analytic review. *Neuropsychol Rev* 2018;28:16–31.

Supporting Data

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

The Parkinson's Disease DNA Variant Browser

Jonggeol J. Kim, BSc,^{1,2} Mary B. Makarios, BSc,¹ Sara Bandres-Ciga, PhD,¹ Jesse Raphael Gibbs, PhD,³ Jinhui Ding, MSc,³ Dena G. Hernandez, PhD,¹ Janet Brooks, BSc,¹ Francis P. Grenn, BSc,¹ Hirotaka Iwaki, MD, PhD,^{1,4} Andrew B. Singleton, PhD,¹ Mike A. Nalls, PhD,^{1,4} Cornelis Blauwendraat, PhD, and^{1*} International Parkinson's Disease Genomics Consortium (IPDGC)

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

*Correspondence to: Dr. Cornelis Blauwendraat, Laboratory of Neurogenetics, NIA, NIH, Building 35, 35 Convent Drive, Bethesda, MD 20892; E-mail: cornelis.blauwendraat@nih.gov

Relevant conflicts of interest/financial disclosures: Nothing to report.

Received: 12 August 2020; **Revised:** 15 December 2020; **Accepted:** 18 December 2020

Published online 26 January 2021 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/mds.28488

¹Molecular Genetics Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, Maryland, USA ²Preventive Neurology Unit, Wolfson Institute of Preventive Medicine, Queen Mary University of London, London, UK ³Computational Biology Core, Laboratory of Neurogenetics, National Institute on Aging, Bethesda, Maryland, USA ⁴Data Tecnica International, Glen Echo, Maryland, USA

ABSTRACT: Background: Parkinson's disease (PD) is a genetically complex neurodegenerative disease with ~20 genes known to contain mutations that cause PD or atypical parkinsonism. Large-scale next-generation sequencing projects have revolutionized genomics research. Applying these data to PD, many genes have been reported to contain putative disease-causing mutations. In most instances, however, the results remain quite limited and rather preliminary. Our aim was to assist researchers on their search for PD-risk genes and variant candidates with an easily accessible and open summary-level genomic data browser for the PD research community.

Methods: Sequencing and imputed genotype data were obtained from multiple sources and harmonized and aggregated.

Results: In total we included a total of 102,127 participants, including 28,453 PD cases, 1650 proxy cases, and 72,024 controls.

Conclusions: We present here the Parkinson's Disease Sequencing Browser: a Shiny-based web application that presents comprehensive summary-level frequency data from multiple large-scale genotyping and sequencing projects <https://pdgenetics.shinyapps.io/VariantBrowser/>. Published © 2021 This article is a U.S. Government work and is in the public domain in the USA. *Movement Disorders* published by Wiley Periodicals LLC on behalf of International Parkinson and Movement Disorder Society.

Key Words: Parkinson's disease; sequencing; data browser; genetics

Parkinson's disease (PD) is a neurodegenerative disease hallmarked by dopaminergic neuron degradation and Lewy-body inclusions in the brain. The exact molecular mechanisms underlying PD remain largely unknown, but the disease is influenced by age, environmental, and complex genetic factors. Putative deleterious and highly functional variants in more than 20 genes and 90 common genetic risk variants have been associated with PD or atypical parkinsonism. However, the population risk of known mutations and risk loci only represents a fraction of the known detectable heritable component of disease, suggesting that additional genetic influence is yet to be identified.^{1,2} Most genes associated with PD have been discovered

through linkage mapping studies in large family studies, such as *SNCA*³ and *LRRK2*.⁴⁻⁶ Some studies contain large sequencing cohort validation analyses such as the one nominating *VPS13C*.⁷ The majority of the recent studies that nominate potential PD genes lack replication of results. Current research aggregation resources such as ClinVar are useful for searching known pathogenic variants, but the information presented often misses the context behind the clinical interpretation and lacks large case-control frequency information. Although other resources such as MDSGene (<https://www.mdsgene.org/>) provide in-depth genotype-phenotype information, however, they lack large study case-control frequencies.⁸

Next-generation sequencing has produced petabytes of genomic data and has transformed genomic medicine. However, databases housing these data, such as gnomAD⁹ and BRAVO variant browser,¹⁰ do not contain disease-specific data (yet), and there is a need for accessible resources that specifically include allele frequencies per disease group. Here, we aggregated multiple genomic data sets based on PD cases and controls and created an exonic summary data user-friendly browser, <https://pdgenetics.shinyapps.io/VariantBrowser/>.

Methods

Data Aggregation

We collected sequencing data from multiple different sources (Table 1). The PD Genome Project includes publicly available whole-genome sequencing data from AMP-PD (<https://amp-pd.org/>) and other sources. The International Parkinson's Disease Genomics Consortium (IPDGC) cohort from Parkinson's Disease Genetics Sequencing Consortium (PDGSC) data was downloaded in November 2019 and was processed using a previously described pipeline, <https://github.com/ipdgc/pdpsc>. The IPDGC resequencing project is a resequencing data set that includes a large number of monogenic genes (*ATP13A2*, *FBXO7*, *GBA*, *LRRK2*, *MAPT*, *PARK7* [DJ-1], *PINK1*, *PLA2G6*, *PRKN*, *SNCA*, and *VPS35*) and genome-wide association study (GWAS) loci regions from a previous PD GWAS.¹¹ The IPDGC genotype data were processed using a previously described quality control pipeline that has been previously described here, <https://github.com/neurogenetics/GWAS-pipeline>.^{1,12} It was imputed using the Haplotype Reference Consortium Panel and filtered with the estimated *r*² (RSQ) threshold of 0.8. UK Biobank (UKB) exome data (field 23160, "Population-level FE variants, PLINK format") were downloaded in May 2019.¹³ The PD status of the UKB participants was based on UKB field number 42033, "Source of Parkinson's disease report," which determined the PD status on 3 criteria: self-report, hospital admission, and death registries. UKB proxy cases were defined

TABLE 1. Description of the data cohorts

Study	Data Type	Cases (n)	Controls (n)	Total (n)
PD Genome Project ^a	WGS	2745	4071	6816
IPDGC Exomes	WES	2110	2978	5088
IPDGC Resequencing project	Resequencing data	3073	2136	5209
IPDGC GWAS Cohort	Imputed array data	21,412	23,894	45,306
UK Biobank	WES	114	38,263	40,027
UK Biobank (proxy cases)	WES	1650	NA	NA
TOTAL (excluding proxy cases)		29,454	71,342	100,796

WGS, whole-genome sequencing; WES, whole-exome sequencing.

^aThis includes AMP-PD release 1 genome data (<https://amp-pd.org/>).

as participants with no PD but with a parent with PD based on UKB field numbers 20107 and 20110, “Illnesses of father” and “Illnesses of mother.” Additional quality control was done to remove participants without case-control status and mean depth of less than 20. Note that the vast majority of data are from European ancestry. Data were trimmed to exome calling regions identical to those used in gnomAD, specifically bait-covered regions plus 50 bp upstream and downstream. Before merging, all hg38 data were mapped to hg19 using CrossMap v0.4.0,¹⁴ and each data set was filtered for relatedness by excluding individuals with PIHAT values > 0.125. After merging, duplicate samples were removed based on either sample ID or PIHAT values > 0.8 using PLINK v1.9¹⁵ (Fig. S1). The data were merged, and allele count and frequency were generated using PLINK. Merged data were annotated using ANNOVAR¹⁶ (Fig. 1).

Browser Design

The IPDGC Sequencing Browser was designed using the Shiny library under R version 3.6.1. All data present in the browser are nonidentifiable aggregate summary-level data. The design was inspired by the gnomAD and BRAVO variant browsers, featuring gene-level information panel and separate variant-level windows. However, this browser increases the information density presented in a single page format with collapsible and information panels to search results facilitating the user visualization and interpretation. It also contains an integrated tutorial function to guide new users. The browser is an open-source project, and the code is available on our GitHub platform, https://github.com/kimjonggeolj/ipdgc_exome_browser.

Results

After quality control, we included a total of 6,126,909 variants from 102,446 participants, specifically 29,454 PD cases, 1650 proxy cases from UKB, and 71,342 controls. Of the 3,581,869 exonic variants (~58%), 2,144,315 (~60%) were nonsynonymous variants, and 1,078,658 (~30%) were synonymous

variants (Table S1). As a positive control, we assessed the allele frequency of *LRRK2* p.G2019S (rs34637584). This variant is one of the most common PD genetic factors associated with both familial and sporadic forms of disease.¹⁷ Our browser shows the minor allele frequency of this variant at 0.007716 for cases, 0.0003201 for controls, and 0.001212 for proxy cases. Zygosity distributions show a similar pattern. Of 23,068 cases and 64,036 controls, there are 354 heterozygous cases and 41 heterozygous controls. There is 1 homozygous carrier case, whereas there are no controls with the same zygosity pattern. An association test (chi square allelic test) without adjusting by any covariate but excluding proxy cases showed an odds ratio of 24.28 (95% CI, 17.57–33.6) and a $P = 1.76 \times 10^{-179}$, which is in line with previous reports for this variant. Another example is the *SNCA* p.A53T (rs104893877) variant, which was the first pathogenic *SNCA* variant described resulting in autosomal-dominant PD.³ The browser shows 2 cases carrying this variant and no controls. The browser can also be used for autosomal-recessive disorders, for example, using *PRKN* (PARK2) p.R275W (rs34424986) as an example, one of the most common *PRKN* pathogenic variants.¹⁸ Six cases were homozygous for this variant, and no controls were identified in a homozygous state. The result showed confidence that the data set could be used to identify or provide evidence for a potential PD causal variant.

Discussion

Here we present the Parkinson's Disease DNA Variant browser, a public platform for the scientific community that allows rapid querying of specific genes and variants in several large case-control cohorts. Provided with a gene name or gene boundaries, the browser will present the user with summarized information on the variants found within the gene, such as the distribution of variants categorized by their functional consequences. Given a specific variant, the browser will present the user with annotated information on the variant including allele

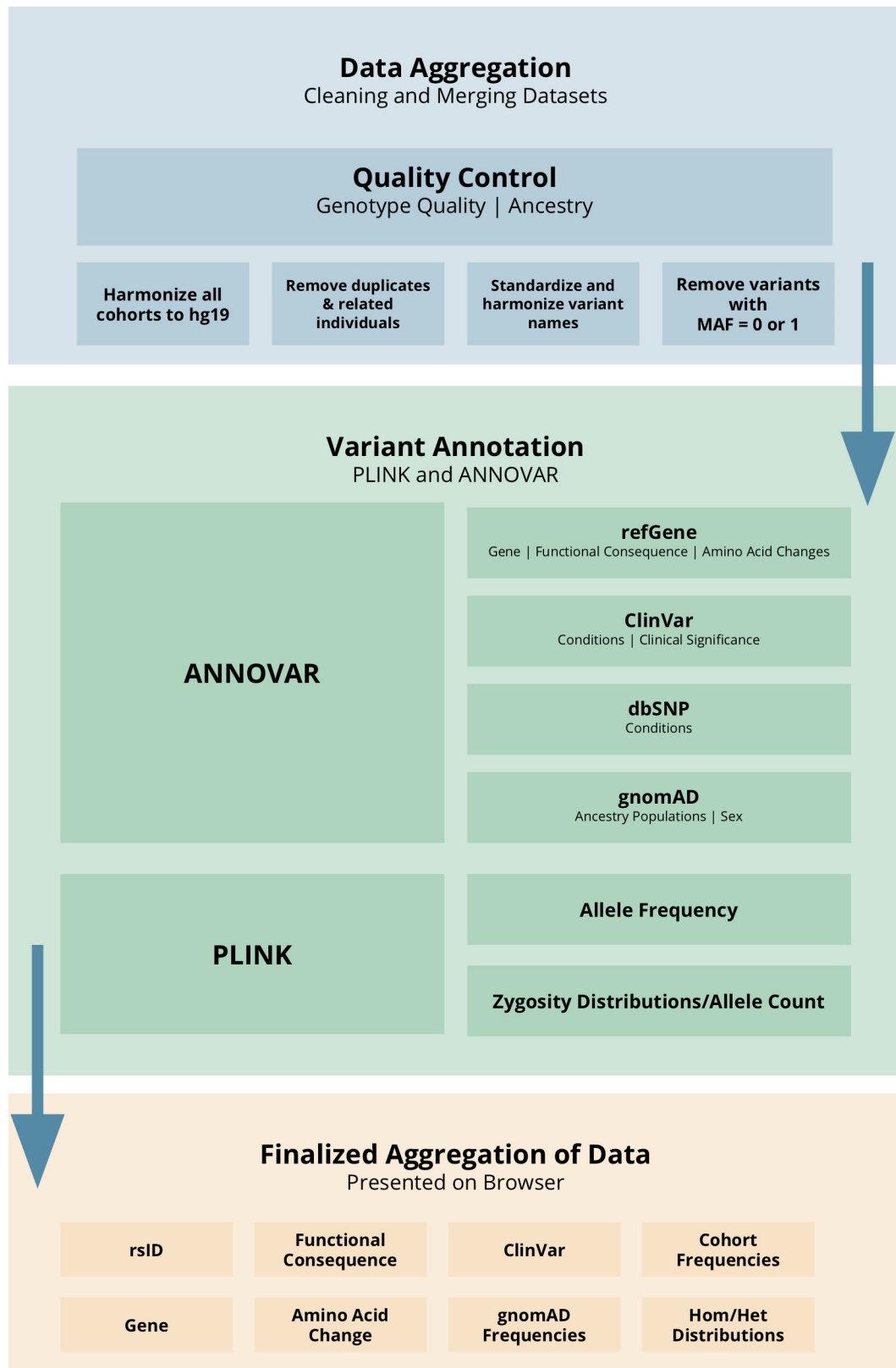


FIG. 1. Parkinson's Disease Sequencing Browser processing pipeline flowchart.

frequency, ClinVar information, and functional consequence. These functionalities can be used, for example, when assessing the frequency of a variant of interest

identified in a PD case or within a family. As shown in the results of the *LRRK2* p.G2019S, *PRKN* p.R275W, and *SNCA* p.A53T examples, allele frequency and

zygosity distribution can give a researcher an idea of whether the reported risk variant may be enriched in PD cases.

Although we performed extensive quality control to ensure high-quality information was used, the data presented in the browser have inherent limitations. The data sets merged different sequencing technologies including whole-genome sequencing, whole-exome sequencing, resequencing, and imputed array-based genotyping and were aligned using different genome builds such as hg19 and hg38. This leaves gaps from low-imputation regions and cross-mapping failures, although the cross-mapping introduced less than 0.1% reduction in the total number of variants (Fig. S1). Although no sequencing or genotyping method can guarantee perfect data, imputed data because of its nature may add additional uncertainties regarding its results. Our quality control filters reduce this uncertainty, but nevertheless users should always consult the specific study-level breakdown of the variant frequency and count. The presented data only include autosomal data and do not contain any variants from sex chromosomes. Furthermore, the data presented only include exonic regions and their immediate flank; thus, it cannot provide information on the majority of the noncoding variants. Although this database contains multiple whole-exome and genome sequencing data, it may still be difficult to identify significant allele frequency and count differences between very rare variants, especially those of recessive inheritance. Researchers should use the annotated information such as ClinVar significance to critically assess any candidate variants. In addition, we only included a very limited amount of phenotypic data with case-control status, which creates potential bias for age-related penetrance. In future versions, we aim to include age of onset and other phenotype data if available. Of note, the majority of the data included are from European ancestry. We hope in future versions to increase the diversity of the data. Last, some genes, regions of interest, and structural variation are very complicated to genotype and sequence (including the *GBA* gene because of the high similarities with the pseudogene), and therefore interpretation of these complex regions should be done with caution. Future larger-scale and targeted studies will hopefully resolve the issue with complex genomic regions.

In summary, we present here an online resource developed for the PD research community to quickly retrieve annotated genomic information on genes and variants in a user-friendly manner, without any required data science or coding experience. Users can access the browser to get information on reported PD risk factors or supplement their own research with data from a large-scale data set. We envisage this browser to be the first step toward easily sharing genomic information that will be continuously updated as new data become available. ■

Acknowledgments: We thank all the subjects who donated their time and biological samples to be part of this study. The authors also thank the Genome Aggregation Database (gnomAD) and the groups that provided exome and genome variant data to this resource. A full list of contributing groups can be found at <https://gnomad.broadinstitute.org/about>. Data used in the preparation of this article were obtained from the AMP PD Knowledge Platform. For up-to-date information on the study, <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, Sanofi, and Verily. Clinical data 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 NeuroDiscovery Biomarker Study (HBS) is a collaboration of HBS investigators. A full list of HBS investigators can be 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, including AbbVie, Allergan, Amathus Therapeutics, Avid Radiopharmaceuticals, Biogen, BioLegend, Bristol-Myers Squibb, Calico, Celgene, Denali, GE Healthcare, Genentech, GlaxoSmithKline, Golub Capital, Handling's Therapeutics, Insitro, Janssen Pharmaceutica, Eli Lilly and Company, Lundbeck, Merck, Meso Scale Discovery, Neurocrine Biosciences, Pfizer, Piramal Imaging, Prevail Therapeutics, Roche, Sanofi Genzyme, Servier, Takeda Pharmaceutical Company, Teva, UCB, Verily, and Voyager Therapeutics. 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 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. This work was supported by the Intramural Research Programs of the National Institute on Aging (NIA). This research has been conducted using the UK Biobank Resource under application number 33601. This study used the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health (<http://hpc.nih.gov>). The authors also thank all members of the International Parkinson Disease Genomics Consortium (IPDGC). For a complete overview of members, acknowledgments, and funding, see <http://pdgenetics.org/partners>.

Full IPDGC Author List

United Kingdom: Alastair J. Noyce (Preventive Neurology Unit, Wolfson Institute of Preventive Medicine, QMUL, London, UK, and Department of Molecular Neuroscience, UCL, London, UK), Rauan Kaiyrzhanov (Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK), Ben Middlehurst (Institute of Translational Medicine, University of Liverpool, Liverpool, UK), Demis A. Kia (UCL Genetics Institute; and Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK), Manuela Tan (Department of Clinical Neuroscience, University College London, London, UK), Henry Houlden (Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK), Catherine Storm (Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, London, UK), Huw R. Morris (Department of Clinical Neuroscience, University College London, London, UK), Helene Plun-Favreau (Department of Molecular

Neuroscience, UCL Institute of Neurology, London, UK), Peter Holmans (Biostatistics & Bioinformatics Unit, Institute of Psychological Medicine and Clinical Neuroscience, MRC Centre for Neuropsychiatric Genetics & Genomics, Cardiff, UK), John Hardy (Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK), Daniah Tratzuni (Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK; Department of Genetics, King Faisal Specialist Hospital and Research Centre, Riyadh, Saudi Arabia), John Quinn (Institute of Translational Medicine, University of Liverpool, Liverpool, UK), Vivien Bubb (Institute of Translational Medicine, University of Liverpool, Liverpool, UK), Kin Y Mok (Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK), Kerri J. Kinghorn (Institute of Healthy Ageing, Research Department of Genetics, Evolution and Environment, University College London, London, UK), Nicholas W. Wood (UCL Genetics Institute; and Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK), Patrick Lewis (University of Reading, Reading, UK), Sebastian R. Schreglmann (Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK), Ruth Lovering (University College London, London, UK), Lea R'Bibo (Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK), Claudia Manzoni (University of Reading, Reading, UK), Mie Rizig (Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK), Mina Ryten (Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK), Sebastian Guelfi (Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK), Valentina Escott-Price (MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University School of Medicine, Cardiff, UK), Viorica Chelban (Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK), Thomas Foltynie (UCL Institute of Neurology, London, UK), Nigel Williams (MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff, UK), Karen E. Morrison (Faculty of Medicine, University of Southampton, UK), Carl Clarke (University of Birmingham, Birmingham, UK, and Sandwell and West Birmingham Hospitals NHS Trust, Birmingham, UK), Kirsten Harvey (UCL School of Pharmacy, UK), Benjamin M. Jacobs (Preventive Neurology Unit, Wolfson Institute of Preventive Medicine, QMUL, London, UK).

France: Alexis Brice (Institut du Cerveau et de la Moelle épinière, ICM, Inserm U 1127, CNRS, UMR 7225, Sorbonne Universités, UPMC University Paris, UMR S 1127, AP-HP, Pitié-Salpêtrière Hospital, Paris, France), Fabrice Danjou (Institut du Cerveau et de la Moelle épinière, ICM, Inserm U 1127, CNRS, UMR 7225, Sorbonne Universités, UPMC University Paris, UMR S 1127, AP-HP, Pitié-Salpêtrière Hospital, Paris, France), Suzanne Lesage (Institut du Cerveau et de la Moelle épinière, ICM, Inserm U 1127, CNRS, UMR 7225,

Sorbonne Universités, UPMC University Paris 06, UMR S 1127, AP-HP, Pitié-Salpêtrière Hospital, Paris, France), Jean-Christophe Corvol (Institut du Cerveau et de la Moelle épinière, ICM, Inserm U 1127, CNRS, UMR 7225, Sorbonne Universités, UPMC University Paris, UMR S 1127, Centre d'Investigation Clinique Pitié Neuropsychiatrie CIC-1422, AP-HP, Pitié-Salpêtrière Hospital, Paris, France), Maria Martinez (INSERM UMR 1220; and Paul Sabatier University, Toulouse, France),

Germany: Claudia Schulte (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tübingen, and DZNE, German Center for Neurodegenerative Diseases, Tübingen, Germany), Kathrin Brockmann (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tübingen, and DZNE, German Center for Neurodegenerative Diseases, Tübingen, Germany), Javier Simón-Sánchez (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tübingen, and DZNE, German Center for Neurodegenerative Diseases, Tübingen, Germany), Peter Heutink (DZNE, German Center for Neurodegenerative Diseases and Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany), Patrizia Rizzu (DZNE, German Center for Neurodegenerative Diseases), Manu Sharma (Centre for Genetic Epidemiology, Institute for Clinical Epidemiology and Applied Biometry, University of Tübingen, Germany), Thomas Gasser (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, and DZNE, German Center for Neurodegenerative Diseases, Tübingen, Germany), Susanne A. Schneider (Department of Neurology, Ludwig-Maximilians-University Munich, München, Germany).

United States: Mark R. Cookson (Laboratory of Neurogenetics, National Institute on Aging, Bethesda, MD), Sara Bandres-Ciga (Laboratory of Neurogenetics, National Institute on Aging, Bethesda, MD), Cornelis Blauwendraat (Laboratory of Neurogenetics, National Institute on Aging, Bethesda, MD), David W. Craig (Department of Translational Genomics, Keck School of Medicine, University of Southern California, Los Angeles, CA), Kimberley Billingsley (Laboratory of Neurogenetics, National Institute on Aging, Bethesda, MD), Mary B. Makarios (Laboratory of Neurogenetics, National Institute on Aging, Bethesda, MD), Derek P. Narendra (Inherited Movement Disorders Unit, National Institute of Neurological Disorders and Stroke, Bethesda, MD), Faraz Faghri (Laboratory of Neurogenetics, National Institute on Aging, Bethesda, MD; Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL), Jesse Raphael Gibbs (Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD), Dena G. Hernandez (Laboratory of

Neurogenetics, National Institute on Aging, Bethesda, MD), Kendall Van Keuren-Jensen (Neurogenomics Division, TGen, Phoenix, AZ), Joshua M. Shulman (Departments of Neurology, Neuroscience, and Molecular & Human Genetics, Baylor College of Medicine, Houston, TX; Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX), Hirotaka Iwaki (Laboratory of Neurogenetics, National Institute on Aging, Bethesda, MD), Hampton L. Leonard (Laboratory of Neurogenetics, National Institute on Aging, Bethesda, MD), Mike A. Nalls (Laboratory of Neurogenetics, National Institute on Aging, Bethesda, MD; CEO/Consultant Data Tecnica International, Glen Echo, MD), Laurie Robak (Baylor College of Medicine, Houston, TX), Jose Bras (Center for Neurodegenerative Science, Van Andel Research Institute, Grand Rapids, MI), Rita Guerreiro (Center for Neurodegenerative Science, Van Andel Research Institute, Grand Rapids, MI), Steven Lubbe (Ken and Ruth Davee Department of Neurology and Simpson Querrey Center for Neurogenetics, Northwestern University Feinberg School of Medicine, Chicago, IL), Steven Finkbeiner (Departments of Neurology and Physiology, University of California, San Francisco; Gladstone Institute of Neurological Disease; Taube/Koret Center for Neurodegenerative Disease Research, San Francisco, CA), Niccolo E. Mencacci (Northwestern University Feinberg School of Medicine, Chicago, IL), Codrin Lungu (National Institutes of Health Division of Clinical Research, NINDS, National Institutes of Health, Bethesda, MD), Andrew B. Singleton (Laboratory of Neurogenetics, National Institute on Aging, Bethesda, MD), Sonja W. Scholz (Neurodegenerative Diseases Research Unit, National Institute of Neurological Disorders and Stroke, Bethesda, MD), Xylena Reed (Laboratory of Neurogenetics, National Institute on Aging, Bethesda, MD), Roy N. Alcalay (Department of Neurology, College of Physicians and Surgeons, Columbia University Medical Center, New York, NY; Taub Institute for Research on Alzheimer's Disease and the Aging Brain, College of Physicians and Surgeons, Columbia University Medical Center, New York, NY), Zbigniew K. Wszolek (Department of Neurology, Mayo Clinic Jacksonville, FL), Ryan J. Uitti (Department of Neurology, Mayo Clinic Jacksonville, FL), Owen A. Ross (Departments of Neuroscience & Clinical Genomics, Mayo Clinic Jacksonville, FL), Francis P. Grenn (Laboratory of Neurogenetics, National Institute on Aging, Bethesda, MD).

Canada: Ziv Gan-Or (Montreal Neurological Institute and Hospital, Department of Neurology & Neurosurgery, Department of Human Genetics, McGill University, Montréal, QC, Canada), Guy A. Rouleau (Montreal Neurological Institute and Hospital, Department of Neurology & Neurosurgery, Department of Human Genetics, McGill University, Montréal, QC, Canada), Lynne Krohn (Montreal

Neurological Institute and Hospital, Department of Neurology & Neurosurgery, Department of Human Genetics, McGill University, Montréal, QC, Canada), Kheireddin Mufti (Montreal Neurological Institute and Hospital, Department of Neurology & Neurosurgery, Department of Human Genetics, McGill University, Montréal, QC, Canada),

The Netherlands: Jacobus J. van Hilten (Department of Neurology, Leiden University Medical Center, Leiden, The Netherlands), Johan Marinus (Department of Neurology, Leiden University Medical Center, Leiden, The Netherlands).

Spain: Astrid D. Adarmes-Gómez (Instituto de Biomedicina de Sevilla [IBiS], Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Seville), Miquel Aguilar (Fundació Docència i Recerca Mútua de Terrassa and Movement Disorders Unit, Department of Neurology, University Hospital Mutua de Terrassa, Terrassa, Barcelona), Ignacio Alvarez (Fundació Docència i Recerca Mútua de Terrassa and Movement Disorders Unit, Department of Neurology, University Hospital Mutua de Terrassa, Terrassa, Barcelona), Victoria Alvarez (Hospital Universitario Central de Asturias, Oviedo), Francisco Javier Barrero (Hospital Universitario San Cecilio de Granada, Universidad de Granada), Jesús Alberto Bergareche Yaza (Instituto de Investigación Sanitaria Biodonostia, San Sebastián), Inmaculada Bernal-Bernal (Instituto de Biomedicina de Sevilla [IBiS], Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Seville), Marta Blazquez (Hospital Universitario Central de Asturias, Oviedo), Marta Bonilla-Toribio (Instituto de Biomedicina de Sevilla [IBiS], Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Seville), Juan A. Botía (Universidad de Murcia, Murcia), María Teresa Boungiorno (Fundació Docència i Recerca Mútua de Terrassa and Movement Disorders Unit, Department of Neurology, University Hospital Mutua de Terrassa, Terrassa, Barcelona), Dolores Buiza-Rueda (Instituto de Biomedicina de Sevilla [IBiS], Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Seville), Ana Cámara (Hospital Clinic de Barcelona), Fátima Carrillo (Instituto de Biomedicina de Sevilla [IBiS], Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Seville), Mario Carrión-Claro (Instituto de Biomedicina de Sevilla [IBiS], Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Seville), Debora Cerdan (Hospital General de Segovia, Segovia), Jordi Clarimón (Memory Unit, Department of Neurology, IIB Sant Pau, Hospital de la Santa Creu i Sant Pau, Universitat Autònoma de Barcelona and Centro de Investigación Biomédica en Red en Enfermedades Neurodegenerativas [CIBERNED], Madrid), Yaroslau Compta (Hospital Clinic de Barcelona), Monica Diez-Fairen (Fundació Docència i Recerca Mútua de Terrassa and Movement Disorders Unit, Department of Neurology, University Hospital Mutua de Terrassa, Terrassa, Barcelona), Oriol Dols-Icardo

(Memory Unit, Department of Neurology, IIB Sant Pau, Hospital de la Santa Creu i Sant Pau, Universitat Autònoma de Barcelona, Barcelona, and Centro de Investigación Biomédica en Red en Enfermedades Neurodegenerativas [CIBERNED], Madrid), Jacinto Duarte (Hospital General de Segovia, Segovia), Raquel Duran (Centro de Investigación Biomedica, Universidad de Granada, Granada), Francisco Escamilla-Sevilla (Hospital Universitario Virgen de las Nieves, Instituto de Investigación Biosanitaria de Granada, Granada), Mario Ezquerro (Hospital Clinic de Barcelona), Cici Feliz (Departamento de Neurología, Instituto de Investigación Sanitaria Fundación Jiménez Díaz, Madrid, Spain), Manel Fernández (Hospital Clinic de Barcelona), Rubén Fernández-Santiago (Hospital Clinic de Barcelona), Ciara Garcia (Hospital Universitario Central de Asturias, Oviedo), Pedro García-Ruiz (Instituto de Investigación Sanitaria Fundación Jiménez Díaz, Madrid), Pilar Gómez-Garre (Instituto de Biomedicina de Sevilla [IBiS], Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Seville), Maria Jose Gomez Heredia (Hospital Universitario Virgen de la Victoria, Malaga), Isabel Gonzalez-Aramburu (Hospital Universitario Marqués de Valdecilla-IDIVAL, Santander), Ana Gorostidi Pagola (Instituto de Investigación Sanitaria Biodonostia, San Sebastián), Janet Hoenicka (Institut de Recerca Sant Joan de Déu, Barcelona), Jon Infante (Hospital Universitario Marqués de Valdecilla-IDIVAL and University of Cantabria, Santander, and Centro de Investigación Biomédica en Red en Enfermedades Neurodegenerativas [CIBERNED]), Silvia Jesús (Instituto de Biomedicina de Sevilla [IBiS], Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Seville), Adriano Jimenez-Escrig (Hospital Universitario Ramón y Cajal, Madrid), Jaime Kulisevsky (Movement Disorders Unit, Department of Neurology, IIB Sant Pau, Hospital de la Santa Creu i Sant Pau, Universitat Autònoma de Barcelona, Barcelona, and Centro de Investigación Biomédica en Red en Enfermedades Neurodegenerativas [CIBERNED]), Miguel A. Labrador-Espinosa (Instituto de Biomedicina de Sevilla [IBiS], Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Seville), Jose Luis Lopez-Sendon (Hospital Universitario Ramón y Cajal, Madrid), Adolfo López de Munain Arregui (Instituto de Investigación Sanitaria Biodonostia, San Sebastián), Daniel Macias (Instituto de Biomedicina de Sevilla [IBiS], Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Seville), Irene Martínez Torres (Department of Neurology, Instituto de Investigación Sanitaria La Fe, Hospital Universitario y Politécnico La Fe, Valencia), Juan Marín (Movement Disorders Unit, Department of Neurology, IIB Sant Pau, Hospital de la Santa Creu i Sant Pau, Universitat Autònoma de Barcelona, Barcelona, and Centro de Investigación Biomédica en Red en Enfermedades

Neurodegenerativas [CIBERNED]), Maria Jose Marti (Hospital Clinic Barcelona), Juan Carlos Martínez-Castrillo (Instituto Ramón y Cajal de Investigación Sanitaria, Hospital Universitario Ramón y Cajal, Madrid), Carlota Méndez-del-Barrio (Instituto de Biomedicina de Sevilla [IBiS], Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Seville), Manuel Menéndez González (Hospital Universitario Central de Asturias, Oviedo), Marina Mata (Department of Neurology, Hospital Universitario Infanta Sofía, Madrid), Adolfo Mínguez (Hospital Universitario Virgen de las Nieves, Granada, Instituto de Investigación Biosanitaria de Granada), Pablo Mir (Instituto de Biomedicina de Sevilla [IBiS], Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Seville), Elisabet Mondragon Rezola (Instituto de Investigación Sanitaria Biodonostia, San Sebastián), Esteban Muñoz (Hospital Clinic Barcelona), Javier Pagonabarraga (Movement Disorders Unit, Department of Neurology, IIB Sant Pau, Hospital de la Santa Creu i Sant Pau, Universitat Autònoma de Barcelona, Barcelona, and Centro de Investigación Biomédica en Red en Enfermedades Neurodegenerativas [CIBERNED]), Pau Pastor (Fundació Docència i Recerca Mútua de Terrassa and Movement Disorders Unit, Department of Neurology, University Hospital Mutua de Terrassa, Terrassa, Barcelona), Francisco Perez Errazquin (Hospital Universitario Virgen de la Victoria, Malaga), Teresa Perinán-Tocino (Instituto de Biomedicina de Sevilla [IBiS], Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Seville), Javier Ruiz-Martínez (Hospital Universitario Donostia, Instituto de Investigación Sanitaria Biodonostia, San Sebastián), Clara Ruz (Centro de Investigación Biomedica, Universidad de Granada, Granada), Antonio Sanchez Rodriguez (Hospital Universitario Marqués de Valdecilla-IDIVAL, Santander), María Sierra (Hospital Universitario Marqués de Valdecilla-IDIVAL, Santander), Esther Suarez-Sanmartin (Hospital Universitario Central de Asturias, Oviedo), Cesar Tabernero (Hospital General de Segovia, Segovia), Juan Pablo Tartari (Fundació Docència i Recerca Mútua de Terrassa and Movement Disorders Unit, Department of Neurology, University Hospital Mutua de Terrassa, Terrassa, Barcelona), Cristina Tejera-Parrado (Instituto de Biomedicina de Sevilla [IBiS], Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Seville), Eduard Tolosa (Hospital Clinic Barcelona), Francesc Valldeoriola (Hospital Clinic Barcelona), Laura Vargas-González (Instituto de Biomedicina de Sevilla [IBiS], Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Seville), Lydia Vela (Department of Neurology, Hospital Universitario Fundación Alcorcón, Madrid), Francisco Vives (Centro de Investigación Biomedica, Universidad de Granada, Granada).

Austria: Alexander Zimprich (Department of Neurology, Medical University of Vienna, Vienna, Austria).

Norway: Lasse Pihlstrom (Department of Neurology, Oslo University Hospital, Oslo, Norway), Mathias Toft (Department of Neurology and Institute of Clinical Medicine, Oslo University Hospital, Oslo, Norway).

Estonia: Pille Taba (Department of Neurology and Neurosurgery, University of Tartu, Tartu, Estonia).

Australia: Sulev Koks (Centre for Molecular Medicine and Innovative Therapeutics, Murdoch University, Murdoch, Perth, Western Australia; Perron Institute for Neurological and Translational Science, Nedlands, Perth, Western Australia).

Israel: Sharon Hassin-Baer (The Movement Disorders Institute, Department of Neurology and Sagol Neuroscience Center, Chaim Sheba Medical Center, Tel-Hashomer, Ramat Gan, Israel, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel).

Finland: Kari Majamaa (Institute of Clinical Medicine, Department of Neurology, University of Oulu, Oulu, Finland; Department of Neurology and Medical Research Center, Oulu University Hospital, Oulu, Finland), Ari Siitonen (Institute of Clinical Medicine, Department of Neurology, University of Oulu, Oulu, Finland; Department of Neurology and Medical Research Center, Oulu University Hospital, Oulu, Finland), Pentti Tienari (Clinical Neurosciences, Neurology, University of Helsinki, Helsinki, Finland, Helsinki University Hospital, Helsinki, Finland).

Nigeria: Njideka U. Okubadejo (University of Lagos, Lagos State, Nigeria), Oluwadamilola O. Ojo (University of Lagos, Lagos State, Nigeria).

Kazakhstan: Rauan Kaiyrzhanov (Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK), Chingiz Shashkin (Kazakh National Medical University named after Asfendiyarov, Almaty, Kazakhstan), Nazira Zharkinbekova (South Kazakhstan Medical Academy, Shymkent, Kazakhstan), Vadim Akhmetzhanov (Astana Medical University, Astana, Kazakhstan), Akbota Aitkulova (National Center for Biotechnology, Astana, Kazakhstan; Al-Farabi Kazakh National University, Almaty, Kazakhstan), Elena Zholdybayeva (National Center for Biotechnology, Astana, Kazakhstan), Zharkyn Zharmukhanov (National Center for Biotechnology, Astana, Kazakhstan), Gulnaz Kaishybayeva (Scientific and Practical Center “Institute of Neurology named after Smagul Kaishibayev,” Almaty, Kazakhstan), Altynay Karimova (Scientific and Practical Center “Institute of Neurology named after Smagul Kaishibayev,” Almaty, Kazakhstan), Talgat Khaibullin (Semey Medical University, Semey, Kazakhstan).

Ireland: Timothy L. Lynch (The Dublin Neurological Institute at the Mater Misericordiae University Hospital, Dublin, Ireland & School of Medicine and Medical Science, University College Dublin, Dublin, Ireland). ■

References

1. Nalls MA, Blauwendraat C, Vallergera 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, Nalls MA, Singleton AB. The genetic architecture of Parkinson's disease. *Lancet Neurol* 2020;19:170–178.
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. Funayama M, Hasegawa K, Kowa H, Saito M, Tsuji S, Obata F. A new locus for Parkinson's disease (PARK8) maps to chromosome 12p11.2-q13.1. *Ann Neurol* 2002;51:296–301.
5. Zimprich A, Biskup S, Leitner P, et al. Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. *Neuron* 2004;44:601–607.
6. Paisán-Ruiz C, Jain S, Evans EW, et al. Cloning of the gene containing mutations that cause PARK8-linked Parkinson's disease. *Neuron* 2004;44:595–600.
7. Lesage S, Drouet V, Majounie E, et al. Loss of VPS13C function in autosomal-recessive parkinsonism causes mitochondrial dysfunction and increases PINK1/Parkin-dependent Mitophagy. *Am J Hum Genet* 2016;98:500–513.
8. 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;33:730–741.
9. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;581:434–443.
10. University of Michigan and NHLBI. The NHLBI trans-omics for precision medicine (TOPMed) whole genome sequencing program. BRAVO. 2018.
11. Nalls MA, Pankratz N, Lill CM, et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat Genet* 2014;46:989–993.
12. Blauwendraat C, Heilbron K, Vallergera CL, Bandres-Ciga S, von Coelln R, Pihlstrom L, et al. Parkinson's disease age at onset genome-wide association study: defining heritability, genetic loci, and α -synuclein mechanisms. *Mov Disord* 2019;34:866–875.
13. Bycroft C, Freeman C, Petkova D, et al. The UKbiobank resource with deep phenotyping and genomic data. *Nature* 2018;562:203–209.
14. Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* 2014;30:1006–1007.
15. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;4:7.
16. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38:e164.
17. Correia Guedes L, Ferreira JJ, Rosa MM, Coelho M, Bonifati V, Sampaio C. Worldwide frequency of G2019S LRRK2 mutation in Parkinson's disease: a systematic review. *Parkinsonism Relat Disord* 2010;16:237–242.
18. Klein C, Djarmati A, Hedrich K, et al. PINK1, Parkin, and DJ-1 mutations in Italian patients with early-onset parkinsonism. *Eur J Hum Genet* 2005;13(9):1086–1093.

Supporting Data

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

Genetic modifiers of risk and age at onset in *GBA* associated Parkinson's disease and Lewy body dementia

 Cornelis Blauwendraat,¹ Xylena Reed,¹  Lynne Krohn,^{2,3} Karl Heilbron,⁴ Sara Bandres-Ciga,¹  Manuela Tan,⁵ J. Raphael Gibbs,¹ Dena G. Hernandez,¹ Ravindran Kumaran,¹ Rebekah Langston,¹ Luis Bonet-Ponce,¹ Roy N. Alcalay,^{6,7} Sharon Hassin-Baer,^{8,9,10,11} Lior Greenbaum,^{8,11,12}  Hirotaka Iwaki,¹ Hampton L. Leonard,¹ Francis P. Grenn,¹ Jennifer A. Ruskey,^{2,3} Marya Sabir,¹³ Sarah Ahmed,¹³ Mary B. Makarios,¹³ Lasse Pihlstrøm,¹⁴ Mathias Toft,¹⁴ Jacobus J. van Hilten,¹⁵ Johan Marinus,¹⁵ Claudia Schulte,^{16,17} Kathrin Brockmann,^{16,17} Manu Sharma,¹⁸ Ari Siitonen,^{19,20} Kari Majamaa,^{19,20} Johanna Eerola-Rautio,²¹ Pentti J. Tienari,²¹ The 23andMe Research Team,⁴ Alexander Pantelyat,²² Argye E. Hillis,²²  Ted M. Dawson,^{22,23} Liana S. Rosenthal,²³ Marilyn S. Albert,²³ Susan M. Resnick,²⁴ Luigi Ferrucci,²⁵ Christopher M. Morris,²⁶ Olga Pletnikova,²⁷ Juan Troncoso,^{23,27} Donald Grosset,²⁸ Suzanne Lesage,²⁹ Jean-Christophe Corvol,²⁹ Alexis Brice,²⁹ Alastair J. Noyce,^{5,30} Eliezer Masliah,¹ Nick Wood,⁵ John Hardy,³¹ Lisa M. Shulman,³² Joseph Jankovic,³³ Joshua M. Shulman,^{33,34,35} Peter Heutink,^{16,17} Thomas Gasser,^{16,17} Paul Cannon,⁴ Sonja W. Scholz,^{13,23} Huw Morris,⁵ Mark R. Cookson,¹ Mike A. Nalls,^{1,36}  Ziv Gan-Or,^{2,3,37,*} and Andrew B. Singleton,^{1,*} on behalf of the International Parkinson's Disease Genomics Consortium (IPDGC)

*These authors contributed equally to this work.

Parkinson's disease is a genetically complex disorder. Multiple genes have been shown to contribute to the risk of Parkinson's disease, and currently 90 independent risk variants have been identified by genome-wide association studies. Thus far, a number of genes (including *SNCA*, *LRRK2*, and *GBA*) have been shown to contain variability across a spectrum of frequency and effect, from rare, highly penetrant variants to common risk alleles with small effect sizes. Variants in *GBA*, encoding the enzyme glucocerebrosidase, are associated with Lewy body diseases such as Parkinson's disease and Lewy body dementia. These variants, which reduce or abolish enzymatic activity, confer a spectrum of disease risk, from 1.4- to >10-fold. An outstanding question in the field is what other genetic factors that influence *GBA*-associated risk for disease, and whether these overlap with known Parkinson's disease risk variants. Using multiple, large case-control datasets, totalling 217 165 individuals (22 757 Parkinson's disease cases, 13 431 Parkinson's disease proxy cases, 622 Lewy body dementia cases and 180 355 controls), we identified 1691 Parkinson's disease cases, 81 Lewy body dementia cases, 711 proxy cases and 7624 controls with a *GBA* variant (p.E326K, p.T369M or p.N370S). We performed a genome-wide association study and analysed the most recent Parkinson's disease-associated genetic risk score to detect genetic influences on *GBA* risk and age at onset. We attempted to replicate our findings in two independent datasets, including the personal genetics company 23andMe, Inc. and whole-genome sequencing data. Our analysis showed that the overall Parkinson's disease genetic risk score modifies risk for disease and decreases age at onset in carriers of *GBA* variants. Notably, this effect was consistent across all tested *GBA* risk variants. Dissecting this signal demonstrated that variants in close proximity to

SNCA and *CTSB* (encoding cathepsin B) are the most significant contributors. Risk variants in the *CTSB* locus were identified to decrease mRNA expression of *CTSB*. Additional analyses suggest a possible genetic interaction between *GBA* and *CTSB* and *GBA* p.N370S induced pluripotent cell-derived neurons were shown to have decreased cathepsin B expression compared to controls. These data provide a genetic basis for modification of *GBA*-associated Parkinson's disease risk and age at onset, although the total contribution of common genetics variants is not large. We further demonstrate that common variability at genes implicated in lysosomal function exerts the largest effect on *GBA* associated risk for disease. Further, these results have implications for selection of *GBA* carriers for therapeutic interventions.

- 1 Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA
- 2 Department of Human Genetics, McGill University, Montreal, Quebec, Canada
- 3 Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada
- 4 23andMe, Inc., Mountain View, CA, USA
- 5 Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, London, UK
- 6 Department of Neurology, College of Physicians and Surgeons, Columbia University, New York, NY, USA
- 7 Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University, New York, NY, USA
- 8 Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel
- 9 Department of Neurology, Sheba Medical Center, Tel Hashomer, Israel
- 10 Movement Disorders Institute, Sheba Medical Center, Tel Hashomer, Israel
- 11 The Joseph Sagol Neuroscience Center, Sheba Medical Center, Tel Hashomer, Israel
- 12 The Danek Gertner Institute of Human Genetics, Sheba Medical Center, Tel Hashomer, Israel
- 13 Neurodegenerative Diseases Research Unit, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA
- 14 Department of Neurology, Oslo University Hospital, Oslo, Norway
- 15 Department of Neurology, Leiden University Medical Center, Leiden, The Netherlands
- 16 Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany
- 17 German Center for Neurodegenerative Diseases (DZNE), Tuebingen, Germany
- 18 Centre for Genetic Epidemiology, Institute for Clinical Epidemiology and Applied Biometry, University of Tübingen, Germany
- 19 Institute of Clinical Medicine, Department of Neurology, University of Oulu, Oulu, Finland
- 20 Department of Neurology and Medical Research Center, Oulu University Hospital, Oulu, Finland
- 21 Department of Neurology, Helsinki University Hospital, and Molecular Neurology, Research Programs Unit, Biomedicum, University of Helsinki, Helsinki, Finland
- 22 Neuroregeneration and Stem Cell Program, Institute for Cell Engineering, Johns Hopkins University Medical Center, Baltimore, MD, USA
- 23 Department of Neurology, Johns Hopkins University Medical Center, Baltimore, MD, USA
- 24 Laboratory of Behavioral Neuroscience, National Institute on Aging, Baltimore, MD, USA
- 25 Longitudinal Studies Section, National Institute on Aging, Baltimore, MD, USA
- 26 Newcastle Brain Tissue Resource, Institute of Neuroscience, Newcastle University, Newcastle upon Tyne, UK
- 27 Department of Pathology (Neuropathology, Johns Hopkins University Medical Center, Baltimore, MD, USA
- 28 Department of Neurology, Institute of Neurological Sciences, Queen Elizabeth University Hospital, Glasgow, UK
- 29 Inserm U1127, Sorbonne Universités, UPMC Univ Paris 06 UMR S1127, Institut du Cerveau et de la Moelle épinière, ICM, Paris, France
- 30 Preventive Neurology Unit, Wolfson Institute of Preventive Medicine, Queen Mary University of London, London, UK
- 31 Department of Neurodegenerative Diseases, UCL Queen Square Institute of Neurology, London, UK
- 32 Department of Neurology, University of Maryland School of Medicine, Baltimore, MD, USA
- 33 Department of Neurology, Baylor College of Medicine, Houston, USA
- 34 Departments of Molecular and Human Genetics and Neuroscience, Baylor College of Medicine, Houston, USA
- 35 Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, USA
- 36 Data Tecnica International, Glen Echo, MD, USA
- 37 Department of Neurology and Neurosurgery, McGill University, Montreal, Quebec, Canada

Correspondence to: Cornelis Blauwendraat
 Laboratory of Neurogenetics
 NIA, NIH, Building 35, 35 Convent Drive
 Bethesda, MD 20892, USA
 E-mail: cornelis.blauwendraat@nih.gov

Keywords: GBA; SNCA; CTSB; Parkinson's disease; modifiers

Abbreviation: GRS = genetic risk score; GWAS = genome-wide association studies; IPDGC = International Parkinson's Disease Genomics Consortium; iPSC = induced pluripotent stem cell; LBD = Lewy body dementia; PC = principal component; PPMI = Parkinson's Progression Marker Initiative

Introduction

Heterozygous functional *GBA* variants are one of the most common genetic risk factors for Parkinson's disease and Lewy body dementia (LBD), found in 3–20% of patients in different populations (Lesage *et al.*, 2011; Gan-Or *et al.*, 2015b; Blauwendraat *et al.*, 2018a; Guerreiro *et al.*, 2018; Rivas *et al.*, 2018). In a homozygous or compound heterozygous state, *GBA* variants may cause Gaucher disease, an autosomal recessive lysosomal storage disorder. *GBA* encodes the lysosomal enzyme glucocerebrosidase (GCase), and it is hypothesized that loss of GCase activity leads to a reduced ability to degrade α -synuclein, encoded by *SNCA*. Aggregated, misfolded α -synuclein protein deposits in the form of Lewy bodies and Lewy neurites are the pathological hallmarks of Parkinson's disease and LBD.

Notably, *GBA* variants that do not cause Gaucher disease but do confer increased risk for Parkinson's disease and LBD have been identified. It is hypothesized that while these variants result in reduced GCase activity, the activity is not low enough to cause Gaucher disease. Multiple rare *GBA* variants have been described in Parkinson's disease in different populations. More common variants include p.E326K, p.T369M, p.N370S and p.L444P, whose frequencies vary with ethnicity and are each found on different haplotypes (Blauwendraat *et al.*, 2018a; Leija-Salazar *et al.*, 2019). In some ethnicities, such as the Ashkenazi Jewish population, certain *GBA* variants are found in about 5% of unaffected individuals, and 17–20% of Parkinson's disease patients (Gan-Or *et al.*, 2008; Sidransky *et al.*, 2009; Ruskey *et al.*, 2019).

In European populations, the p.E326K risk variant is found in ~2% of the general population (<https://gnomad.broadinstitute.org/>). However, the majority of *GBA* variant carriers will not develop Parkinson's disease, implying that there are other genetic and/or environmental factors that affect the penetrance of these variants. Studies that have examined the penetrance of *GBA* variants in carriers suggest it is age-related and is typically between 10% and 30% (Anheim *et al.*, 2012; Rana *et al.*, 2013). Furthermore, penetrance differs across *GBA* variants, with high-risk variants leading to earlier disease onset compared to lower risk variants (Gan-Or *et al.*, 2015a). Interestingly, it was demonstrated that various *GBA* variants, including p.E326K, p.T369M and p.N370S have similar effects on GCase activity in humans, reducing it by 18–46% on average (Alcalay *et al.*, 2015). Additionally, genotype-phenotype studies have shown that Parkinson's disease patients with *GBA* variants have an earlier age at onset, faster disease progression, and higher rates of non-motor symptoms,

such as rapid eye movement (REM) sleep behaviour disorder (RBD), autonomic dysfunction, hallucinations and cognitive decline, compared to those with non-*GBA* associated Parkinson's disease (Gan-Or *et al.*, 2018).

In recent years, *GBA* has become a prominent target for therapeutic development, and the first gene-specific phase 2 clinical trial in Parkinson's disease is currently ongoing for *GBA*-positive Parkinson's disease patients (ClinicalTrials.gov Identifier: NCT02906020). One of the concerns in performing such a trial is that, despite the randomization for treatment and placebo groups, these groups will remain unbalanced in terms of factors that affect their progression, which can be a significant confounder of trial outcome (Leonard *et al.*, 2018). In addition, it is likely that future preventive clinical trials will also target populations that are at high risk for Lewy body diseases, such as *GBA* risk variant carriers with prodromal symptoms. Therefore, identifying factors that can affect the penetrance and clinical presentation of *GBA*-associated Parkinson's disease will be crucial for trial design. Furthermore, modifiers of penetrance age at onset may also become drug targets.

In the current study, we gathered data from multiple sources, including the International Parkinson's Disease Genomics Consortium (IPDGC) and the UK Biobank (UKB), to identify *GBA* variants in both cases and controls and used 23andMe and whole-genome sequencing data for further validation. Subsequently, we used genome-wide association studies (GWAS) and genetic risk scoring to identify genetic variants that modify the penetrance and age at onset of *GBA*-associated Parkinson's disease. We then examined protein levels of the top hits in forebrain neurons differentiated from induced pluripotent stem cells of individuals with and without *GBA* variants.

Materials and methods

Genotyping data

International Parkinson Disease Genomics Consortium genotyping data

Genotyping data (all Illumina platform based) was obtained from IPDGC members, collaborators, and public resources (Supplementary Tables 1 and 2). All datasets underwent quality control separately, both on individual-level data and variant-level data before imputation as previously described (Nalls *et al.*, 2018; Blauwendraat *et al.*, 2019). Refer to the online Supplementary material for a detailed description of data processing.

For GWAS analyses, variants passing the following post-imputation criteria were included: $R^2 > 0.8$, minor allele

frequency > 0.05 , and genotype missingness < 0.15 across subsets. Case-control GWAS were performed using RVTESTS with default settings (Zhan *et al.*, 2016) using logistic regression with sex, age (if not available the mean age of all data were used), principal components (PC) 1–5, and dataset origin as covariates. age at onset GWAS was performed in GBA carriers only using RVTESTS linear regression with age at onset as a continuous phenotype and sex, PCs 1–5 and dataset origin as covariates. Cases without age information were excluded from the age at onset GWAS, and individuals with two GBA variants were excluded from all analyses to prevent bias (Supplementary Table 2).

Lewy body dementia genotyping data

LBD cases and controls were genotyped for ongoing projects at the Neurodegenerative Diseases Research Unit (NDRU) using the NeuroChip genotyping array (Illumina). Genotyping was performed as previously described (Blauwendraat *et al.*, 2017). All cases met the McKeith criteria for probable or definite disease (McKeith, 2006). For more details on this dataset see Supplementary Table 2. Genotype quality control, processing, and analysis was performed similar to the above described IPDGC genotype data. For analyses, variants that passed the following post-imputation criteria were included: imputation $R^2 > 0.8$ and minor allele frequency > 0.05 . Case-control analyses were performed using RVTESTS using logistic regression on genotype dosages with sex, and PCs 1–5 as covariates. Age was not used since the majority of the individuals had only age of death information available.

McGill GBA genotyping data

Ashkenazi Jewish Parkinson's disease cases were genotyped at McGill University using the Illumina Human OmniExpress Array and custom SNPs of the NeuroX array (Nalls *et al.*, 2015a). Sanger sequencing or targeted next-generation sequencing confirmed GBA p.N370S status. For more details on this dataset, see Supplementary Table 2. GBA variant carriers from the McGill dataset included samples from Columbia University, NY, USA (Alcalay *et al.*, 2015) and Sheba Medical Center, Tel Hashomer, Israel (Ruskey *et al.*, 2019). Genotype quality control, processing, and analysis was performed similar to the above described genotyping data.

UK Biobank data

Imputed UKB genotype data (v3) were downloaded (April 2018) under application number 32663 (Sudlow *et al.*, 2015; Bycroft *et al.*, 2018). All European ancestry individuals (based on code 22006) were screened for the common GBA coding variants p.E326K, p.T369M (both directly genotyped) and p.N370S (imputed $R^2 = 0.81$). Parkinson's disease cases were identified using ICD10 code 'G20' from data fields 41202 (Diagnoses - main ICD10), 41204 (Diagnoses - secondary ICD10) and 40002 [Contributory (secondary) causes of death: ICD10]. Additionally, self-report Parkinson's disease cases were included from data field 20002 with code '1262'. 'Proxy' Parkinson's disease cases were included by using individuals who reported a parent with Parkinson's disease (paternal Parkinson's disease, data field 20107 and maternal Parkinson's disease, data field 20110), which have been previously shown to share genetic risk with Parkinson's

disease cases (Nalls *et al.*, 2018). Controls were set as subjects with no report of Parkinson's disease, no parent with Parkinson's disease and no sibling with Parkinson's disease (data field 20111). Covariates were obtained from the data fields: genetic sex (22001) batch (22000), age of recruitment (21022), and Townsend index (189). Individuals were filtered for relatedness ($PIHAT > 0.125$) based on the pre-imputed data. Imputed genotypes were converted to PLINK2 .pgen files using PLINK2 (version v2.00) and filtered for missingness (< 0.05) and minor allele frequency (> 0.05), Hardy-Weinberg equilibrium of $P \geq 1 \times 10^{-6}$ in controls and imputation quality ($R^2 > 0.8$). GWAS was performed using PLINK2 logistic regression with covariates including batch, age of recruitment, Townsend index, and five PCs created using FlashPCA (Abraham *et al.*, 2017). All UKB analyses were performed using population controls with age of recruitment ≥ 60 and for GBA carrier case-control analyses, the case-control ratio was set to 1:9 by randomly selected controls.

Additional datasets

Additional results for variants of interest were obtained from 23andMe and whole-genome sequencing data obtained from various resources. Analyses were performed using similar and identical methods as described above. Associations in 23andMe were performed using logistic regression correcting for age, sex, genotyping platform (v4 or v5) and five PCs. Genome-sequencing data were obtained from multiple ongoing projects at the Laboratory of Neurogenetics, see Supplementary material for a detailed description. Sample sizes can be found in Supplementary Table 3.

Genetic risk score analyses

To assess the influence of the genetic risk score (GRS), we calculated the GRS using the effect sizes from the most recent Parkinson's disease GWAS variants (Nalls *et al.*, 2018) (see Supplementary Table 4). Note that most of the data are included in the latest Parkinson's disease GWAS; however, for the calculation of the GRS in the UKB data we excluded the UKB summary statistics. Overall, at least 65 of these variants of interest were imputed with high quality ($R^2 > 0.8$) in the included datasets, allowing for a maximum of 15% variant missingness across any constituent data subset due to differing genotyping backbones. For the IPDGC dataset, rs356182 was replaced by rs356219, due to a higher imputation quality ($R^2 > 0.82$ and $D' 0.959$ between variants in the IPDGC reference panel). Variants from the GBA locus and LRRK2 p.G2019 were excluded from the GRS in analyses including GBA carriers only. The GRS was calculated and processed using PLINK (v1.90) for each individual as described previously (Nalls *et al.*, 2015b). To make the scale of analyses relating to the Parkinson's disease GRS more easily interpretable, Parkinson's disease GRSs were converted to Z-scores. GRS associations were performed with a logistic regression in R (v3.5.1) using case-control status or linear regression with age at onset as the outcome phenotype. As covariates, we included: sex, age (in case-control status comparison only), the first 10 PCs, and the dataset origin (due to the fact that the IPDGC data are derived from multiple sources see Supplementary Table 2). Area under the curve risk predictions

based on the GRS were performed using the pROC package (Robin *et al.*, 2011).

Additional analyses

GBA variants were annotated using ANNOVAR (Wang *et al.*, 2010). Nucleotide positions in *GBA* were based on build hg19 and amino acid positions were based on NM_001005742. Note that historically these *GBA* variants are also known as p.E365K → p.E326K (rs2230288), p.T408M → p.T369M (rs75548401), and p.N409S → p.N370S (rs76763715). Where possible, association results were meta-analysed with METAL (v.2011-03-25) using default settings (Willer *et al.*, 2010) or using the metafor R package (2.0) using fixed effects models (Viechtbauer, 2010). We excluded SNPs with a heterogeneity I^2 statistic > 80% and variants that were present in only one GWAS. Limited population stratification was seen despite low sample numbers in some GWAS (Supplementary Table 5). Differences in age at onset between *GBA* variants was tested using linear regression in R using age at onset as the dependent variable, correcting for sex, PCs 1–5, and the dataset origin as covariates. Conditional GCTA-cojo analyses (Yang *et al.*, 2011, 2012) were performed to identify whether there were multiple independent signals in loci of interest, using the previously described IPDGC reference panel (Nalls *et al.*, 2018; Blauwendraat *et al.*, 2019). Genetic interaction analyses were performed using R (version 3.5) using an interaction term created using $\text{SNP1} + 1 \times \text{SNP2} + 1$, followed by a regression using both variants and covariates (PCs 1–5, sex and age if available).

Exploratory gene expression values and functional annotation of identified risk variants were obtained by mining datasets including: the largest brain expression quantitative trait loci (eQTL) study to date (Qi *et al.*, 2018), the GTEx QTL portal V7 (<https://gtexportal.org/>) (Consortium, 2013) and the NABEC dataset [phs001300.v1.p1 (Gibbs *et al.*, 2010)]. LocusZoom (Pruim *et al.*, 2010) and LocusCompare plots (Liu *et al.*, 2019) were generated for the regions of interest and were compared to the latest published Parkinson's disease GWAS. Single-nuclei RNAseq expression of three frozen prefrontal cortex samples and four frozen hippocampal samples were obtained from dbGaP accession phs000424.v8.p1 (Habib *et al.*, 2017). Data were processed using Seurat and genes of interest were highlighted to explore expression patterns (Butler *et al.*, 2018).

GBA variant validation

To validate the genotyping and imputation of *GBA* variants, we obtained next-generation sequencing (targeted resequencing and genome sequencing) data from two parallel IPDGC projects. *GBA* carrier genotypes were merged using PLINK (v1.90) with both the resequencing and whole genome sequencing data. Overlapping individuals were identified using > 0.9 PIHAT values. Additionally, the majority of the PROBAND dataset individuals were Sanger sequenced for the full *GBA* gene as described elsewhere (Malek *et al.*, 2018). In the McGill dataset, *GBA* variants were confirmed using Sanger sequencing or targeted next generation sequencing as previously described (Ruskey *et al.*, 2019). Overall, the concordance rate between the imputed/genotype data and the sequence data for the presence of at least one p.N370S and p.T369M allele

was 100% and for p.E326K was 98% and 94% in the next-generation and Sanger sequencing data, respectively (see Supplementary material for more details).

Induced pluripotent stem cells and neuron analysis

Forebrain neuron differentiation

Induced pluripotent stem cells (iPSCs) from the Parkinson's Progression Marker Initiative (PPMI) were selected based on their genotype at *GBA* and rs1293298 then differentiated to forebrain type neurons. iPSCs were grown in E8 media (ThermoFisher) on Matrigel® until 90% confluent. When cells had grown to confluence they were transitioned to N3 media (50% DMEM/F12, 50% Neurobasal™ with $1 \times$ penicillin-streptomycin, $0.5 \times$ B-27™ minus vitamin A, $0.5 \times$ N2 supplement, $1 \times$ GlutaMAX™, $1 \times$ NEAA, 0.055 mM 2-mercaptoethanol and 1 µg/ml Insulin) plus 1.5 µM dorsomorphin (Tocris Bioscience) and 10 µM SB431542 (Stemgent) as previously reported (Burkhardt *et al.*, 2013). Media was replaced every day for 11 days. On Day 12 dorsomorphin and SB431542 were removed and cells continued to be fed each day with N3. N3 was supplemented with 0.05 µM retinoic acid from Days 16 to 20. The differentiating cells were split 1:2 with Accutase® and seeded with ROCK inhibitor onto poly-L-ornithine (Sigma), fibronectin (2 µg/ml) and laminin (0.2 µg/ml) coated plates on Day 20. Cells were then fed daily with N4 media (same as N3 plus 0.05 µM retinoic acid, 2 ng/ml BDNF and 2 ng/ml GDNF) until Day 26 when they were frozen in Synth-a-Freeze™. Vials were thawed as needed in N4 media. Differentiation state was confirmed by immunocytochemistry for neuronal markers β -III tubulin (Novus Biologicals, NB100-1612) and MAP2 (Santa Cruz Biotechnology, sc-20172) with nuclei counterstained for Hoechst 33342 (ThermoFisher Scientific, H3570).

Protein analysis

Protein lysates were collected in buffer with 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM EDTA, 0.3% Triton™ 10% glycerol, $1 \times$ protease inhibitors, and $1 \times$ phosphatase inhibitors and run as previously reported (Blauwendraat *et al.*, 2018b) on 4–20% Criterion™ TGX™ gels (Bio-Rad) in Tris-glycine running buffer and transferred to nitrocellulose membranes (Bio-Rad) using the Trans-blot® Turbo system (Bio-Rad). Membranes were blocked in Odyssey® Blocking buffer (LI-COR) diluted by 50% in phosphate-buffered saline. Primary antibodies for cathepsin B (Abcam, ab92955) and α -synuclein (BD Biosciences, 610787) were incubated overnight at 4°C. Primary antibody to β -III tubulin (Novus Biologicals, NB100-1612) was incubated for 1 h at room temperature. LI-COR secondary IRDye® antibodies were incubated 1 h at room temperature. Proteins were visualized on the Odyssey® CLx Imaging System (LI-COR) and quantified using Image Studio Lite (LI-COR).

Data availability

Summary statistics from the GWAS are available at <https://pdgenetics.org/resources>.

Results

Initial data overview

We included 21 478 Parkinson's disease cases and 24 388 controls from IPDGC datasets; 1176 Parkinson's disease cases, 13 431 proxy Parkinson's disease cases, and 155 325 population controls from the UKB dataset; 103 GBA p.N370S Parkinson's disease cases from the McGill dataset; and 622 LBD cases and 782 controls from the NDRU dataset (Supplementary Table 2). In the Haplotype Reference Consortium (HRC) reference panel, 10 coding GBA variants are present (Supplementary Table 6), with varying imputation qualities as common variants are more easily imputed than variants with lower frequencies. Additionally, some GBA variants were directly genotyped in some datasets (Supplementary Table 1). The three most common variants (p.E326K, p.T369M and p.N370S) had the highest quality, we therefore selected these three variants for all subsequent analyses. This resulted in a total of 1691 GBA carriers with Parkinson's disease, 81 with LBD, 711 Parkinson's disease proxy cases, and 7624 control GBA carriers. We excluded individuals carrying two GBA variants to ensure conservative estimates and reduce bias due to the association with Gaucher disease (Supplementary Table 2). All included GBA variants (p.E326K, p.T369M and p.N370S) have previously been shown to be enriched in cases in all datasets.

Parkinson's disease genetic risk score modifies disease risk in GBA carriers

Exploring the potential effect of Parkinson's disease case-control GWAS loci, we calculated the Parkinson's disease GRS, both including and excluding GBA variants and LRRK2 p.G2019S for all three datasets. To compare the effect size associated with the GRS, we examined four groups within each dataset: controls without a GBA risk variant, Parkinson's disease cases without a GBA risk variant, controls with a GBA risk variant, and Parkinson's disease cases with a GBA risk variant. Additionally, for the UKB dataset, we included proxy-Parkinson's disease cases (having a parent with Parkinson's disease) (Fig. 1). As expected, the GRS is significantly associated with risk in GBA-negative Parkinson's disease cases compared to controls and interestingly, also in the NDRU-LBD dataset (Fig. 1 and Table 1). This effect is highly similar between GBA carrier controls and GBA carrier cases, when excluding GBA and LRRK2 variants from the GRS (Fig. 2 and Table 1). This effect is consistent in subsets based on all three GBA variants (Supplementary Fig. 1). Although this was not significant in the (smaller) p.N370S group, the direction and effect size were highly similar to the other GBA variants [Supplementary Fig. 1, odds ratios (ORs) ranging from 1.31 to 1.55]. Notably, we do not observe a consistent difference in genetic risk between Parkinson's

disease or LBD cases with and without GBA variants when excluding GBA and LRRK2 in the genetic risk (Table 1).

When assessing the predictive power of the GRS in GBA positive cases versus GBA-positive controls (excluding GBA and LRRK2 variants in GRS calculations), these values were slightly lower compared to the reported predictive power of the GRS in Parkinson's disease cases as previously described by Nalls *et al.* (2018) [area under the curve (AUC) = 0.651, 95% confidence interval (CI) 0.617–0.684], with AUC of 0.614 (95% CI 0.591–0.637) in the IPDGC dataset and 0.622 (95% CI 0.561–0.684) in the UKB dataset (Supplementary Table 7 and Supplementary Figs 2–4). As expected, the NDRU-LBD dataset performed poorer than the Parkinson's disease cases with an AUC of 0.570 (95% CI 0.540–0.600), supporting the notion of partial overlapping and shared genetic risk between Parkinson's disease and LBD.

SNCA and CTSB loci as genetic modifiers of disease risk in GBA carriers

To identify potential genetic modifiers of GBA-associated Parkinson's disease penetrance, we performed a GWAS comparing Parkinson's disease patients with GBA variants and controls with GBA variants. We first examined the association using all three GBA variants together (p.E326K, p.T369M and p.N370S) in the IPDGC and the UKB dataset, followed by a meta-analysis of both. After meta-analysing the GBA data in both datasets, no genome-wide significant variants were identified (Supplemental Fig. 5).

When solely examining the 90 Parkinson's disease variants previously identified in a Parkinson's disease case-control GWAS (Nalls *et al.*, 2018), variants in two Parkinson's disease loci passed Bonferroni multiple test correction for association with Parkinson's disease, SNCA and CTSB. The strongest association was observed with rs356219 in the SNCA locus (meta- $P = 9.26 \times 10^{-7}$, OR = 1.375, 95% CI = 1.211–1.563, Table 2 and Supplementary Fig. 6). In addition to SNCA, the CTSB locus rs1293298 (meta- $P = 9.74 \times 10^{-6}$, OR = 0.725, 95% CI = 0.629–0.836, Table 2 and Supplementary Fig. 7) was also significantly associated with GBA penetrance after correction for multiple testing (Supplementary Table 8). Additional investigation of these two variants was performed in the 23andMe Parkinson's disease case-control dataset and in a WGS dataset (Table 2). Meta-analyses of these results showed an increase of significance level for the CTSB variant ($P = 6.34 \times 10^{-6}$) and a genome-wide significant P -value ($P = 2.23 \times 10^{-11}$) for the SNCA variant (Table 2 and Supplementary Figs 9 and 10). The majority of these effects are mostly driven by p.E326K carriers likely due to the larger sample size, but the effect size was consistent for each of the three variants when tested

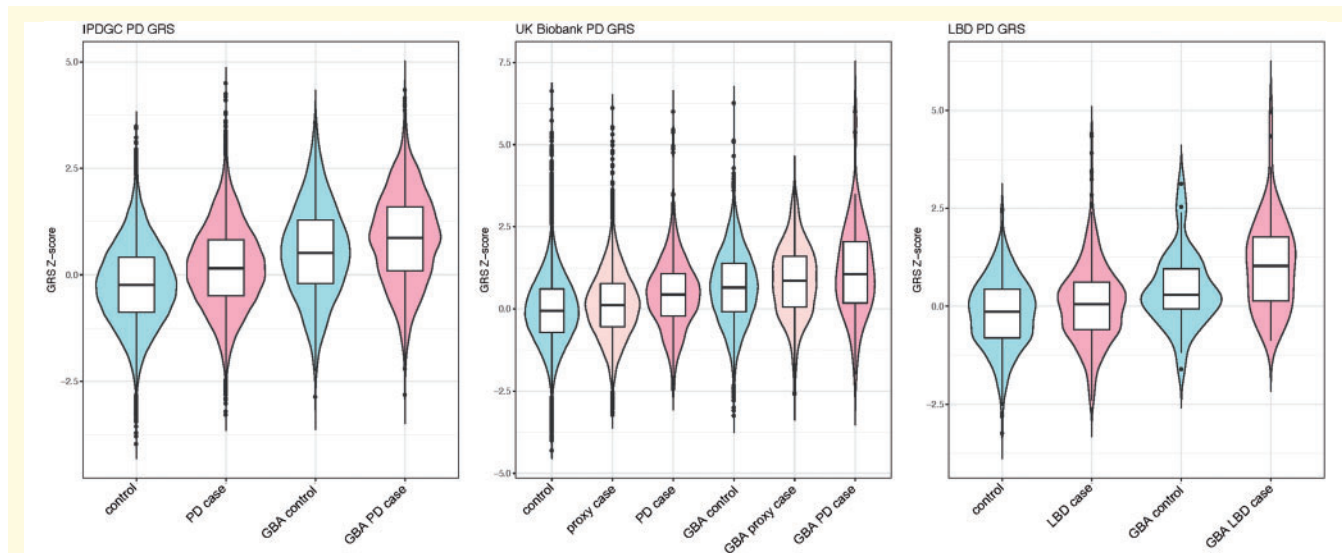


Figure 1 Parkinson disease GRS box plots divided by disease status and GBA carrier status. GRS differences between GBA negative controls, GBA negative Parkinson's disease or LBD cases, GBA positive controls and GBA positive cases. In each dataset, a highly similar effect is observed clearly separating the subgroups. PD = Parkinson's disease.

individually in the IPDGC dataset (Supplementary Table 9 and Supplementary Figs 6 and 7).

Evidence for a potential genetic interaction between *CTSB* and *GBA*

Interestingly, the OR of the *CTSB* variant rs1293298 association in the *GBA* GWAS is lower compared to the Parkinson's disease case-control GWAS [Supplementary Table 4 (Nalls *et al.*, 2018) OR = 0.911, 95% CI = 0.891–0.932 (Supplementary Fig. 8)], while for the *SNCA* variant this was not the case. This suggests a potential interaction between *CTSB* and *GBA*. Examining this possibility further resulted in a significant interaction term after meta-analysing effects of the IPDGC and UKB data (Table 2). As expected, no consistent significant interaction term was identified between *SNCA* and *GBA* (Table 2 and Supplementary Fig. 11). Replication of these results was investigated in the 23andMe and LNG Genomes data for both interactions. In both datasets the interaction terms were not significant, although the direction of effect was the same for *CTSB* × *GBA* (Table 2 and Supplementary Fig. 12).

Genetic dissection of the *CTSB* GWAS locus

To evaluate the potential molecular mechanism for increased risk of Parkinson's disease in *GBA* carriers, we explored gene expression and eQTL datasets for the two significant associated loci, *SNCA* (rs356219) and *CTSB* (rs1293298). For the *SNCA* locus, several previous reports showed that risk variants in this locus likely increase *SNCA*

expression (Soldner *et al.*, 2016; Pihlstrøm *et al.*, 2018). For the *CTSB*, no comprehensive assessment has yet been performed. In both the Parkinson's disease GWAS and the *GBA* GWAS, rs1293298 appears to be the index variant and both GWAS signals appears to be highly similar (Fig. 3A–C).

The protective allele (A) of the *CTSB* variant rs1293298 increases *CTSB* expression in the brain in both GTEx and NABEC datasets and was significant in the latest meta-analyses of brain eQTL data (Supplementary Table 10). Results from the NABEC dataset suggest that specific isoforms of *CTSB* are differentially regulated and the QTL effect is mainly for the most highly expressed transcript (isoform: uc003wuq.3, Supplementary Fig. 13). Interestingly the QTL signal shows high correlation with the Parkinson's disease GWAS signal (Fig. 3D).

Exploring post-mortem brain-derived single nuclei RNAseq data showed that *CTSB* is mainly expressed in neurons and microglia in a pattern that is similar to *SNCA* expression (Supplementary Fig. 14). This suggests that the biological effects of these variants possibly take place in these cell types, which is in line with pathology and previous literature. We attempted to examine *GBA* expression levels, but they were too low to assess.

Given the function of cathepsin B (encoded by *CTSB*) as lysosomal protease, which is of interest for Parkinson's disease pathogenesis in general (McGlinchey and Lee, 2015), we explored other members of the cathepsin family for: (i) genetic association in the most recent Parkinson's disease GWAS; (ii) genetic association in *GBA* GWAS; and (iii) the presence of an eQTL at each locus. In addition to the peak at *CTSB*, only *CTSE* has a Parkinson's disease GWAS peak in close proximity; however, this locus is already nominated as the *NUCKS1/RAB29* locus (Supplementary

Table 1 Associations between the Parkinson disease GRS and case-control status

Outcome	Case-control		GBA +, Case-control ^a		GBA + case versus GBA – case ^a	
	P	Cases, n; controls, n	OR (95%CI)	P	OR (95%CI)	GBA + cases, n; GBA – cases, n
IPDGC	1.877 × 10 ^{−267}	21 478; 24 388	1.531 (1.51–1.56)	2.75 × 10 ^{−12}	0.926 (0.87–0.98)	1392; 18 062
UKB	9.156 × 10 ^{−55}	1176; 155 185	1.557 (1.5–1.61)	1.96 × 10 ^{−4}	1.050 (0.82–1.28)	82; 1094
UKB proxy-case	1.284 × 10 ^{−35}	13 291; 155 185	1.138 (1.12–1.16)	2.12 × 10 ^{−5}	NA	NA
NDRU-LBD	6.88 × 10 ^{−8}	622; 782	1.370 (1.26–1.48)	0.0798	1.017 (0.78–1.26)	81; 541

GRS associations performed per group. Odds ratios are based on a Z-score scale as described in the methods. NA = not applicable; SE = standard error; UKB = UK Biobank.

^aFor these associations, the GBA region and LRRK2 p.G2019S were excluded when calculating the GRS.

Table 11). Interestingly, for almost all cathepsins a brain eQTL has been identified. However, with the exception of *CTSB*, none of the eQTLs were associated with Parkinson's disease (Supplementary Table 11). These results suggest that expression differences in other cathepsins, including *CTSD* and *CTSL*, which were previously suggested to degrade α -synuclein (McGlinchey and Lee, 2015), do not modulate risk for disease.

Cathepsin B and α -synuclein levels in iPSC-derived neurons from GBA carriers and non-carriers

We examined the potential relationship between *GBA* and *CTSB* using PPMI iPSCs differentiated to forebrain neurons. For this analysis we used iPSCs from individuals with and without the *GBA* p.N370S variant to interrogate cathepsin B and α -synuclein protein levels in the context of *GBA* variants. All lines used were heterozygous for rs1293298 at the *CTSB* locus and neuronal differentiation was confirmed by immunocytochemistry for β -III tubulin and MAP2 (Supplementary Fig. 15). Previous work has shown that cathepsin B protein is highly cleaved and processed before becoming active (Ritonja *et al.*, 1985; Chan *et al.*, 1986). The protein can be cleaved to a final heavy (24 kDa) and light chain (5 kDa) or they can remain uncleaved and run at 29 kDa. Processing of cathepsin B is cell type-dependent (Chen *et al.*, 2019) and in our cells we see a double band pattern corresponding to the heavy plus light chain and the heavy chain alone. When quantifying both bands we observed a significantly lower ($P = 0.0291$; unpaired *t*-test, $n = 6$) level of active cathepsin B protein in *GBA* variant carriers compared to cells without *GBA* variants (*GBA* wild-type = 0.528 ± 0.0930 ; *GBA* p.N370S = 0.263 ± 0.0467) (Fig. 4). These results show that the *GBA* variant carriers have reduced levels of active cathepsin B, which may result in even lower lysosomal protease activity and increased accumulation of protein aggregates in neurons. Although the forebrain neurons carrying the *GBA* variant did not have increased levels of α -synuclein (Supplementary Fig. 16) it is possible that the overall reduction in lysosomal proteases allows for a faster accumulation of α -synuclein aggregates as neurons age.

Genetic modifiers of age at onset in GBA associated Parkinson's disease cases

Information on age at onset was available for the majority of the IPDGC *GBA* Parkinson's disease cases [89.9% $n = 1353$ cases, average = 60.54 years, standard deviation (SD) = 11.42]. Cases with two *GBA* variants (either homozygous or compound heterozygous, $n = 38$) had a lower age at onset compared to single variant carriers; however, no statistically significant difference was detected [$P = 0.107$; linear regression, Beta = -2.93 years, standard error (SE)

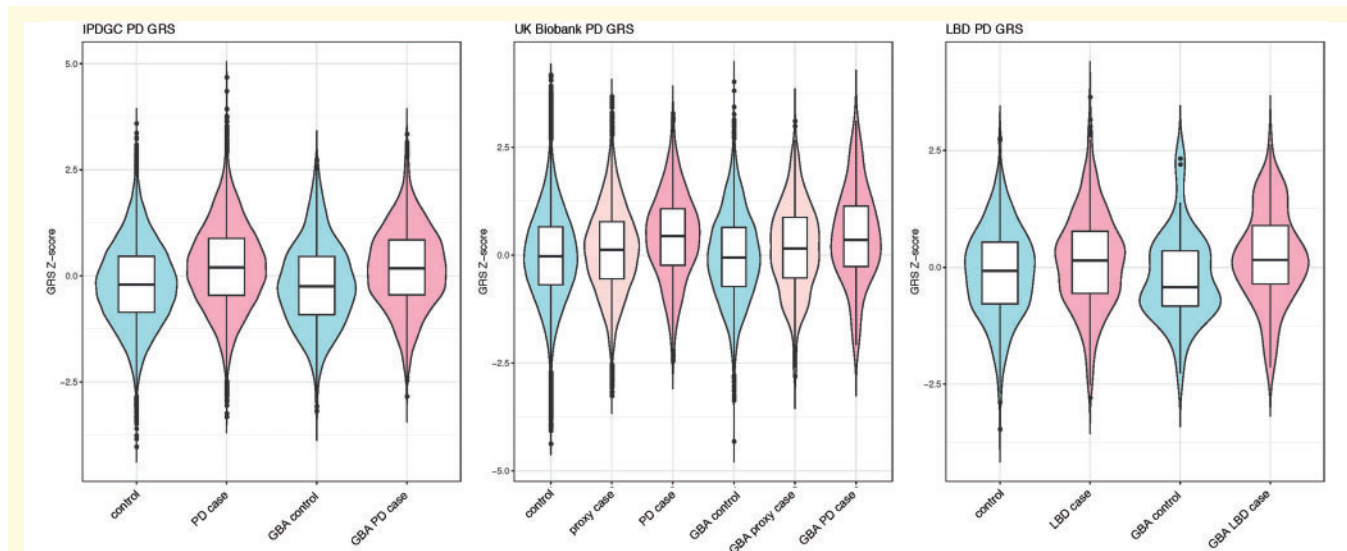


Figure 2 Parkinson disease GRS box plots divided by disease status and GBA carrier status excluding GBA in the GRS calculation. GRS differences between GBA negative controls, GBA negative Parkinson's disease or Lewy body dementia cases, GBA positive controls and GBA positive cases when excluding the GBA region and *LRRK2* p.G2019S from the calculations. In each dataset, a highly similar effect is observed clearly separating the cases and controls. However, there is no clear separation between GBA cases and GBA negative cases or GBA negative controls and GBA positive controls. PD = Parkinson's disease.

= 1.82, two GBA variant case average = 56.81, SD = 10.50 versus single GBA variant case average = 60.54, SD = 11.42]. The age at onset of GBA Parkinson's disease cases was significantly lower compared to non-GBA Parkinson's disease cases [linear regression, $P = 9.70 \times 10^{-6}$, Beta = -1.43, SE = 0.324, GBA Parkinson's disease case average (one GBA mutation) = 60.54, SD = 11.42 versus non-GBA case average = 62.06, SD = 12.01]. On average p.N370S cases had the earliest age at onset followed by p.E326K and p.T369M. Of note, the *LRRK2* p.G2019S variant was also identified in 14 GBA carriers (nine p.N370S and five p.E326K carriers) and did not seem to influence the age at onset $P = 0.44$ (linear regression, Beta = 2.41, SE = 3.15, GBA Parkinson's disease *LRRK2* p.G2019S carriers average = 63.86, SD = 10.87 versus GBA Parkinson's disease non-*LRRK2* p.G2019S average = 60.29, SD = 11.80), which is consistent with previous findings, although power was limited given the small number of included individuals (Yahalom *et al.*, 2019).

To identify potential GBA age at onset modifiers, we performed a GWAS using age at onset as phenotype in all GBA positive Parkinson's disease cases ($n = 1353$). No genome-wide hits were identified, most likely because of the relatively low sample size. When investigating previously identified Parkinson's disease loci, no variants passed Bonferroni correction. However, nominally significant signals were identified for several Parkinson's disease loci, including *SNCA* and *TMEM175* (Supplementary Table 12). Similar to Parkinson's disease in general, the Parkinson's disease GRS is significantly associated with a lower age at onset in GBA carriers (Fig. 5) (Blauwendraat *et al.*, 2019). This effect was mainly driven by p.E326K

carriers. However, both p.T369M and p.N370S carriers also show the same trend towards lower age at onset with a similar effect size that is close to significance likely due to lower sample numbers (Fig. 5). This association was similar in an independent p.N370S Ashkenazi Jewish ancestry dataset, again demonstrating the same general trend and similar effect size. A meta-analysis of the three tested GBA variants resulted in a significant P -value of 0.0050 (Fig. 5 and Supplementary Fig. 17).

Discussion

In the current study, we analysed multiple large datasets to examine whether genetic variants affect the penetrance and age at onset of GBA-associated neurodegeneration. Our data show that at least 7% of the European Parkinson's disease population carries a common protein coding GBA risk variant versus ~4.5% of the general population. It is noteworthy that this number is an underestimation because in some datasets p.T369M and p.N370S were not genotyped or reliably imputed, and rarer disease-linked GBA variants were not captured. For LBD, these numbers are likely to be higher as seen in the NDRU-LBD dataset with ~13% of the LBD cases carrying a GBA variant.

One of the major findings of this study is that Parkinson's disease and LBD cases with GBA variants often also carry a substantial number of other Parkinson's disease associated risk variants, which modify disease risk and age at onset. The GRS is significantly higher in cases with a GBA variant compared to controls with a GBA variant (Table 1) with effect sizes that are very

Table 2 Meta-analyses of significant GWAS signals and interaction effects

	IPDGC			UKB			Meta			23andMe			LNG Genomes			Meta_all		
	P	OR (CI)	P	P	OR (CI)	P	P	OR (CI)	P	P	OR (CI)	P	P	OR (CI)	P	P	OR (CI)	P
Association																		
rs356219 (SNCA)	6.37×10^{-5}	1.315 (1.150–1.504)	3.99×10^{-4}	2.085 (1.388–3.132)	9.26 $\times 10^{-7}$	1.376 (1.211–1.563)	6.89×10^{-7}	1.273 (1.157–1.400)	0.6645	0.921 (0.635–1.336)	2.23×10^{-11}	0.6645	0.921 (0.635–1.336)	2.23×10^{-11}	0.6645	0.921 (0.635–1.336)	1.290 (1.198–1.391)	0.6645
rs1293298 (CTSB)	8.27×10^{-5}	0.740 (0.637–0.860)	0.0304478	0.614 (0.394–0.955)	9.74×10^{-6}	0.726 (0.630–0.837)	0.0564	0.898 (0.805–1.003)	0.05908	0.6616 (0.431–1.016)	6.34×10^{-6}	0.05908	0.6616 (0.431–1.016)	6.34×10^{-6}	0.05908	0.6616 (0.431–1.016)	0.822 (0.755–0.895)	0.05908
Interaction																		
rs356219 $\times$ GBA	0.7025	1.029 (0.889–1.190)	0.00639	1.600 (1.141–2.244)	0.152	1.103 (0.965–1.261)	0.773	0.985 (0.891–1.089)	0.3697	0.823 (0.537–1.261)	0.653	0.3697	0.823 (0.537–1.261)	0.653	0.3697	0.823 (0.537–1.261)	1.018 (0.941–1.102)	0.3697
rs1293298 $\times$ GBA	0.00647	0.796 (0.676–0.938)	0.28833	0.802 (0.534–1.205)	0.00347	0.797 (0.685–0.928)	0.554	0.966 (0.860–1.084)	0.0663	0.625 (0.379–1.032)	0.0113	0.0663	0.625 (0.379–1.032)	0.0113	0.0663	0.625 (0.379–1.032)	0.890 (0.813–0.974)	0.0113

Two variants were significant after correction of multiple testing: rs356219 (SNCA) and rs1293298 (CTSB) in the initial used datasets. Additional datasets were assessed and overall both P-values became more significant. The interaction effects between the two variants and GBA were not replicated in the external datasets.

similar to general Parkinson's disease compared to controls. These findings suggest that GRS could be considered when designing clinical trials for GBA Parkinson's disease/LBD. Pre-trial genetic analysis could assist in stratifying patients according to Parkinson's disease-related genetic variants, on top of GBA. Therefore, future clinical trials involving GBA carriers should include a comprehensive genetic assessment prior to enrolment, as previously suggested (Leonard *et al.*, 2018). Disease predictions based on current genetic knowledge should be combined with non-genetic data, such as family history and pre-symptomatic phenotypes including RBD and smell test to maximize disease prediction (Nalls *et al.*, 2015c).

The link between Parkinson's disease and lysosomal pathways has already been established by both genetic and functional studies (Gan-Or *et al.*, 2015c; Robak *et al.*, 2017; Ysselstein *et al.*, 2019). Moreover, in the latest Parkinson's disease GWAS (Nalls *et al.*, 2018) numerous lysosomal and autophagy-related genes were associated with Parkinson's disease, emphasizing the central role of the lysosomal-autophagy pathway in disease. Interestingly, the two main loci identified that influence disease risk in GBA carriers are variants in close proximity to SNCA and CTSB (rs356219 and rs1293298), both implicated in the lysosomal autophagy pathway. The SNCA locus is well-studied and the current hypothesis is that variants in this locus increase SNCA expression of either total α -synuclein or specific mRNA isoforms, resulting in an increased risk of Parkinson's disease (Soldner *et al.*, 2016; Pihlström *et al.*, 2018).

For the CTSB locus, no functional studies have been performed to dissect this locus in Parkinson's disease to date, although several genetic studies have nominated CTSB as causal gene in this region via either QTL analyses or Mendelian randomization (Nalls *et al.*, 2018; Li *et al.*, 2019). Therefore, it remains unclear whether CTSB is the causal gene in this locus. However, there is evidence linking CTSB to Parkinson's disease pathways. In this study, we see a decrease in active cathepsin B protein levels in iPSC-derived neurons from GBA variant carriers compared to non-carriers, suggesting a further reduction in lysosomal protease function in these cases. Depletion of TMEM175, another lysosomal gene associated with Parkinson's disease, significantly decreases protein expression and enzyme activity of cathepsin B in rat hippocampal neurons (Jinn *et al.*, 2017, 2019). Cathepsin B has also been suggested to be involved in α -synuclein lysosomal degradation (McGlinchey and Lee, 2015). Moreover, ceramide, which is one of the lipids produced by GCase, has been shown to be an activator of cathepsin B-related pathways (Liu *et al.*, 2016). Overall, these results fit with the general model that GBA risk variant carriers have impaired lysosomal function due to reduced or abolished GCase activity (Alcalay (Alcalay *et al.*, 2015). The independent associations of the SNCA and CTSB loci with GBA associated Parkinson's disease suggest that increased SNCA expression and/or decreased CTSB mRNA expression may further impair

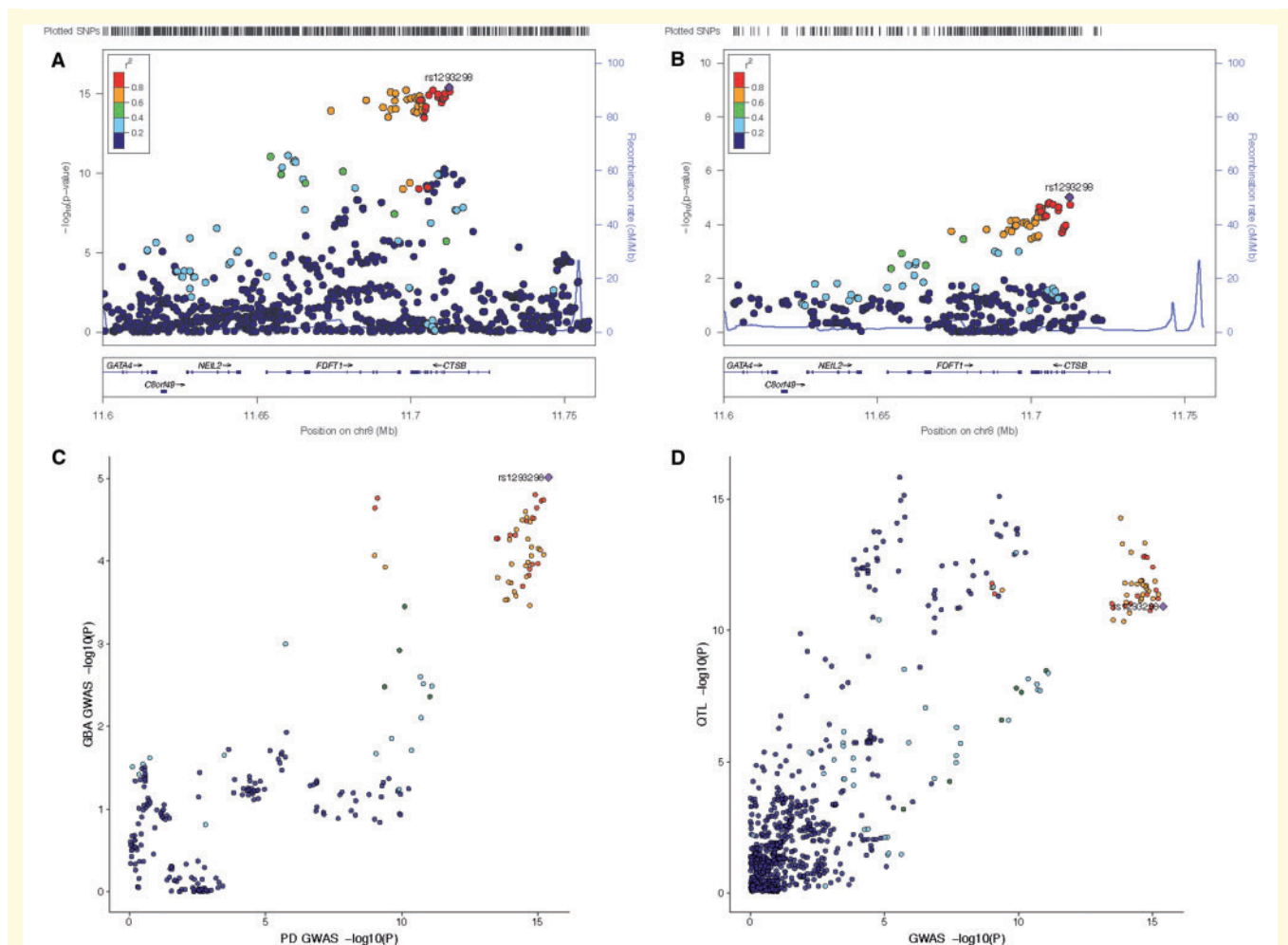


Figure 3 Genetic dissection of the *CTSB* locus. (A) LocusZoom plot of the *CTSB* region from the latest Parkinson's disease GWAS (Nalls et al., 2018). (B) LocusZoom plot of the *CTSB* region from the latest current GBA carrier GWAS. (C) LocusCompare plot of the *CTSB* region showing high correlation between the Parkinson's disease GWAS (x-axis) and the GBA carrier GWAS (y-axis). (D) LocusCompare plot of the *CTSB* region showing high correlation between the Parkinson's disease GWAS (x-axis) and the brain QTL data (Qi et al., 2018) (y-axis).

lysosomal function, resulting in an increased risk of Parkinson's disease. Additionally, we identified that cathepsin B protein levels are reduced in p.N370S neurons compared to wild-type. Overall, these findings suggest that cathepsin B might play a larger role in Parkinson's disease than previously thought and that increasing *CTSB*/cathepsin B levels could be a potential therapeutic strategy. However, extensive functional *in vitro* and *in vivo* work is needed to validate these effects.

Despite the large datasets included in this study, it has several limitations. First, only one genome-wide significant locus was identified (*SNCA*) after meta-analysing with the replication datasets, most likely because of the lack of power caused by the relatively low number of *GBA* variant carriers available. However, when we sought to reduce the burden of genome-wide correction by examining previously implicated Parkinson's disease risk loci, a significant association with the *CTSB* locus was identified. Second, the majority of the analyses were based on aggregated

genotype data derived from several different genotyping arrays and centres/countries. As such, some datasets did not contain data on all three of the *GBA* variants of interest. It is also important to consider that different centres/countries might have used different measures of age at onset or age at diagnosis and that included controls may develop Parkinson's disease at a later age. However, using the current data, we are still able to identify large effects, such as those driven by the GRS. Larger future initiatives specifically focused on *GBA* carriers will resolve these limitations and are needed to confirm these associations. Third, because the majority of the data included here were also used in the most recent Parkinson's disease GWAS meta-analysis (Nalls et al., 2018); there could be a bias towards the identification of known Parkinson's disease risk signals in this *GBA* dataset. We addressed this potential bias by looking for replication in several datasets that have not been included in the most recent Parkinson's disease GWAS meta-analysis, such as the UKB Parkinson's disease

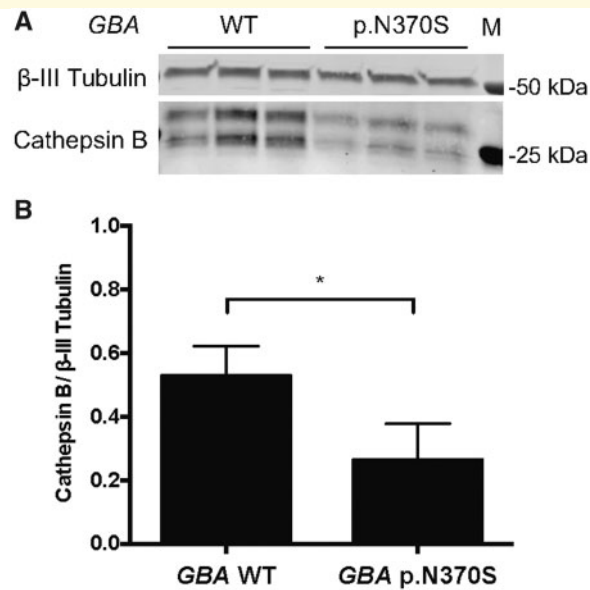


Figure 4 Cathepsin B protein expression is decreased in iPSC derived neurons from *GBA* p.N370S carriers. Forebrain neurons differentiated from patient iPSCs with *GBA* mutations have decreased cathepsin B protein expression. **(A)** Representative western blot showing protein levels of active cathepsin B (CTSB, 24 and 29 kDa) and neuron associated β -III tubulin (50 kDa) in forebrain neurons differentiated from individuals with and without Parkinson's disease-associated *GBA* variants ($n = 2$ lines from each genotype; all lines tested are heterozygous at rs1293298). **(B)** Quantification of the mean active cathepsin B protein levels relative to β -III tubulin shows a significant decrease ($*P < 0.05$; unpaired t -test, $n = 2$ lines of each genotype with three technical replicates each) in cathepsin B levels in forebrain neurons differentiated from *GBA* p.N370S carriers compared to neurons from wild-type (WT) *GBA* carriers. M = protein marker; error bars represent standard error of the mean. Note that both the heavy chain and heavy plus light chain bands were quantified in this analysis.

cases, the LBD dataset and the McGill *GBA* dataset. Unfortunately, we did not have a replication dataset for the *GBA* age at onset results, so these results require replication in an independent dataset. Fourth, some datasets might have excluded known *GBA* carriers prior to genotyping, which would result in an underestimation of total *GBA* carriers and the associated risk estimates. Of note, another relatively common *GBA* variant p.L444P (rs421016) is currently not imputable, as it is not included in the HRC reference panel and is in a complex region of the genome with a highly similar pseudogene. Finally, the functional data presented in Fig. 4 suggests a functional interaction between *GBA* and CTSB; however, it does not functionally validate the expression changes associated with the rs1293298 variant. Additional, larger experiments are needed to explore these findings further.

Despite these limitations, several conclusions can be drawn from our results. First, the risk of Parkinson's disease in *GBA* carriers is influenced by specific variants at

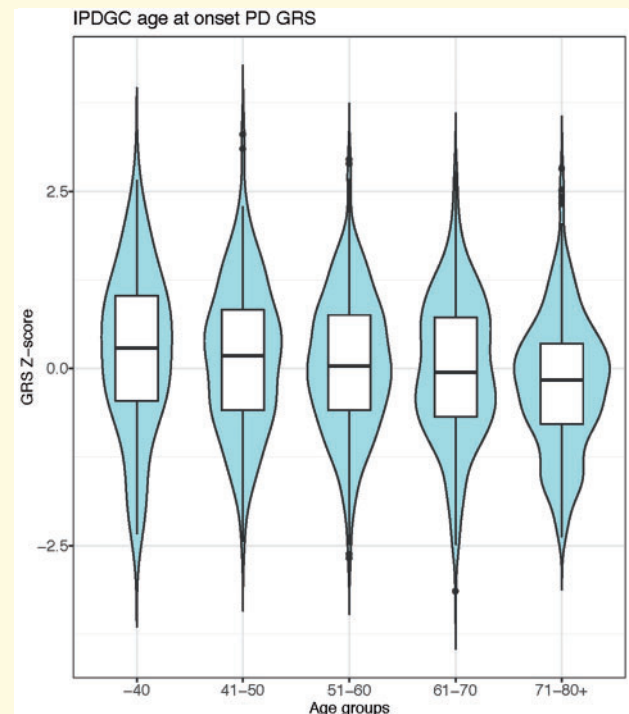


Figure 5 Parkinson's disease GRS significantly influences the age at onset of *GBA* positive cases. Parkinson's disease cases carrying a *GBA* variant were grouped by ages and a clear and significant decline was seen for the GRS. PD = Parkinson's disease.

loci that are known to be associated more generally with Parkinson's disease risk, including *SNCA* and *CTSB*, the latter possibly through interaction with *GBA*. In addition, *GBA* associated risk is significantly affected by the GRS; however, the total contribution of the GRS is not large. In other words, common genetic factors do not fully explain the partial penetrance of *GBA* variants in Parkinson's disease, hence rare variants, environmental and ageing-related factors probably contribute to *GBA* penetrance. Furthermore, variants in known Parkinson's disease loci, such as *SNCA* and *TMEM175*, as well as the GRS, might affect the age at onset of *GBA*-associated parkinsonism at a similar scale compared to the effect seen in typical Parkinson's disease. Overall, these results confirm the importance of the lysosomal pathway in Parkinson's disease and further supports *CTSB* and *SNCA* as interesting candidate genes for functional analysis in both *GBA*-associated and general Parkinson's disease.

Acknowledgements

We would like to thank all of the subjects who donated their time and biological samples to be part of this study. We would also like to thank all members of the International Parkinson Disease Genomics Consortium

(IPDGC). The authors thank Ole Andreassen and the DemGene consortium for genotyping of the Oslo cohort. Samples used in the analyses presented in this article were obtained from the Golub Capital iPSC Parkinson's Progression Markers Initiative (PPMI) Sub-study (www.ppmi-info.org/cell-lines). As such, the investigators within PPMI contributed to the design and implementation of PPMI and/or provided data and collected samples but did not participate in the analysis or writing of this report. For up-to-date information on the study, visit www.ppmi-info.org. Data used in the preparation of this article were obtained from the PPMI database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. Data and biospecimens used in preparation of this manuscript were obtained from the Parkinson's Disease Biomarkers Program (PDBP) Consortium, part of the National Institute of Neurological Disorders and Stroke at the National Institutes of Health. Investigators include: Roger Albin, Roy Alcalay, Alberto Ascherio, DuBois Bowman, Alice Chen-Plotkin, Ted Dawson, Richard Dewey, Dwight German, Xuemei Huang, Rachel Saunders-Pullman, Liana Rosenthal, Clemens Scherzer, David Vaillancourt, Vladislav Petyuk, Andy West and Jing Zhang. The PDBP Investigators have not participated in reviewing the data analysis or content of the manuscript. DNA panels from the NINDS Human Genetics Resource Center DNA and Cell Line Repository (<http://ccr.coriell.org/ninds>) were used in this study, as well as clinical data. The Newcastle Brain Tissue Resource, which is funded in part by a grant from the UK Medical Research Council and the Brains for Dementia research, a joint venture between Alzheimer's Society and Alzheimer's Research UK. The research was also partly funded by the National Institute for Health Research (NIHR) Newcastle Biomedical Research Centre based at Newcastle Hospitals NHS Foundation Trust and Newcastle University. We thank the following brain banks for providing brain tissues: Banner Sun Health Research Institute, New York Brain Bank, Newcastle Brain Tissue Resource, University of Michigan Brain Bank, University of California San Diego Brain Bank, Duke University Brain Bank, Virginia Commonwealth University Brain Bank, and the Georgetown University Brain Bank. We would like to thank the NIA Baltimore Longitudinal Study of Aging for contributing tissue samples to the Johns Hopkins Alzheimer's Disease Research Center. We are grateful to members of the North American Brain Expression Consortium (NABEC) for contributing DNA samples. This research has been conducted using the UK Biobank Resource under Application Number 33601. We would like to thank the NIH Neuro Brain Bank (<https://neurobiobank.nih.gov>) for contributing tissue samples. Tissue samples for genotyping were provided by the Johns Hopkins Morris K. Udall Center of Excellence for Parkinson's Disease Research (NIH P50 NS38377) and the Johns Hopkins Alzheimer's Disease Research Center. Members of the

23andMe Research Team are provided in the Supplementary material.

Funding

For a complete overview of members, acknowledgements and funding, please see the Supplemental material and/or <http://pdgenetics.org/partners>. This work was supported in part by the Intramural Research Programs of the National Institute of Neurological Disorders and Stroke (NINDS), the National Institute on Aging (NIA), and the National Institute of Environmental Health Sciences both part of the National Institutes of Health, Department of Health and Human Services; project numbers 1ZIA-NS003154, Z01-AG000949-02 and Z01-ES101986. This work was also financially supported by the Michael J. Fox Foundation, the Canadian Consortium on Neurodegeneration in Aging (CCNA), the Canada First Research Excellence Fund (CFREF), awarded to McGill University for the Healthy Brains for Healthy Lives (HBHL) program. The Columbia University cohort is supported by the Parkinson's Foundation, the National Institutes of Health [K02NS080915, and UL1 TR000040], and the Brookdale Foundation. PPMI, a public-private partnership, is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including AbbVie, Avid, Biogen, Bristol-Myers Squibb, Covance, GE Healthcare, Genentech, GlaxoSmithKline, Lilly, Lundbeck, Merck, Meso Scale Discovery, Pfizer, Piramal, Roche, Servier, Teva, UCB, and Golub Capital. This study was supported in part by grants from the National Institutes of Health: U19-AG03365, P50 NS38377, and P50-AG005146. We are grateful for the support of the entire BIOCARD study team at Johns Hopkins University. Additionally, we acknowledge the contributions of the Geriatric Psychiatry Branch (GPB) in the intramural program of NIMH who initiated the BIOCARD study.

Competing interests

M.A.N. reported receiving support from a consulting contract between Data Tecnica International and the National Institute on Aging (NIA), National Institutes of Health (NIH), and consulting for the Michael J. Fox Foundation, Vivid Genomics, Lysosomal Therapeutics Inc., and Neuron23, Inc, among others. No other disclosures were reported. Z.G.-O. has received consultancy fees from Lysosomal Therapeutics Inc., Idorsia, Denali, Prevail Therapeutics and Inception Sciences. K.H., P.C., and members of the 23andMe Research Team are employees of 23andMe, Inc. and hold stock or stock options in 23andMe.

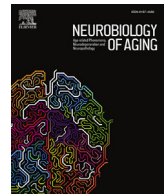
Supplementary material

Supplementary material is available at *Brain* online.

References

- Abraham G, Qiu Y, Inouye M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* 2017; 33: 2776–8.
- Alcalay RN, Levy OA, Waters CC, Fahn S, Ford B, Kuo S-H, et al. Glucocerebrosidase activity in Parkinson's disease with and without GBA mutations. *Brain* 2015; 138: 2648–58.
- Anheim M, Elbaz A, Lesage S, Durr A, Condroyer C, Viallet F, et al. Penetrance of Parkinson disease in glucocerebrosidase gene mutation carriers. *Neurology* 2012; 78: 417–20.
- Blauwendraat C, Bras JM, Nalls MA, Lewis PA, Hernandez DG, Singleton AB, et al. Coding variation in GBA explains the majority of the SYT11-GBA Parkinson's disease GWAS locus. *Mov Disord* 2018a; 33: 1821–3.
- Blauwendraat C, Faghri F, Pihlstrom L, Geiger JT, Elbaz A, Lesage S, et al. NeuroChip, an updated version of the NeuroX genotyping platform to rapidly screen for variants associated with neurological diseases. *Neurobiol Aging* 2017; 57: 247.e9–13.
- Blauwendraat C, Heilbron K, Vallerga CL, Bandres-Ciga S, von Coelln R, Pihlström L, et al. Parkinson's disease age at onset genome-wide association study: Defining heritability, genetic loci, and alpha-synuclein mechanisms. *Mov Disord* 2019; 34: 866–75.
- Blauwendraat C, Reed X, Kia DA, Gan-Or Z, Lesage S, Pihlstrom L, et al. Frequency of Loss of function variants in LRRK2 in Parkinson disease. *JAMA Neurol* 2018b; 75: 1416–22.
- Burkhardt MF, Martinez FJ, Wright S, Ramos C, Volfson D, Mason M, et al. A cellular model for sporadic ALS using patient-derived induced pluripotent stem cells. *Mol Cell Neurosci* 2013; 56: 355–64.
- Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018; 36: 411–20.
- Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018; 562: 203–9.
- Chan SJ, San Segundo B, McCormick MB, Steiner DF. Nucleotide and predicted amino acid sequences of cloned human and mouse preprocathepsin B cDNAs. *Proc Natl Acad Sci U S A* 1986; 83: 7721–5.
- Chen S, Zhou C, Yu H, Tao L, An Y, Zhang X, et al. 27-Hydroxycholesterol contributes to lysosomal membrane permeabilization-mediated pyroptosis in co-cultured SH-SY5Y cells and C6 cells. *Front Mol Neurosci* 2019; 12: 14.
- Consortium GT. The genotype-tissue expression (GTEx) project. *Nat Genet* 2013; 45: 580–5.
- Gan-Or Z, Amshalom I, Bar-Shira A, Gana-Weisz M, Mirelman A, Marder K, et al. The Alzheimer disease BIN1 locus as a modifier of GBA-associated Parkinson disease. *J Neurol* 2015a; 262: 2443–7.
- Gan-Or Z, Amshalom I, Kilarski LL, Bar-Shira A, Gana-Weisz M, Mirelman A, et al. Differential effects of severe vs mild GBA mutations on Parkinson disease. *Neurology* 2015b; 84: 880–7.
- Gan-Or Z, Dion PA, Rouleau GA. Genetic perspective on the role of the autophagy-lysosome pathway in Parkinson disease. *Autophagy* 2015c; 11: 1443–57.
- Gan-Or Z, Giladi N, Rozovski U, Shifrin C, Rosner S, Gurevich T, et al. Genotype-phenotype correlations between GBA mutations and Parkinson disease risk and onset. *Neurology* 2008; 70: 2277–83.
- Gan-Or Z, Liong C, Alcalay RN. GBA-associated Parkinson's disease and other synucleinopathies. *Curr Neurol Neurosci Rep* 2018; 18: 44.
- Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai SL, et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet* 2010; 6: e1000952.
- Guerreiro R, Ross OA, Kun-Rodrigues C, Hernandez DG, Orme T, Eicher JD, et al. Investigating the genetic architecture of dementia with Lewy bodies: a two-stage genome-wide association study. *Lancet Neurol* 2018; 17: 64–74.
- Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat Methods* 2017; 14: 955–8.
- Jinn S, Blauwendraat C, Toolan D, Gretzula CA, Drolet RE, Smith S, et al. Functionalization of the TMEM175 p.M393T variant as a risk factor for Parkinson disease. *Hum Mol Genet* 2019. doi: 10.1093/hmg/ddz136.
- Jinn S, Drolet RE, Cramer PE, Wong AH-K, Toolan DM, Gretzula CA, et al. TMEM175 deficiency impairs lysosomal and mitochondrial function and increases α -synuclein aggregation. *Proc Natl Acad Sci U S A* 2017; 114: 2389–94.
- Leija-Salazar M, Sedlazeck FJ, Toffoli M, Mullin S, Mokretar K, Athanasopoulou M, et al. Evaluation of the detection of GBA missense mutations and other variants using the Oxford Nanopore MinION. *Mol Genet Genomic Med* 2019; 7: e564.
- Leonard H, Blauwendraat C, Krohn L, Faghri F, Iwaki H, Furgeson G, et al. Genetic variability and potential effects on clinical trial outcomes: perspectives in Parkinson's disease. *bioRxiv* 2018. doi: 10.1101/427385.
- Lesage S, Anheim M, Condroyer C, Pollak P, Durif F, Dupuits C, et al. Large-scale screening of the Gaucher's disease-related glucocerebrosidase gene in Europeans with Parkinson's disease. *Hum Mol Genet* 2011; 20: 202–10.
- Li YI, Wong G, Humphrey J, Raj T. Prioritizing Parkinson's disease genes using population-scale transcriptomic data. *Nat Commun* 2019; 10: 994.
- Liu B, Gloudemans MJ, Rao AS, Ingelsson E, Montgomery SB. Abundant associations with gene expression complicate GWAS follow-up. *Nat Genet* 2019; 51: 768–9.
- Liu F, Li X, Lu C, Bai A, Bielawski J, Bielawska A, et al. Ceramide activates lysosomal cathepsin B and cathepsin D to attenuate autophagy and induces ER stress to suppress myeloid-derived suppressor cells. *Oncotarget* 2016; 7: 83907–25.
- Malek N, Weil RS, Bresner C, Lawton MA, Grosset KA, Tan M, et al. Features of -associated Parkinson's disease at presentation in the UK study. *J Neurol Neurosurg Psychiatry* 2018; 89: 702–9.
- McGlinchey RP, Lee JC. Cysteine cathepsins are essential in lysosomal degradation of α -synuclein. *Proc Natl Acad Sci U S A* 2015; 112: 9322–7.
- McKeith IG. Consensus guidelines for the clinical and pathologic diagnosis of dementia with Lewy bodies (DLB): report of the Consortium on DLB International Workshop. *J Alzheimers Dis* 2006; 9: 417–23.
- Nalls MA, Blauwendraat C, Vallerga CL, Heilbron K, Bandres-Ciga S, Chang D, et al. Parkinson's disease genetics: identifying novel risk loci, providing causal insights and improving estimates of heritable risk. *bioRxiv* 2018. doi: 10.1101/388165.
- Nalls MA, Bras J, Hernandez DG, Keller MF, Majounie E, Renton AE, et al. NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases. *Neurobiol Aging* 2015a; 36: 1605.e7–12.
- Nalls MA, Escott-Price V, Williams NM, Lubbe S, Keller MF, Morris HR, et al. Genetic risk and age in Parkinson's disease: continuum not stratum. *Mov Disord* 2015b; 30: 850–4.
- Nalls MA, McLean CY, Rick J, Eberly S, Hutten SJ, Gwinn K, et al. Diagnosis of Parkinson's disease on the basis of clinical and genetic classification: a population-based modelling study. *Lancet Neurol* 2015c; 14: 1002–9.
- Pihlström L, Blauwendraat C, Cappelletti C, Berge-Seidl V, Langmyhr M, Henriksen SP, et al. A comprehensive analysis of SNCA -related genetic risk in sporadic Parkinson disease. *Ann Neurol* 2018; 84: 117–29.
- Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 2010; 26: 2336–7.
- Qi T, Wu Y, Zeng J, Zhang F, Xue A, Jiang L, et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat Commun* 2018; 9: 2282.

- Rana HQ, Balwani M, Bier L, Alcalay RN. Age-specific Parkinson disease risk in GBA mutation carriers: information for genetic counseling. *Genet Med* 2013; 15: 146–9.
- Ritonja A, Popovic T, Turk V, Wiedenmann K, Machleidt W. Amino acid sequence of human liver cathepsin B. *FEBS Lett* 1985; 181: 169–72.
- Rivas MA, Avila BE, Koskela J, Huang H, Stevens C, Pirinen M, et al. Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population. *PLoS Genet* 2018; 14: e1007329.
- Robak LA, Jansen IE, van Rooij J, Uitterlinden AG, Kraaij R, Jankovic J, et al. Excessive burden of lysosomal storage disorder gene variants in Parkinson's disease. *Brain* 2017; 140: 3191–203.
- Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S to analyze and compare ROC curves. *BMC Bioinform* 2011; 12: 77.
- Ruskey JA, Greenbaum L, Roncière L, Alam A, Spiegelman D, Liong C, et al. Increased yield of full GBA sequencing in Ashkenazi Jews with Parkinson's disease. *Eur J Med Genet* 2019; 62: 65–9.
- Sidransky E, Nalls MA, Aasly JO, Aharon-Peretz J, Annesi G, Barbosa ER, et al. Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. *N Engl J Med* 2009; 361: 1651–61.
- Soldner F, Stelzer Y, Shivalila CS, Abraham BJ, Latourelle JC, Barrasa MI, et al. Parkinson-associated risk variant in distal enhancer of α -synuclein modulates target gene expression. *Nature* 2016; 533: 95–9.
- Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015; 12: e1001779.
- Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw* 2010; 36: 1–48.
- Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010; 38: e164.
- Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; 26: 2190–1.
- Yahalom G, Greenbaum L, Israeli-Korn S, Fay-Karmon T, Livneh V, Ruskey JA, et al. Carriers of both GBA and LRRK2 mutations, compared to carriers of either, in Parkinson's disease: Risk estimates and genotype-phenotype correlations. *Parkinsonism Relat Disord* 2019; 62: 179–84.
- Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ATC, Replication DIG, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* 2012; 44: 369–75, S1–3.
- Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011; 88: 76–82.
- Ysselstein D, Shulman JM, Krainc D. Emerging links between pediatric lysosomal storage diseases and adult parkinsonism. *Mov Disord* 2019; 34: 614–24.
- Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* 2016; 32: 1423–6.



Replication assessment of *NUS1* variants in Parkinson's disease



Bernabe I. Bustos^{a,**}, Sara Bandres-Ciga^b, J. Raphael Gibbs^c, Dimitri Krainc^a, Niccolo E. Mencacci^a, Ziv Gan-Or^{d,e,f}, Steven J. Lubbe^{a,*}, for the International Parkinson's Disease Genomics Consortium (IPDGC)¹

^a Ken and Ruth Davee Department of Neurology and Simpson Querrey Center for Neurogenetics, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

^b Molecular Genetics Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA

^c Computational Biology Group, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA

^d Department of Human Genetics, McGill University, Montréal, QC, Canada

^e Montreal Neurological Institute, McGill University, Montréal, QC, Canada

^f Department of Neurology and Neurosurgery, McGill University, Montréal, QC, Canada

ARTICLE INFO

Article history:

Received 27 July 2020

Received in revised form 5 October 2020

Accepted 7 November 2020

Available online 13 November 2020

Keywords:

NUS1

Rare-variant burden

Parkinson's disease

ABSTRACT

The *NUS1* gene was recently associated with Parkinson's disease (PD) in the Chinese population. Here, as part of the International Parkinson's Disease Genomics Consortium, we have leveraged large-scale PD case-control cohorts to comprehensively assess damaging *NUS1* variants in individuals of European descent. Burden analysis of rare nonsynonymous damaging variants across case-control individuals from whole-exome and -genome data sets did not find evidence of *NUS1* association with PD. Overall, single-variant tests for rare (minor allele frequency < 0.01) and common (minor allele frequency > 0.01) variants, including 15 PD-GWAS cohorts and summary statistics from the largest PD GWAS meta-analysis to date, also did not uncover any associations. Our results indicate a lack of evidence for a role of rare damaging nonsynonymous *NUS1* variants in PD in unrelated case-control cohorts of European descent, suggesting that the previously observed association could be driven by extremely rare population-specific variants.

© 2020 Elsevier Inc. All rights reserved.

1. Introduction

The *NUS1* gene encodes an endoplasmic dehydrodolichyl diphosphate synthase subunit that promotes trafficking of low-density lipoprotein-derived cholesterol (Harrison et al., 2009, 2011). A significant burden of rare (minor allele frequency [MAF] < 0.01) damaging nonsynonymous *NUS1* variants was recently identified in 2 large independent Chinese case-control cohorts (Guo et al., 2018). Replication in a smaller Chinese cohort found no association with Parkinson's disease (PD) (Chen et al., 2020). Here, as part of the International Parkinson's Disease Genomics Consortium

(IPDGC), we assessed a compendium of large-scale PD case-control cohorts to investigate damaging *NUS1* variants in PD populations of European descent.

2. Results

Using the IPDGC whole-exome sequencing (WES) data (Jansen et al., 2017) and the Accelerating Medicines Partnership–Parkinson's disease initiative (AMP-PD) whole-genome sequencing (WGS) data (<https://amp-pd.org/whole-genome-data>), we only identified 2 heterozygous rare (MAF < 0.01), damaging (CADD > 12.37) nonsynonymous *NUS1* variants: p.P169 R and p.K214 E (Table 1). None of the originally reported 18 variants were observed. We identified single heterozygous carriers of p.P169 R and p.K214 E in PD cases and none in controls in the IPDGC-WES data (1/1040 cases vs. 0/452 controls). In the AMP-PD WGS data, we observed 5 PD cases and 5 controls harboring p.P169 R (5/1647 cases vs. 5/1050 controls) and one case and control carrying p.K214 E (1/1647 vs. 1/1050 controls). Assuming a global disease prevalence of 0.5% and the originally reported carriership frequency of 0.003 (28/9512), we have a 100% power to detect an odds ratio of

* Corresponding author at: Ken and Ruth Davee Department of Neurology and Simpson Querrey Center for Neurogenetics, Northwestern University Feinberg School of Medicine, 320 East Superior Avenue, Morton 7-605, Chicago, IL 60611, USA. Tel.: (+1) 312-503-5298.

** Corresponding author at: Ken and Ruth Davee Department of Neurology and Simpson Querrey Center for Neurogenetics, Northwestern University Feinberg School of Medicine, 303 E Superior St, Chicago, IL 60611, USA. Tel.: (+1) 312-503-5297.

E-mail addresses: bernabe.bustos@northwestern.edu (B.I. Bustos), steven.lubbe@northwestern.edu (S.J. Lubbe).

¹ The full list of IPDGC members is in pdgenetics.org/partners.

Table 1
Burden analysis of rare damaging nonsynonymous variants within the *NUS1* gene in the IPDGC-WES and AMP-PD WGS data sets

CHR	BP	REF	ALT	LOC	Exon	Type	Variant	CADD	dbSNP	gnomAD Eur	IPDGC-WES		AMP-PD WGS		Combined	
											Cases	Controls	Cases	Controls	Cases	Controls

Key: ALT, alternative allele; AMP-PD, Accelerating Medicines Partnership–Parkinson's disease initiative (<https://amp-pd.org/>); BP, base pair; CADD, Combined Annotation Dependent Depletion (<https://cadd.gs.washington.edu/>); CHR, chromosome; dbSNP, The Single Nucleotide Polymorphism Database (<https://www.ncbi.nlm.nih.gov/snp/>); Exon, exon number; Freq, frequency; gnomAD Eur, European (non-Finnish) frequency from the Genome aggregation consortium (<https://gnomad.broadinstitute.org/>); IPDGC, International Parkinson's Disease Genomics Consortium (<https://pdgenetics.org/>); LOC, variant location; nsn, nonsynonymous; REF, reference allele; WES, whole-exome sequencing; WGS, whole-genome sequencing.

6	118,014,295	C	G	exonic	2	nsn	p.P169 R	19.26	rs150646335	0.002	1	0	5	5	6	5
6	118,015,292	A	G	exonic	3	nsn	p.K214 E	18.02	rs146171115	0.0006	1	0	1	1	2	1

11.5 at $p = 0.049$, for both IPDGC WES and AMP-PD cohorts. No significant burden differences were observed between cases and controls (Table 1). Additional burden tests (gene- and allele-based) failed to uncover any significant enrichment (see [Supplementary Information](#)).

We next explored single-variant associations, using rare and common variants separately, within *NUS1* plus 10kb flanking regions. None of the 6 and 328 rare variants observed in the IPDGC-WES and AMP-PD WGS data, respectively, showed significant association (See [Supplementary Information](#)). Two rare intronic variants in the IPDGC-GWAS cohorts (total number of variants observed is 36) survived multiple testing correction: rs144827068 ($p=1.6 \times 10^{-3}$) and rs187218668 ($p=1.7 \times 10^{-3}$) ([Supplementary Table 1](#)). Although suggestive associations with common variants were observed in the AMP-PD WGS cohort (rs74498762: $p=1.07 \times 10^{-3}$, OR=0.40, SE=0.28; rs143797072: $p=1.51 \times 10^{-3}$, OR=0.63, SE=0.15; [Supplementary Table 2](#) and [Supplementary Fig. 1](#), upper panel), assessment of *NUS1* common variants in the IPDGC-GWAS data and summary statistics from the most recent PD GWAS meta-analysis (Nalls et al., 2019) failed to uncover conclusive associations ([Supplementary Fig. 1](#), middle and lower panel).

3. Conclusions

Here, we performed a comprehensive assessment of *NUS1* variants using the largest available PD genetic data sets of European descent. We did not observe any evidence of enrichment of rare damaging nonsynonymous variants in PD cases compared with controls. At the single-variant level, although we observed some suggestive associations, the largest meta-analysis of PD GWAS data (Nalls et al., 2019) did not uncover any associated *NUS1* variants. The lack of evidence here supports the previous failure to replicate in Chinese population (Chen et al., 2020); however, we note that it was somewhat statistically underpowered because of the small sample size, a similar limitation we faced in our rare single-variant tests. Although we cannot exclude a possible role for rare *NUS1* variants in the European population, the fact that most of the original variants are extremely rare and mostly present in Asian populations ([Supplementary Table 3](#)) suggests that they are likely population specific. Additional large-scale familial and case-control studies in non-European ancestry populations are needed to further evaluate the role of *NUS1* in PD etiology.

4. Methods

We used 4 large sets of available PD case-control cohorts of European ancestry (See [Supplementary Information](#)). Variant annotation and selection are described in the [Supplementary Information](#). For rare nonsynonymous variant burden tests, we leveraged carrier counts in cases and controls using Fisher's exact test. For single rare variant tests ($MAF < 0.01$), we assessed allelic, genotypic, trend, dominant, and recessive models with Fisher's exact test. For single common variant tests ($MAF > 0.01$), we used a logistic regression. Multiple testing was accounted for by Bonferroni correction. More detailed descriptions are available in the [Supplementary Information](#).

Disclosure

D.K. is the Founder and Scientific Advisory Board Chair of Lysosomal Therapeutics Inc. D.K. serves on the scientific advisory boards of The Silverstein Foundation, Intellia Therapeutics, and Prevail Therapeutics and is a Venture Partner at OrbiMed. B.I.B., S.B.

C, J.R.G, N.E.M, Z.G-O and S.J.L. declare that they do not have conflict of interests.

CRedit authorship contribution statement

Bernabe I. Bustos: Conceptualization, Methodology, Formal analysis, Resources, Writing - original draft, Writing - review & editing. **Sara Bandres-Ciga:** Resources, Writing - review & editing. **J. Raphael Gibbs:** Resources, Formal analysis, Writing - review & editing. **Dimitri Krainc:** Funding acquisition, Writing - review & editing. **Niccolo E. Mencacci:** Conceptualization, Writing - review & editing. **Ziv Gan-Or:** Conceptualization, Writing - review & editing. **Steven J. Lubbe:** Conceptualization, Supervision, Writing - original draft, Writing - review & editing.

Acknowledgment

The funding source of our work is coming from the Simpson Querrey Center for Neurogenetics to Dimitri Krainc.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <https://doi.org/10.1016/j.neurobiolaging.2020.11.007>.

References

- Chen, X., Xiao, Y., Zhou, M., Lin, Y., Guo, W., Huang, S., Qiu, J., Peng, G., Mo, M., Li, Z., Zhu, X., Xu, P., 2020. Genetic analysis of NUS1 in Chinese patients with Parkinson's disease. *Neurobiol. Aging* 86, 202 e205–202 e206.
- Guo, J.F., Zhang, L., Li, K., Mei, J.P., Xue, J., Chen, J., Tang, X., Shen, L., Jiang, H., Chen, C., Guo, H., Wu, X.L., Sun, S.L., Xu, Q., Sun, Q.Y., Chan, P., Shang, H.F., Wang, T., Zhao, G.H., Liu, J.Y., Xie, X.F., Jiang, Y.Q., Liu, Z.H., Zhao, Y.W., Zhu, Z.B., Li, J.D., Hu, Z.M., Yan, X.X., Fang, X.D., Wang, G.H., Zhang, F.Y., Xia, K., Liu, C.Y., Zhu, X.W., Yue, Z.Y., Li, S.C., Cai, H.B., Zhang, Z.H., Duan, R.H., Tang, B.S., 2018. Coding mutations in NUS1 contribute to Parkinson's disease. *Proc. Natl. Acad. Sci. U. S. A.* 115, 11567–11572.
- Harrison, K.D., Miao, R.Q., Fernandez-Hernando, C., Suarez, Y., Davalos, A., Sessa, W.C., 2009. Nogo-B receptor stabilizes Niemann-Pick type C2 protein and regulates intracellular cholesterol trafficking. *Cell Metab.* 10, 208–218.
- Harrison, K.D., Park, E.J., Gao, N., Kuo, A., Rush, J.S., Waechter, C.J., Lehrman, M.A., Sessa, W.C., 2011. Nogo-B receptor is necessary for cellular dolichol biosynthesis and protein N-glycosylation. *EMBO J.* 30, 2490–2500.
- Jansen, I.E., Gibbs, J.R., Nalls, M.A., Price, T.R., Lubbe, S., van Rooij, J., Uitterlinden, A.G., Kraaij, R., Williams, N.M., Brice, A., Hardy, J., Wood, N.W., Morris, H.R., Gasser, T., Singleton, A.B., Heutink, P., Sharma, M., International Parkinson's Disease Genomics, C., 2017. Establishing the role of rare coding variants in known Parkinson's disease risk loci. *Neurobiol. Aging* 59, 220.e211–220.e218.
- Nalls, M.A., Blauwendraat, C., Vallerga, C.L., Heilbron, K., Bandres-Ciga, S., Chang, D., Tan, M., Kia, D.A., Noyce, A.J., Xue, A., Bras, J., Young, E., von Coelln, R., Simon-Sanchez, J., Schulte, C., Sharma, M., Krohn, L., Pihlstrom, L., Siitonen, A., Iwaki, H., Leonard, H., Faghri, F., Gibbs, J.R., Hernandez, D.G., Scholz, S.W., Botia, J.A., Martinez, M., Corvol, J.C., Lesage, S., Jankovic, J., Shulman, L.M., Sutherland, M., Tienari, P., Majamaa, K., Toft, M., Andreassen, O.A., Bangale, T., Brice, A., Yang, J., Gan-Or, Z., Gasser, T., Heutink, P., Shulman, J.M., Wood, N.W., Hinds, D.A., Hardy, J.A., Morris, H.R., Gratten, J., Visscher, P.M., Graham, R.R., Singleton, A.B., Me Research, T., System Genomics of Parkinson's Disease, C., International Parkinson's Disease Genomics, C., 2019. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* 18, 1091–1102.

Acta Neuropathol

. 2020 Sep;140(3):401-402.

doi: 10.1007/s00401-020-02172-4. Epub 2020 Jun 15.

ATP10B and the risk for Parkinson's disease

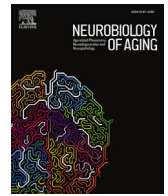
[Raquel Real](#)^{1,2}, [Anni Moore](#)³, [Cornelis Blauwendraat](#)³, [Huw R Morris](#)^{1,2}, [Sara Bandres-Ciga](#)⁴, [International Parkinson's Disease Genomics Consortium \(IPDGC\)](#)

Affiliations collapse

Affiliations

- ¹Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, London, WC1N 3BG, UK.
- ²UCL Movement Disorders Centre, UCL Queen Square Institute of Neurology, London, WC1N 3BG, UK.
- ³Molecular Genetics Section, Laboratory of Neurogenetics, National Institute On Aging, National Institutes of Health, Bethesda, MD, 20892, USA.
- ⁴Molecular Genetics Section, Laboratory of Neurogenetics, National Institute On Aging, National Institutes of Health, Bethesda, MD, 20892, USA. sara.bandresciga@nih.gov.
- PMID: 32556962
- PMCID: PMC7540943 (available on 2021-09-01)
- DOI: 10.1007/s00401-020-02172-4

ABSTRACT NOT AVAILABLE



Assessment of *LIN28A* variants in Parkinson's disease in large European cohorts

Monica Diez-Fairen^{a,1}, Mary B. Makarious^{b,1}, Sara Bandres-Ciga^b, Cornelis Blauwendraat^{b,*}, on behalf of the International Parkinson's Disease Genomics Consortium (IPDGC)

^a Department of Neurology, Fundació Docència i Recerca MútuaTerrassa and Movement Disorders Unit, University Hospital MútuaTerrassa, Terrassa, Barcelona, Spain

^b Molecular Genetics Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA



ARTICLE INFO

Article history:

Received 13 July 2020

Received in revised form 2 December 2020

Accepted 2 December 2020

Available online 11 December 2020

Keywords:

Parkinson's disease

Risk factor

LIN28A

Loss of function

ABSTRACT

Parkinson's disease (PD) is a complex neurodegenerative disease with a strong genetic component. To date, several genes have been associated with monogenic forms of the disease, but these only explain a small fraction of the observed familial aggregation in PD. Recently, a heterozygous loss-of-function variant in *LIN28A* was associated with PD pathogenesis in the Asian population. Here, we comprehensively investigate the role of *LIN28A* variants in PD patients of European ancestry and assess susceptibility using individual-level genotyping data from 14,671 PD cases and 17,667 controls, as well as whole-genome sequencing data from 1647 patients with PD and 1050 controls. In addition, we further assess the summary statistics from the most recent genome-wide association studies meta-analyses to date for PD risk and age at onset. After evaluating these data, we did not find evidence to support a role for *LIN28A* as a major causal gene for PD. However, additional large-scale familial and case-control studies in non-European ancestry populations are necessary to further evaluate the role of *LIN28A* in PD etiology.

Published by Elsevier Inc.

1. Introduction

Parkinson's disease (PD) is a complex neurodegenerative disorder with a strong genetic component in which both rare and common genetic variants contribute to disease risk, onset, and progression. Mutations in several genes have been associated with familial forms of disease, most of which are highly penetrant rare variants resulting in early onset presentation. Although significant progress in understanding the genetic basis of PD has been made, validation of novel genes associated with PD remains extremely challenging (Bandres-Ciga et al., 2020; Blauwendraat et al., 2020).

We read with great interest the recently published article by Chang et al. in which the authors report and claim that a rare variant in *LIN28A*, which encodes for a highly conserved RNA-binding protein (Lin28a), contributes to PD pathogenesis (Chang et al., 2019). Through whole-exome sequencing, the authors identified a

heterozygous missense variant in *LIN28A* (p.Arg192Gly, rs558060339), shown to result in gene loss-of-function (LOF), in 2 patients of East Asian ancestry with early-onset PD (EOPD). In addition, midbrain dopamine neurons differentiated from patient-derived human embryonic stem cells showed developmental defects and PD-specific pathological features that could be rescued by expression of wild-type Lin28a. Indeed, previous work from the same group had shown that loss of *LIN28A* function could ultimately lead to a PD-like phenotype coupled with midbrain dopamine neuronal loss (Rhee et al., 2016).

There are some known similarities and differences in the genetic contributions of PD on a population genetics scale. An example of this is the identification of the p.Arg1628Pro and p.Gly2385Arg *LRRK2* mutations found only in Asian-specific cohorts, whereas European cohorts identified the p.Gly2019Ser *LRRK2* variant. Studies looking at the genetic risk of PD in European populations identified the *MAPT* H2 haplotype as protective, which was absent in Asian populations (Foo et al., 2020). Here, we sought to investigate the role of *LIN28A* variants in PD susceptibility in Europeans as part of the International Parkinson's Disease Genomics Consortium (IPDGC) efforts to screen reported risk factors for PD.

* Corresponding author at: Cornelis Blauwendraat, Porter Neuroscience Center, 35 Convent Drive, Bethesda, MD 20892, USA. Tel.: +1 301 451 6079; fax: +1 301 451 5466.

E-mail address: cornelis.blauwendraat@nih.gov (C. Blauwendraat).

¹ These authors contributed equally to this work.

2. Methods

2.1. Cohort description

We used publicly available whole-genome sequencing (WGS) data including 1647 patients with PD without known disease-causing mutations (mean age at onset (AAO) 64.2 ± 9.6) and 1050 controls (mean age 60.3 ± 11.9) from the Accelerating Medicines Partnership–Parkinson's disease initiative (www.amp-pd.org). All individuals were of white European descent and were obtained from 3 different cohorts (Biofind (<https://biofind.ionis.usc.edu/>), Parkinson's Disease Biomarker Program (<https://pdpb.ninds.nih.gov/>), and Parkinson's Progression Markers Initiative (<https://www.ppmi-info.org/>)) included in the Accelerating Medicines Partnership–Parkinson's disease initiative resource. Among the PD cases, 145 (8.8%) had EOPD (mean AAO 45.2 ± 4.6) and 527 (32%) had a family history. More detailed cohort characteristics as well as quality control procedures are further described in <https://amp-pd.org/whole-genome-data>.

In addition, we analyzed individual-level genotyping data from genome-wide association studies (GWAS) from the IPDGC consisting of 14,671 PD cases and 17,667 neurologically healthy individuals of white European ancestry. Quality control procedures on both individual and variant levels were performed before imputation as previously described (Nalls et al., 2019; Blauwendraat et al., 2019). We further assessed data from the most recent GWAS meta-analyses to date (excluding 23andMe data) for PD risk consisting of 15,056 PD cases, 18,618 UK Biobank proxy-cases (i.e., participants with a first degree relative with PD) and 449,056 controls (Nalls et al., 2019), as well as for AAO comprising 17,996 PD cases (Blauwendraat et al., 2019). All participants were of European ancestry, and details on genotyping quality and imputation are described in the original sources (Blauwendraat et al., 2019; Nalls et al., 2019). Summary statistics used in this study are public and available via <https://pdgenetics.org>.

For both sequencing and genotyping data sets, the entire *LIN28A* region (NM_024674) was annotated using ANNOVAR (Wang et al., 2010) and variant allele frequencies were determined using PLINK v1.9 (Chang et al., 2015). Additional allele frequencies were obtained from the Genome Aggregation Database (gnomAD v2.1.1; <https://gnomad.broadinstitute.org/>) (Lek et al., 2016).

2.2. Statistical analyses

Association analyses of *LIN28A* variants and risk for PD were carried out using single-variant score tests in RVTESTS package v.2.1.0 (Zhan et al., 2016) for sequencing data and logistic regression for genotyping data. All analyses were adjusted by sex, age, family history, education level, 10 principal components (PCs) to account for population stratification, and cohort to control for chip bias.

To assess the cumulative effect of multiple rare variants on the risk for PD, we performed gene-based burden analyses (SKAT, sequence Kernel association tests; and SKAT-O, optimized SKAT) on WGS data. We used the RVTESTS package v.2.1.0 following default parameters (Zhan et al., 2016) and adjusted for covariates age, gender, 10 PCs and data set to control for chip bias. All code used can be found here <https://github.com/ipdgc/IPDGC-Trainees/blob/master/LIN28A.md>.

Table 1
LIN28A coding variants in whole-genome sequencing data from PD and controls

Variant	rsID	Frequency (gnomAD) ^a	Frequency (cases)	Frequency (controls)	OR	95% CI	p-value
p.Lys127Lys	rs138964629	0.0008	0.0009107	0	NA	NA	NA
p.Thr189Ile	rs201124162	0.0007	0.0009107	0.0004762	2.26	–1.70 to 3.33	0.524
p.Pro205Pro	rs201055562	0.0009	0.0006072	0.0009524	0.45	–3.21 to 1.63	0.521

Key: CI, confidence interval; OR, odds ratio; PD, Parkinson's disease.

^a Allele frequencies in accordance with non-Finnish European gnomAD exomes.

3. Results

We identified a total of 83 variants within the *LIN28A* gene in WGS data, of which only 3 were coding, including 2 synonymous (p.Lys127Lys and p.Pro205Pro) and 1 nonsynonymous variant (p.Thr189Ile) (Table 1). All 3 coding variants are extremely rare, with a minor allele frequency (MAF) $\leq 1\%$ in accordance with public databases. Case-control association testing using a single-variant score test adjusted by covariates (including sex, age, family history, education level, 10 PCs, and cohort) showed no significant differences in the frequency of *LIN28A* variants between PD cases and controls after Bonferroni correction (significance threshold = $0.05/83 = 6.02 \times 10^{-4}$) (Table 1, Supplementary Table 1). Gene-based burden analyses did not detect an enrichment of rare variants in PD cases versus controls (Supplementary Table 2).

Similarly, we performed a case-control association analysis using logistic regression in imputed individual-level genotyping data from IPDGC. All 34 *LIN28A* variants found were noncoding, including 28 intronic and 6 in the 3'-untranslated region (3'-UTR). None of the variants were significantly associated with the disease (Supplementary Table 3). We further analyzed summary statistics from the abovementioned risk and AAO PD GWAS meta-analyses. No evidence for an association between *LIN28A* common genetic variation and PD was detected for either PD risk or AAO (Supplementary Fig. 1; Supplementary Table 4). Among the 97 *LIN28A* variants identified within summary statistics, 81 were intronic, 14 were in the 3'-UTR, and only 2 were coding nonsynonymous variants (p.Arg123Gln and p.Thr189Ile) (Supplementary Table 4). A total of 46 variants (47.4%) were rare (MAF $\leq 3\%$). Of note, we did not find the original reported mutation (NC_000001.10:g.26752893 C>G, p.Arg192Gly) by Chang et al. in any of our data sets.

We also explored the frequency of LOF variants in *LIN28A* by looking at the gnomAD database v.2.1.1 (<https://gnomad.broadinstitute.org/>), which comprises a total of 125,748 exomes and 15,708 genomes. Six LOF variants were identified (4 stop-gain and 2 frameshift, all heterozygous), resulting in an estimate of 0.0056% of the population having haploinsufficiency for *LIN28A*. The reported p.Arg192Gly variant was identified in 6 Korean individuals out of a total of 1909 included, setting the allele frequency at 0.15% in the general population.

4. Discussion

In conclusion, based on the current data presented, our analyses do not support a role for *LIN28A* variants as a major causal gene or risk factor for PD. However, the vast majority of our data are of European ancestry, in which the p.Arg192Gly variant is not present in accordance with the gnomAD database, and we have a relatively small number of EOPD cases in the cohorts used. Noteworthy, the frequency of this variant is relatively high in the Korean population (0.15%). Overall, we cannot rule out a potential role of very rare variants in *LIN28A* in PD among other populations. Additional large-scale familial and case-control studies in non-European ancestry populations are necessary to further evaluate the role of *LIN28A* in PD etiology.

Disclosure statement

The authors declare that they have no conflict of interest.

CRedit authorship contribution statement

Monica Diez-Fairen: Conceptualization, Formal analysis, Writing - original draft. **Mary B. Makarious:** Conceptualization, Formal analysis, Data curation, Writing - original draft. **Sara Bandres-Ciga:** Conceptualization, Formal analysis, Data curation, Writing - review & editing. **Cornelis Blauwendraat:** Conceptualization, Supervision, Writing - review & editing.

Acknowledgements

We would like to thank all the participants who donated their time and biological samples to be a part of this study. We also would like to thank all members of the International Parkinson's Disease Genomics Consortium (IPDGC). For a complete overview of members, acknowledgements, and funding, please see <http://pdgenetics.org/partners>. This work was supported in part by the Intramural Research Programs of the National Institute of Neurological Disorders and Stroke (NINDS), the National Institute on Aging (NIA), and the National Institute of Environmental Health Sciences both part of the National Institutes of Health, Department of Health and Human Services: project numbers 1ZIA-NS003154, Z01-AG000949-02 and Z01-ES101986. In addition, this work was supported by the Department of Defense (award W81XWH-09-2-0128) and The Michael J Fox Foundation for Parkinson's Research. 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). This work used the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.neurobiolaging.2020.12.002>.

References

Bandres-Ciga, S., Diez-Fairen, M., Kim, J.J., Singleton, A.B., 2020. Genetics of Parkinson's disease: an introspection of its journey towards precision medicine. *Neurobiol. Dis.* 137, 104782.

- Blauwendraat, C., Heilbron, K., Vallerga, C.L., Bandres-Ciga, S., von Coelln, R., Pihlström, L., Simón-Sánchez, J., Schulte, C., Sharma, M., Krohn, L., Siitonen, A., Iwaki, H., Leonard, H., Noyce, A.J., Tan, M., Gibbs, J.R., Hernandez, D.G., Scholz, S.W., Jankovic, J., Shulman, L.M., Lesage, S., Corvol, J.-C., Brice, A., van Hilten, J.J., Marinus, J., 23andMe Research Team, Eerola-Rautio, J., Tienari, P., Majamaa, K., Toft, M., Grosset, D.G., Gasser, T., Heutink, P., Shulman, J.M., Wood, N., Hardy, J., Morris, H.R., Hinds, D.A., Gratten, J., Visscher, P.M., Gan-Or, Z., Nalls, M.A., Singleton, A.B., International Parkinson's Disease Genomics Consortium (IPDGC), 2019. Parkinson's disease age at onset genome-wide association study: defining heritability, genetic loci, and α -synuclein mechanisms. *Mov. Disord.* 34, 866–875.
- Blauwendraat, C., Nalls, M.A., Singleton, A.B., 2020. The genetic architecture of Parkinson's disease. *Lancet Neurol.* 19, 170–178.
- Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Giga-science* 4, 7.
- Chang, M.-Y., Oh, B., Choi, J.-E., Sulistio, Y.A., Woo, H.-J., Jo, A., Kim, J., Kim, E.-H., Kim, S.W., Hwang, J., Park, J., Song, J.-J., Kwon, O.-C., Henry Kim, H., Kim, Y.-H., Ko, J.Y., Heo, J.Y., Lee, M.J., Lee, M., Choi, M., Chung, S.J., Lee, H.-S., Lee, S.-H., 2019. LIN28A loss of function is associated with Parkinson's disease pathogenesis. *EMBO J.* 38, e101196.
- Foo, J.N., Chew, E.G.Y., Chung, S.J., Peng, R., Blauwendraat, C., Nalls, M.A., Mok, K.Y., Satake, W., Toda, T., Chao, Y., Tan, L.C.S., Tandiono, M., Lian, M.M., Ng, E.Y., Prakash, K.M., Au, W.-L., Meah, W.-Y., Annuar, S.Q., Mok, Annuar, A.A., Chan, A.Y.Y., Chen, L., Chen, Y., Jeon, B.S., Jiang, L., Lim, J.L., Lin, J.-J., Liu, C., Mao, C., Mok, V., Pei, Z., Shang, H.-F., Shi, C.-H., Song, K., Tan, A.H., Wu, Y.-R., Xu, Y.-M., Xu, R., Yan, Y., Yang, J., Zhang, B., Koh, W.-P., Lim, S.-Y., Khor, C.C., Liu, J., Tan, E.-K., 2020. Identification of risk loci for Parkinson's disease in Asians and comparison of risk between Asians and Europeans: A Genome-Wide Association Study. *JAMA Neurol.* 77, 746–754.
- Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., Tukiainen, T., Birnbaum, D.P., Kosmicki, J.A., Duncan, L.E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D.N., DeFlaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M.I., Moonshine, A.L., Natarajan, P., Orozco, L., Peloso, G.M., Poplin, R., Rivas, M.A., Ruano-Rubio, V., Rose, S.A., Ruderfer, D.M., Shakir, K., Stenson, P.D., Stevens, C., Thomas, B.P., Tiao, G., Tusie-Luna, M.T., Weisburd, B., Won, H.-H., Yu, D., Altshuler, D.M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J.C., Gabriel, S.B., Getz, G., Glatt, S.J., Hultman, C.M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M.J., McGovern, D., McPherson, R., Neale, B.M., Palotie, A., Purcell, S.M., Saleheen, D., Scharf, J.M., Sklar, P., Sullivan, P.F., Tuomilehto, J., Tsuang, M.T., Watkins, H.C., Wilson, J.G., Daly, M.J., MacArthur, D.G., Exome Aggregation Consortium, 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291.
- Nalls, M.A., Blauwendraat, C., Vallerga, C.L., Heilbron, K., Bandres-Ciga, S., Chang, D., Tan, M., Kia, D.A., Noyce, A.J., Xue, A., Bras, J., Young, E., von Coelln, R., Simón-Sánchez, J., Schulte, C., Sharma, M., Krohn, L., Pihlström, L., Siitonen, A., Iwaki, H., Leonard, H., Faghri, F., Gibbs, J.R., Hernandez, D.G., Scholz, S.W., Botia, J.A., Martinez, M., Corvol, J.-C., Lesage, S., Jankovic, J., Shulman, L.M., Sutherland, M., Tienari, P., Majamaa, K., Toft, M., Andreassen, O.A., Bangale, T., Brice, A., Yang, J., Gan-Or, Z., Gasser, T., Heutink, P., Shulman, J.M., Wood, N.W., Hinds, D.A., Hardy, J.A., Morris, H.R., Gratten, J., Visscher, P.M., Graham, R.R., Singleton, A.B., 23andMe Research Team, System Genomics of Parkinson's Disease Consortium, International Parkinson's Disease Genomics Consortium, 2019. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* 18, 1091–1102.
- Rhee, Y.-H., Kim, T.-H., Young Jo, A., Chang, M.-Y., Park, C.-H., Kim, S.-M., Song, J.-J., Oh, S.-M., Yi, S.-H., Kim, H.H., You, B.-H., Nam, J.-W., Lee, S.-H., 2016. LIN28A enhances the therapeutic potential of cultured neural stem cells in a Parkinson's disease model. *Brain* 139 (Pt 10), 2722–2739.
- Wang, K., Li, M., Hakonarson, H., 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164.
- Zhan, X., Hu, Y., Li, B., Abecasis, G.R., Liu, D.J., 2016. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* 32, 1423–1426.



Negative results

Lack of evidence for association of *UQCRC1* with Parkinson's disease in EuropeansKonstantin Senkevich^{a,b,*}, Sara Bandres-Ciga^c, Ziv Gan-Or^{a,b,d}, Lynne Krohn^{a,d}, on behalf of the International Parkinson's Disease Genomics Consortium (IPDGC)^a Montreal Neurological Institute, McGill University, Montréal, Quebec, Canada^b Department of Neurology and neurosurgery, McGill University, Montréal, Quebec, Canada^c Molecular Genetics Section, Laboratory of Neurogenetics, NIA, NIH, Bethesda, MD, USA^d Department of Human Genetics, McGill University, Montréal, Quebec, Canada

ARTICLE INFO

Article history:

Received 4 September 2020

Received in revised form 27 October 2020

Accepted 28 October 2020

Available online 2 November 2020

Keywords:

Parkinson disease

Genetics

UQCRC1

ABSTRACT

Recently, a novel variant p.Y314S in *UQCRC1* has been implicated as pathogenic in Parkinson's disease (PD). In the present study, we aimed to examine the association of *UQCRC1* with PD in large cohorts of European origin. We examined common and rare genetic variation in *UQCRC1* using genome-wide association study data from the International Parkinson Disease Genomics Consortium, including 14,671 cases and 17,667 controls, and whole-genome sequencing data from the Accelerating Medicines Partnership–Parkinson's disease initiative, including 1647 patients with PD and 1050 controls. No common variants were consistently associated with PD, and a variety of burden analyses did not reveal an association between rare variants in *UQCRC1* and PD. Therefore, our results do not support a major role for *UQCRC1* in PD in the European population, and additional studies in other populations are warranted.

© 2020 Elsevier Inc. All rights reserved.

1. Introduction

The genetics of Parkinson's disease (PD) has been extensively studied over the last 20 years (Bandres-Ciga et al., 2020). A recent genome-wide association study (GWAS) identified 90 independent risk variants (Nalls et al., 2019); however, these variants explain less than 50% of the heritability of PD (Nalls et al., 2019), suggesting that other, unknown common and rare genetic variants affect the risk of PD. A novel variant p.Y314S in *UQCRC1* encoding the ubiquinol-cytochrome c reductase core protein (UQCRC1) has been recently identified in 5 Taiwanese family members with parkinsonism by whole-exome sequencing (Lin et al., 2019). UQCRC1 is a mitochondrial protein and part of the respiratory chain III complex (Hoffman et al., 1993), which may play a role in mitochondrial respiration (Lin et al., 2019; Shan et al., 2019).

The purpose of this work is to examine the role of *UQCRC1* in a large-scale European PD population using GWAS and whole-genome sequencing (WGS) data from the International Parkinson Disease Genomics Consortium (IPDGC) and Accelerating Medicines Partnership–Parkinson's disease (AMP-PD) initiative.

* Corresponding author at: Montreal Neurological Institute, McGill University 1033 Pine Avenue, West, Ludmer Pavilion, room 308-309 Montreal, Quebec, H3A 1A1, Canada. Tel.: +1-438-979-5052; fax: +1-514-398-8248.

E-mail address: Konstantin.senkevich@mcgill.ca (K. Senkevich).

2. Methods

The study populations included 14,671 patients with PD and 17,667 controls from IPDGC and 1647 patients with PD and 1050 controls from AMP-PD (<https://amp-pd.org/>). Quality control of IPDGC GWAS data was performed on both individual and variant levels as previously described (Nalls et al., 2019). Similar quality control procedures were performed in the AMP-PD WGS data, as described by AMP-PD (<https://amp-pd.org/whole-genome-data>). We extracted *UQCRC1* genotyping data from both data sets using gene coordinates determined by UQCRC1 position (+/− 100kb): hg19: chr3:48,536,432–48,747,098; hg38: chr3:48,499,002–48,709,646. We used ANNOVAR (Wang et al., 2010) to annotate both data sets. PLINK 1.9 (Chang et al., 2015) was used for logistic regression (adjusted for age, sex, and first 10 principal components) to study the association between common *UQCRC1* variants (with minor allele frequency [MAF] > 0.01) and PD in the IPDGC cohort. In the AMP-PD cohort, to study the burden of rare variants (defined as variants with MAF < 0.03 in the current data), a variety of methods were applied, including sequence kernel association test and its optimized version, combined multivariate and collapsing, Zeggini and Madsen-Browning tests, as a part of the Rvtest package (Zhan et al., 2016). Bonferroni correction was applied to correct for multiple comparisons. All codes used in the present study are available

UQCRC1 and Parkinson disease risk

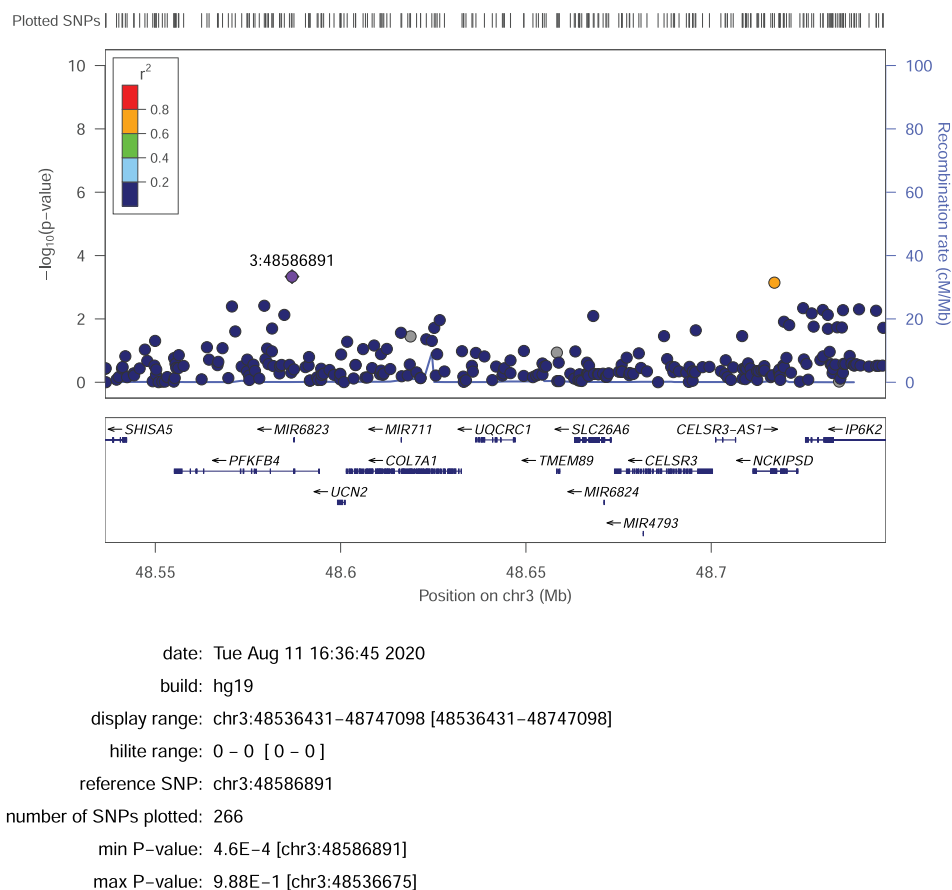


Fig. 1. LocusZoom plot of IPDGC GWAS summary statistics showing variants with MAF more than 1% near the *UQCRC1* gene. Abbreviations: GWAS, genome-wide association study; IPDGC, International Parkinson Disease Genomics Consortium; MAF, minor allele frequency.

at our GitHub at https://github.com/ipdgc/IPDGC-Trainees/blob/master/UQCRC1_IPDGC_trainee.md.

3. Results

Using the IPDGC GWAS data, we identified 140 common variants in the selected region (Fig. 1). None of the common variants annotated to *UQCRC1* were associated with PD (Supplementary Table 1). In the AMP-PD WGS data, we identified 94 variants with MAF <0.03 within or close to *UQCRC1* (Supplementary Table 2), including 9 nonsynonymous variants (Supplementary Table 2). We performed burden tests for 3 categories of variants: 1) all rare variants, 2) all rare coding variants, and 3) all rare nonsynonymous variants. None of the burden tests showed association between rare

UQCRC1 variants and PD (Table 1). In both cohorts, we did not find the *UQCRC1* p.Y314S variant described in the original study (Lin et al., 2019).

4. Discussion

In the present study, we performed a comprehensive analysis of common and rare variants in *UQCRC1* using GWAS and WGS data from large cohorts of patients with PD and controls.

In the original study in which *UQCRC1* variants were implicated, the authors described a family in which 5 carriers of *UQCRC1* p.Y314S had late-onset levodopa responsive parkinsonism and axonal type sensorimotor polyneuropathy (Lin et al., 2019). In addition, autosomal dominant inheritance was demonstrated and

Table 1
 Burden tests for *UQCRC1* in the AMP-PD cohort

Burden test	All variants			Coding variants			Nonsynonymous variants		
	NumVar	NumPolyVar	p-value	NumVar	NumPolyVar	p-value	NumVar	NumPolyVar	p-value
Zeggini	73	54	0.562	18	13	0.619	12	9	0.721
SKAT-O	73	54	0.77	18	13	0.586	12	9	0.712
SKAT	73	54	0.865	18	13	0.476	12	9	0.594
Madson-Browning	73	54	0.487	18	13	0.046	12	9	0.109
Fp	73	54	0.712	18	13	0.489	12	9	0.898
CMC	73	54	0.596	18	13	0.629	12	9	0.731

Key: AMP-PD, Accelerating Medicines Partnership–Parkinson's disease; CMC, combined multivariate and collapsing; NumPolyVar, number of polymorphic genotypes; NumVar, number of variants; SKAT, sequence kernel association test; SKAT-O, sequence kernel association test optimized version.

there were no asymptomatic carriers of this variant, based on a family tree authors present (Lin et al., 2019). Recent replication study in an eastern Chinese population did not reveal association between *UQCRC1* and PD (Lin et al., 2020).

It is possible that this variant is associated with a specific form of atypical parkinsonism, but not with typical PD. It has been shown that a number of genes previously reported as PD-associated (such as *DNAJC13*, *UCHL1*, *HTRA2*, *GIGYF2*, and *EIF4G1*) do not play a role in PD and thus should not be regarded as PD genes (Foo et al., 2014; Krüger et al., 2011; Lesage et al., 2010; Saini et al., 2020). Furthermore, other genes (e.g., *ATP13A2*, *FBXO7*) that are associated with atypical forms of parkinsonism are often cited as PD-associated genes (Dehay et al., 2012; Deng et al., 2015; Weissbach et al., 2019). It is important to properly define which genes are involved in typical PD and which genes are not, particularly in the era of targeted drug development.

Overall, we did not find any evidence to support an important role for *UQCRC1* in patients with PD of European origin. However, additional studies in other populations are required to further study the potential role of *UQCRC1* in PD, and we cannot rule out the possibility that very rare, specific *UQCRC1* variants are associated with PD or with atypical forms of parkinsonism.

Disclosure statement

Z.G.-O. has received consulting fees from Lysosomal Therapeutics Inc., Idorsia, Prevail Therapeutics, Denali, Ono Therapeutics, Neuron23, Handl Therapeutics, Deerfield and Inception Sciences (now Ventus). None of these companies were involved in any parts of preparing, drafting, and publishing this study. Other authors have no additional disclosures to report.

CRediT authorship contribution statement

Konstantin Senkevich: Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing - original draft. **Sara Bandres-Ciga:** Data curation, Resources, Formal analysis, Software, Writing - review & editing. **Ziv Gan-Or:** Conceptualization, Funding acquisition, Methodology, Resources, Supervision, Visualization, Writing - review & editing. **Lynne Krohn:** Conceptualization, Methodology, Project administration, Resources, Supervision, Visualization, Writing - review & editing.

Acknowledgements

We thank the participants for contributing to the study. We would like to also thank all members of the International Parkinson Disease Genomics Consortium (IPDGC). For a complete overview of members, acknowledgements, and funding, please see <http://pdgenetics.org/partners>. 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, Sanofi, 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). 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). For up-to-date information on the study, visit michaeljfox.org/biofind. PPMI—a public-private partnership—is funded by the Michael J. Fox

Foundation for Parkinson's Research and funding partners, including [list the full names of all of the PPMI funding partners found at www.ppmi-info.org/fundingpartners]. For up-to-date information on the study, visit www.ppmi-info.org. 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 BioFIND, PPMI, and PDBP Investigators have not participated in reviewing the data analysis or content of the manuscript. This work was financially supported by grants from the Michael J. Fox Foundation, the Canadian Consortium on Neurodegeneration in Aging (CCNA), the Canada First Research Excellence Fund (CFREF), awarded to McGill University for the Healthy Brains for Healthy Lives initiative (HBHL), and Parkinson Canada. This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging. K.S. is supported by a postdoctoral fellowship from the Canada First Research Excellence Fund (CFREF), awarded to McGill University for the Healthy Brains for Healthy Lives initiative (HBHL). Z.G.-O. is supported by the Fonds de recherche du Québec - Santé (FRQS) Chercheurs-boursiers award, in collaboration with Parkinson Quebec and by the Young Investigator Award by Parkinson Canada.

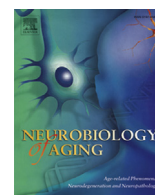
Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.neurobiolaging.2020.10.030>.

References

- Bandres-Ciga, S., Diez-Fairen, M., Kim, J.J., Singleton, A.B., 2020. Genetics of Parkinson's disease: an introspection of its journey towards precision medicine. *Neurobiol. Dis.* 137, 104782.
- Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Giga-science* 4, 7.
- Dehay, B., Ramirez, A., Martinez-Vicente, M., Perier, C., Canron, M.-H., Doudnikoff, E., Vital, A., Vila, M., Klein, C., Bezard, E., 2012. Loss of P-type ATPase ATP13A2/PARK9 function induces general lysosomal deficiency and leads to Parkinson disease neurodegeneration. *Proc. Natl. Acad. Sci.* 109, 9611–9616.
- Deng, H., Wu, Y., Jankovic, J., 2015. The EIF 4G1 gene and Parkinson's disease. *Acta Neurol. Scand.* 132, 73–78.
- Foo, J.N., Liany, H., Tan, L.C., Au, W.-L., Prakash, K.-M., Liu, J., Tan, E.-K., 2014. DNAJ mutations are rare in Chinese Parkinson's disease patients and controls. *Neurobiol. Aging* 35, 935.e931–935.e932.
- Hoffman, G., Lee, S., Christiano, A., Chung-Honet, L., Cheng, W., Katchman, S., Uitto, J., Greenspan, D., 1993. Complete coding sequence, intron/exon organization, and chromosomal location of the gene for the core I protein of human ubiquinol-cytochrome c reductase. *J. Biol. Chem.* 268, 21113–21119.
- Krüger, R., Sharma, M., Riess, O., Gasser, T., Van Broeckhoven, C., Theuns, J., Aasly, J., Annesi, G., Bentivoglio, A.R., Brice, A., 2011. A large-scale genetic association study to evaluate the contribution of Omi/HtrA2 (PARK13) to Parkinson's disease. *Neurobiol. Aging* 32, 548.e549–548.e518.
- Lesage, S., Coudroy, C., Lohman, E., Troiano, A., Tison, F., Viallet, F., Damier, P., Tranchant, C., Vidhaillet, M., Ouvrard-Hernandez, A.-M., 2010. Follow-up study of the GIGYF2 gene in French families with Parkinson's disease. *Neurobiol. Aging* 31, 1069–1071.
- Lin, C.H., Chen, P.L., Tai, C.H., Lin, H.I., Chen, C.S., Chen, M.L., Wu, R.M., 2019. A clinical and genetic study of early-onset and familial parkinsonism in taiwan: an integrated approach combining gene dosage analysis and next-generation sequencing. *Mov. Disord.* 34, 506–515.
- Lin, Z.H., Zheng, R., Ruan, Y., Gao, T., Jin, C.Y., Xue, N.J., Dong, J.X., Yan, Y.P., Tian, J., Pu, J.L., Zhang, B.R., 2020. The lack of association between ubiquinol-cytochrome c reductase core protein 1 (UQCRC1) variants and Parkinson's disease in an eastern Chinese population. *CNS Neurosci. Ther.* 26, 990–992.
- Nalls, M.A., Blauwendraat, C., Valleria, C.L., Heilbron, K., Bandres-Ciga, S., Chang, D., Tan, M., Kia, D.A., Noyce, A.J., Xue, A., Bras, J., Young, E., von Coelln, R., Simon-Sanchez, J., Schulte, C., Sharma, M., Krohn, L., Pihlstrom, L., Siitonen, A., Iwaki, H., Leonard, H., Faghri, F., Gibbs, J.R., Hernandez, D.G., Scholz, S.W., Botia, J.A., Martinez, M., Corvol, J.C., Lesage, S., Jankovic, J., Shulman, L.M., Sutherland, M., Tienari, P., Majamaa, K., Toft, M., Andreassen, O.A., Bangale, T., Brice, A., Yang, J., Gan-Or, Z., Gasser, T., Heutink, P., Shulman, J.M., Wood, N.W., Hinds, D.A., Hardy, J.A., Morris, H.R., Gratten, J., Visscher, P.M., Graham, R.R., Singleton, A.B., Me Research, T., System Genomics of Parkinson's Disease, C., International

- Parkinson's Disease Genomics, C., 2019. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* 18, 1091–1102.
- Saini, P., Rudakou, U., Yu, E., Ruskey, J., Asayesh, F., Laurent, S., Spiegelman, D., Fahn, S., Waters, C., Monchi, O., 2020. Association of DNAJC13, UCHL1, HTRA2, GIGYF2, and EIF4G1 with Parkinson's disease. medRxiv.
- Shan, W., Li, J., Xu, W., Li, H., Zuo, Z., 2019. Critical role of UQCRC1 in embryo survival, brain ischemic tolerance and normal cognition in mice. *Cell Mol. Life Sci.* 76, 1381–1396.
- Wang, K., Li, M., Hakonarson, H., 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164.
- Weissbach, A., Wittke, C., Kasten, M., Klein, C., 2019. 'Atypical'Parkinson's Disease—Genetic, International Review of Neurobiology. Elsevier, pp. 207–235. <https://www.sciencedirect.com/science/article/pii/S0074774219300972?via%3Dihub>. (Accessed 2 August 2020).
- Zhan, X., Hu, Y., Li, B., Abecasis, G.R., Liu, D.J., 2016. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* 32, 1423–1426.



Negative results

The role of *RHOT1* and *RHOT2* genetic variation on Parkinson disease risk and onset

María Teresa Perinán^a, Pilar Gómez-Garre^a, Cornelis Blauwendraat^b, Pablo Mir^a, Sara Bandres-Ciga^{b,*}, on behalf of the International Parkinson's Disease Genomics Consortium (IPDGC)

^a Unidad de Trastornos del Movimiento, Servicio de Neurología y Neurofisiología Clínica, Instituto de Biomedicina de Sevilla, Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla; CIBERNED, Seville, Spain

^b Molecular Genetics Section, Laboratory of Neurogenetics, NIA, NIH, Bethesda, MD, USA

ARTICLE INFO

Article history:

Received 1 June 2020

Received in revised form 30 June 2020

Accepted 3 July 2020

Available online 14 July 2020

Keywords:

Parkinson disease

Genetics

Risk

Mitochondrial pathway

RHOT1

RHOT2

ABSTRACT

Genetic variation within the mitochondrial pathway contributes to the risk of Parkinson's disease (PD). Recent genetic analyses have investigated the association between the *RHOT1* and *RHOT2* genes and PD etiology. Furthermore, 4 mutations in the *RHOT1* gene (p.R272Q, p.R450C, p.T351A, p.T610A) have been reported to be potentially associated with disease risk. As part of the International Parkinson Disease Genomics Consortium efforts to evaluate reported PD risk factors, we assessed the role of common and low frequency variants in both *RHOT1* and also *RHOT2* according to the high degree of homology in their amino acid sequences. Utilizing large-scale genotyping and whole-genome sequencing data from the International Parkinson Disease Genomics Consortium and the Accelerating Medicines Partnership – Parkinson Disease initiative, our analyses did not identify evidence to support the hypothesis that *RHOT1* and *RHOT2* are disease causing or modifying genes for PD risk or age at onset.

© 2020 Elsevier Inc. All rights reserved.

1. Introduction

Mitochondrial function was the first biological process to be associated with Parkinson's disease (PD) (Hardy et al., 2010). Overexpression of Miro was reported to cause significant loss of dopaminergic neurons in *Drosophila* and knockout resulted in absence of mitochondria from axons and synapses (Guo et al., 2005; Liu et al., 2012). As a consequence, *RHOT1* and *RHOT2*, encoding Miro1 and Miro2 respectively, were initially investigated as possible candidate genes associated with PD risk. A recent genetic screening of *RHOT1* identified 2 PD associated variants (p.R272Q, p.R450C) in 2 unrelated German late-onset cases (Grossmann et al., 2019). An independent study reported 2 additional *RHOT1* mutations in 2 German PD patients (p.T351A, p.T610A) (Berenguer-Escuder et al., 2019). Due to the high homology between Miro1 and Miro2 (Fransson et al., 2003), we comprehensively investigate whether genetic variation within *RHOT1* and *RHOT2* influences PD risk and age at onset utilizing large-scale genotyping and whole-

genome sequencing data from the International Parkinson Disease Genomics Consortium (IPDGC) and Accelerating Medicines Partnership – Parkinson Disease (AMP-PD) initiative.

2. Methods

We analyzed IPDGC individual-level genome-wide genotyping data consisting of 14,671 cases and 17,667 healthy controls of European ancestry (Supplementary Table 1). Additionally, we utilized AMP-PD whole-genome sequencing (WGS) data consisting of 1647 cases and 1050 healthy controls from European descent. Variants were annotated using ANNOVAR (Wang et al., 2010) and Fisher's exact test was used to examine the differences between allele frequencies using PLINK 1.9 (Chang et al., 2015).

To assess the cumulative effect of multiple low frequency variants on the risk for PD, gene-based burden analyses were performed on both the IPDGC Genome wide association study (GWAS) data and AMP-PD WGS data by using RVTESTS (Zhan et al., 2016). Summary statistics from the latest meta-analyses on PD (Nalls et al., 2019) and age at onset (Blauwendraat et al., 2019) were further assessed.

* Corresponding author at: Porter Neuroscience Center, 35 Convent Drive, Bethesda, MD 20892, USA. Tel.: +1 301 451 6079; fax: +1 301 451 5466.

E-mail address: sara.bandresciga@nih.gov (S. Bandres-Ciga).

3. Results

Leveraging the IPDGC genotyping data, 62 variants were identified within the *RHOT1* gene, of which all were noncoding variants. Additionally, we detected 17 variants in the *RHOT2* gene, including 4 missense variants (Supplementary Table 2). We further explored summary statistics from the latest PD risk (Nalls et al., 2019) and age at onset (Blauwendraat et al., 2019) GWAS meta-analyses and did not identify a significant association between *RHOT1* and *RHOT2* common genetic variation and PD risk or age at onset (Fig. 1; Supplementary Table 2). Gene-based burden analysis was performed to assess the cumulative effect of low-frequency genetic variants in *RHOT1* and PD. Following this, we found no evidence for an association between *RHOT1* and PD risk (N variants = 7; CMC p = 0.594, Zeggini p = 0.444, MB p = 0.381, SKAT p = 0.385, SKAT-O p = 0.388, Fp p = 0.375). Similarly, we did not identify a link between the *RHOT2* gene and PD risk when the aggregated effect of low-frequency genetic variants was assessed (N variants = 3; CMC p = 0.276, Zeggini p = 0.205, MB p = 0.226, SKAT p = 0.224, SKAT-O p = 0.226, Fp p = 0.226).

Finally, we identified a total of 973 variants within the *RHOT1* gene in WGS data, of which 10 were coding variants, including 4 synonymous and 6 nonsynonymous variants. A total of 118 variants

were detected in the *RHOT2* gene, of which 48 were coding variants, including 25 synonymous and 23 nonsynonymous variants (Supplementary Table 3). Fisher's exact test did not show a significant enrichment for any of the *RHOT1* variants in PD patients versus controls after Bonferroni correction (threshold for significance = $0.05/973 = 5.1 \times 10^{-5}$) nor for any of the *RHOT2* variants (threshold for significance = $0.05/118 = 4.2 \times 10^{-4}$). Additionally, we found no evidence for an association between *RHOT1* low-frequency genetic variation and PD risk, when gene-based burden analyses were applied (N variants = 9; CMC p = 0.452, Zeggini p = 0.452, MB p = 0.650, SKAT p = 0.616, SKAT-O p = 0.678, Fp p = 0.330). Similarly, the *RHOT2* gene did not show a consistent cumulative effect on PD risk across all the tests applied (N variants = 40; CMC p = 0.044, Zeggini p = 0.166, MB p = 0.404, SKAT p = 0.125, SKAT-O p = 0.210, Fp p = 0.659).

4. Discussion

Our study aimed at exploring the role of *RHOT1* and *RHOT2* in PD disease risk and onset by screening large datasets from the IPDGC and AMP-PD initiative. Following our analyses we did not identify evidence to support the hypothesis that *RHOT1* and *RHOT2* are disease causing or modifying genes for PD in the European

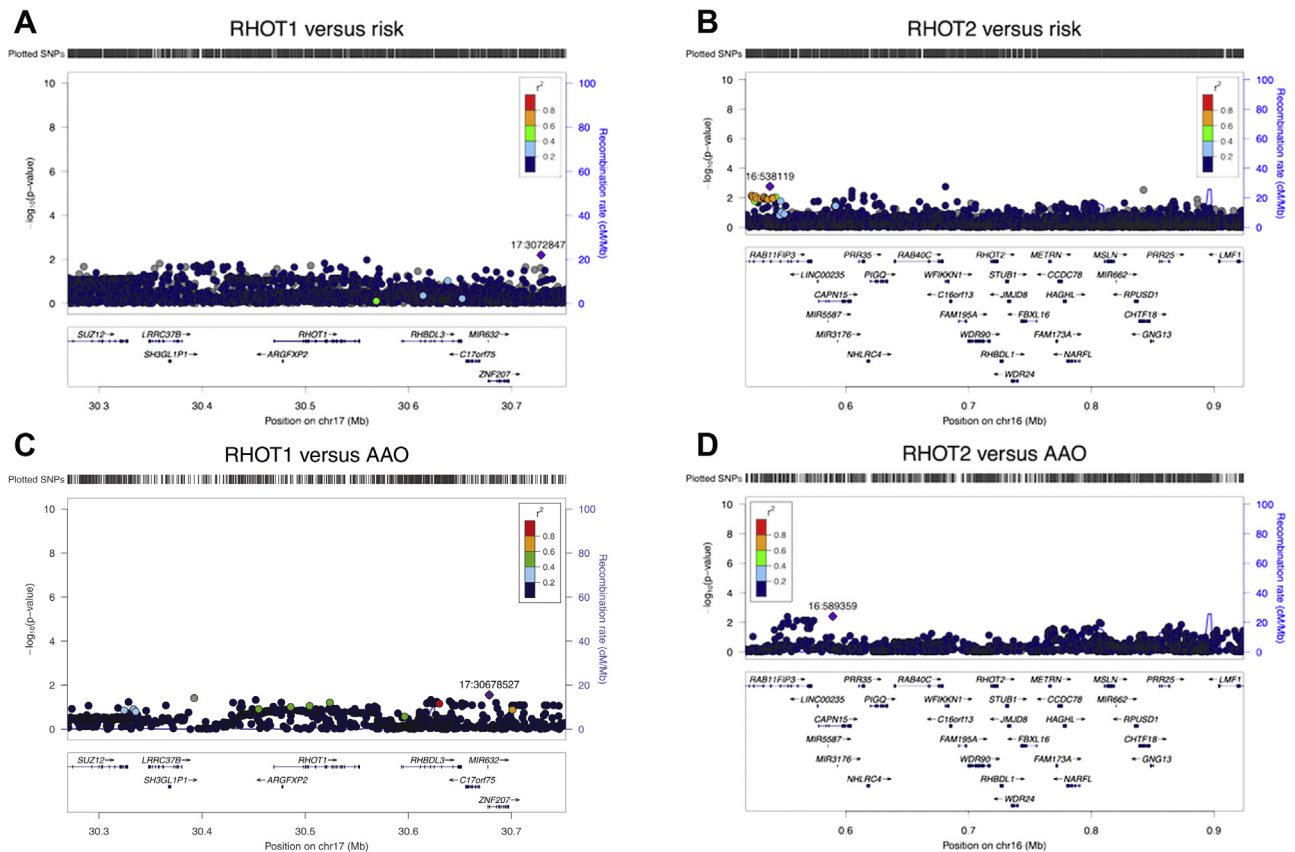


Fig. 1. (A) Locus zoom plot for *RHOT1* variants versus PD risk. Data were obtained from the latest PD GWAS meta-analysis excluding 23andMe data, consisting of 15,056 PD cases, 18,618 UK Biobank proxy-cases, and 449,056 healthy controls (Nalls et al., 2019). (B) Locus zoom plot for *RHOT2* variants versus PD risk. (C) Locus zoom plot for *RHOT1* variants versus PD age at onset. Data were obtained from the most recent age at onset PD GWAS meta-analysis excluding 23andMe data, consisting of 17,996 patients (Blauwendraat et al., 2019). (D) Locus zoom plot for *RHOT2* variants versus PD age at onset. p -values on the log-10 scale are plotted on the left vertical axis and the chromosomal position is plotted along the horizontal axis with the gene names and size of flanking region. The most strongly associated SNPs are indicated by a purple diamond and pairwise LD (r^2) with these SNP are indicated by dotted color as described in the legend in the upper right corner. The right vertical axis indicates the regional recombination rate (cM/Mb) which is overlaid in blue. Abbreviations: LD, linkage disequilibrium; PD, Parkinson's disease; SNP, single nucleotide polymorphism. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

population. Utilizing the largest case-control cohorts publicly available to date, we failed to detect a significant enrichment of *RHOT1* and *RHOT2* risk alleles in PD patients. Moreover, we did not identify a cumulative effect of multiple genetic variants within *RHOT1* and *RHOT2* on the risk for PD when burden analyses including low-frequency variants were performed. The latter is assuming the limitation that when using WGS data we were unpowered to prove a null effect. The *RHOT1* p.T610A variant originally reported by Berenguer-Escuder et al. was the only reported variant present in our WGS data; however, the risk allele was found absent in our cohort of cases and controls (meaning that both cases and controls were carriers for the reference allele). As previously discussed (Blauwendraat et al., 2017), considering that the lifetime risk of developing PD is of 1.3%–2% (depending on sex) (Elbaz et al., 2002), one would expect ~1300 non-Finnish European individuals from the gnomAD population to develop PD. The reported *RHOT1* p.T610A variant has been detected in 3 individuals out of a total of 64,336 included in the gnomAD database, which would represent around ~0.005% of all estimated PD cases. We did not identify this variant in any of the screened cases or controls. Given the low frequency of the *RHOT1* p.T610A variant, we cannot completely discard a possible effect on PD risk. In summary, although genetic discoveries have lent support to the notion that mitochondrial impairment plays a major role on PD disease and onset, our results do not provide enough evidence to confirm *RHOT1* and *RHOT2* as major contributors to disease etiology in the European population.

Disclosure statement

The authors declare that they have no conflict of interest.

CRediT authorship contribution statement

María Teresa Perrián: Formal analysis, Writing - original draft. **Pilar Gómez-Garre:** Investigation. **Cornelis Blauwendraat:** Investigation. **Pablo Mir:** Investigation. **Sara Bandres-Ciga:** Conceptualization, Methodology, Formal analysis, Writing - review & editing.

Acknowledgements

We would like to thank all of the subjects who donated their time and biological samples to be a part of this study. We also would like to thank all members of the International Parkinson's Disease Genomics Consortium (IPDGC). For a complete overview of members, acknowledgments, and funding, please see <http://pdgenetics.org/partners>. This work was supported in part by the Intramural Research Programs of the National Institute of Neurological Disorders and Stroke (NINDS), the National Institute on Aging (NIA), and the National Institute of Environmental Health Sciences both part of the National Institutes of Health, Department of Health and Human Services; project numbers 1ZIA-NS003154, Z01-AG000949-02, and Z01-ES101986. In addition, this work was supported by the Department of Defense (award W81XWH-09-2-0128), and The Michael J. Fox Foundation for Parkinson's Research. 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). This work utilized the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>). María Teresa Perrián is supported by the Spanish Ministry of Education, Culture, and Sports [FPU16/05061].

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <https://doi.org/10.1016/j.neurobiolaging.2020.07.003>.

References

- Berenguer-Escuder, C., Grossmann, D., Massart, F., Antony, P., Burbulla, L.F., Glaab, E., Imhoff, S., Trinh, J., Seibler, P., Grünwald, A., Krüger, R., 2019. Variants in Miro1 cause alterations of ER-mitochondria contact sites in fibroblasts from Parkinson's disease patients. *J. Clin. Med.* 8, 2226.
- Blauwendraat, C., Kia, D.A., Pihlström, L., Gan-Or, Z., Lesage, S., Gibbs, J.R., Ding, J., Alcalay, R.N., Hassin-Baer, S., Pittman, A.M., Brooks, J., Edsall, C., Chung, S.J., Goldwurm, S., Toft, M., Schulte, C., International Parkinson's Disease Genomics Consortium (IPDGC), COURAGE-PD Consortium, Hernandez, D., Singleton, A.B., Nalls, M.A., Brice, A., Scholz, S.W., Wood, N.W., 2017. Insufficient evidence for pathogenicity of SNCA His50Gln (H50Q) in Parkinson's disease. *Neurobiol. Aging* 64, 159.e5–159.e8.
- Blauwendraat, C., Heilbron, K., Vallerger, C.L., Bandres-Ciga, S., von Coelln, R., Pihlström, L., Simón-Sánchez, J., Schulte, C., Sharma, M., Krohn, L., Siitonen, A., Iwaki, H., Leonard, H., Noyce, A.J., Tan, M., Gibbs, J.R., Hernandez, D.G., Scholz, S.W., Jankovic, J., Shulman, L.M., Lesage, S., Corvol, J.-C., Brice, A., van Hilten, J.J., Marinus, J., Eerola-Rautio, J., Tienari, P., Majamaa, K., Toft, M., Grosset, D.G., Gasser, T., Heutink, P., Shulman, J.M., Wood, N., Hardy, J., Morris, H.R., Hinds, D.A., Gratten, J., Visscher, P.M., Gan-Or, Z., Nalls, M.A., Singleton, A.B., 2019. Parkinson's disease age at onset genome-wide association study: defining heritability, genetic loci, and α -synuclein mechanisms. *Mov. Disord.* 34, 866–875.
- Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Giga-science* 4, 7.
- Elbaz, A., Bower, J.H., Maraganore, D.M., McDonnell, S.K., Peterson, B.J., Ahlskog, J.E., Schaid, D.J., Rocca, W.A., 2002. Risk tables for parkinsonism and Parkinson's disease. *J. Clin. Epidemiol.* 55, 25–31.
- Fransson, A., Ruusala, A., Aspenström, P., 2003. Atypical Rho GTPases have roles in mitochondrial homeostasis and apoptosis. *J. Biol. Chem.* 278, 6495–6502.
- Grossmann, D., Berenguer-Escuder, C., Bellet, M.E., Scheibner, D., Bohler, J., Massart, F., Rapaport, D., Skupin, A., Fouquier d'Hérouël, A., Sharma, M., Ghelfi, J., Raković, A., Lichtner, P., Antony, P., Glaab, E., May, P., Dimmer, K.S., Fitzgerald, J.C., Grünwald, A., Krüger, R., 2019. Mutations in *RHOT1* disrupt endoplasmic reticulum-mitochondria contact sites interfering with calcium homeostasis and mitochondrial dynamics in Parkinson's disease. *Antioxid. Redox Signal.* 31, 1213–1234.
- Guo, X., Macleod, G.T., Wellington, A., Hu, F., Panchumarthi, S., Schoenfeld, M., Marin, L., Charlton, M.P., Atwood, H.L., Zinsmaier, K.E., 2005. The GTPase dMiro is required for axonal transport of mitochondria to drosophila synapses. *Neuron* 47, 379–393.
- Hardy, J., 2010. Genetic analysis of pathways to Parkinson disease. *Neuron* 68, 201–206.
- Liu, S., Sawada, T., Lee, S., Yu, W., Silverio, G., Alapatt, P., Millan, I., Shen, A., Saxton, W., Kanoo, T., Takahashi, R., Hattori, N., Imai, Y., Lu, B., 2012. Parkinson's disease-associated kinase PINK1 regulates miro protein level and axonal transport of mitochondria. *PLoS Genet.* 8, e1002537.
- Nalls, M.A., Blauwendraat, C., Vallerger, C.L., Heilbron, K., Bandres-Ciga, S., Chang, D., Tan, M., Kia, D.A., Noyce, A.J., Xue, A., Bras, J., Young, E., von Coelln, R., Simón-Sánchez, J., Schulte, C., Sharma, M., Krohn, L., Pihlström, L., Siitonen, A., Iwaki, H., Leonard, H., Faghri, F., Gibbs, J.R., Hernandez, D.G., Scholz, S.W., Botia, J.A., Martinez, M., Corvol, J.C., Lesage, S., Jankovic, J., Shulman, L.M., Sutherland, M., Tienari, P., Majamaa, K., Toft, M., Andreassen, O.A., Bangale, T., Brice, A., Yang, J., Gan-Or, Z., Gasser, T., Heutink, P., Shulman, J.M., Wood, N.W., Hinds, D.A., Hardy, J.A., Morris, H.R., Gratten, J., Visscher, P.M., Graham, R.R., Singleton, A.B., 2019. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* 18, 1091–1102.
- Wang, K., Li, M., Hakonarson, H., 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164.
- Zhan, X., Hu, Y., Li, B., Abecasis, G.R., Liu, D.J., 2016. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* 32, 1423–1426.

Letter to the Editor

Letter to the editor regarding "TGM6 variants in Parkinson's disease: clinical findings and functional evidence"

Ashley Hall¹, John P. Quinn¹, Kimberley J. Billingsley^{2,*} on behalf of the International Parkinson's Disease Genomics Consortium (IPDGC)

¹Neurobiology Research Group, Molecular and Clinical Pharmacology, Institute of Systems, Molecular & Integrative Biology, University of Liverpool, L69 3BX, Liverpool, United Kingdom

²Molecular Genetics Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, 20892, MD, USA

*Correspondence: kimberley.billingsley@nih.gov (Kimberley J. Billingsley)

DOI: [10.31083/j.jin.2020.04.261](https://doi.org/10.31083/j.jin.2020.04.261)

This is an open access article under the CC BY 4.0 license (<https://creativecommons.org/licenses/by/4.0/>).

Dear Editor:

We read with interest the recent article by [Chen et al. \(2020\)](#) published in the *Journal of Integrative Neuroscience* in which the authors suggest that three rare variants in *TGM6*, encoding transglutaminase 6 (TG6), are associated with Parkinson's Disease (PD). This is notable because the *TGM6* locus was not associated with PD in the most recent meta-analysis of PD GWAS ([Nalls et al., 2019](#)). The authors used targeted gene sequencing to identify eleven rare variants of *TGM6* in idiopathic PD patients, including two previously known function-altering rare variants of *TGM6*, p.R111C and p.L517W. The authors demonstrated that overexpression of these three *TGM6* mutants correlated with reduced TG6 activity, increased levels of α -synuclein, and decreased autophagy flux compared to the wild-type, suggesting a link between *TGM6* variants and PD pathogenesis.

As part of the International Parkinson's Disease Genomics Consortium's (IPDGC) efforts to examine reported genetic risk factors for PD, we assessed genetic variation in *TGM6* using publicly available whole-genome sequencing (WGS) data from the Accelerating Medicines Partnership - Parkinson's disease initiative (AMP-PD) consisting of 1,647 PD patients without known disease-causing mutations (mean age at onset (AAO) 64.2 ± 9.6), of which 145 cases had early-onset PD (mean AAO 45.2 ± 4.6), and 1,050 neurologically healthy controls of European ancestry (mean age 60.3 ± 11.9)¹. Our WGS analysis identified 15 rare coding variants (MAF < 5%): 12 nonsynonymous, of which 1 variant was only found in controls ([Table 1](#)). None of the variants described by [Chen et al. \(2020\)](#) were detected in our analysis.

Association analyses of *TGM6* variants and PD risk were then performed using RVTESTS package v.2.1.0 ([Zhan et al., 2016](#)). Case-control association testing using a single-variant score test adjusted by covariates (including sex, age, and 10 principal components) identified no significant differences in *TGM6* variants' frequency between PD cases and controls after Bonferroni correc-

tion (significance threshold = $0.05/362 = 1.38E-04$). Gene-based burden analyses, again using RVTESTS v.2.1.0, did not detect an enrichment of rare *TGM6* variants in PD cases versus controls (SKAT-O with MAF < 0.03 *P*-values were 0.7733 for all variants and 0.6890 for coding variants). All code used can be found here: (https://github.com/ipdgc/IPDGC-Trainees/blob/master/TGM6_RARE_VARIANT_ANALYSIS.md).

In conclusion, our analyses presented here do not support a role for *TGM6* variants as a risk factor for PD in our cohort. We did not find any *TGM6* variants enriched in PD versus controls, and the additive effects of multiple *TGM6* variants were not found to represent a significant burden for PD. Notably, none of the variants identified by [Chen et al. \(2020\)](#) were observed in our dataset. However, the vast majority of our data is of European ancestry, while [Chen et al. \(2020\)](#) carried out their work in a Chinese population. PD is genetically complex and heterogeneous, and different genetic contexts may influence the importance of variants in the disease. Therefore, our analysis cannot rule out a role for rare *TGM6* variants in non-European populations and highlights the need for further study in diverse ancestries.

Author contributions

Writing and original draft preparation by AH. Methodology, formal analysis and data interpretation by KJB. Review and editing by JPQ. Datasets provided by the IPDGC. All authors have read and agreed to the published version of the manuscript.

Acknowledgment

We want to thank all of the subjects who donated their time and biological samples to participate in this study. We also would like to thank all members of the International Parkinson's Disease Genomics Consortium (IPDGC). For a complete overview of members, acknowledgments and funding, please see <http://pdgenetics.org/partners>. This work was supported in part by the Intramural Research Programs of the National Institute of Neurological Disorders and Stroke (NINDS), the National Insti-

¹ The Accelerating Medicine Partnership in Parkinson's Disease (AMP-PD). (2020) Available at: <https://amp-pd.org/>

Table 1. Frequency of rare coding variants detected in AMP-PD WGS data (grey).

Variant	rsID	gnomAD		AMP-PD				
		Frequency (European non-Finnish)	Frequency (East Asian)	Frequency (cases)	Frequency (controls)	OR	CI 95	P-value (single-variant test)
p.W11G	rs141178275	0.00003149	0	0.001518	0.001429	1.063	0.2537-4.451	0.64927
p.S14S	rs149394698	0.005994	0.00005017	0.006982	0.01095	0.635	0.3553-1.135	0.827472
p.S39C	rs144201778	0.0007248	0	0.0006072	0.0009524	0.6373	0.0897-4.528	0.18099
p.M93V	rs146529216	0.00005423	0	0.003643	0.001429	2.556	0.7204-9.067	0.0433451
p.A141E	rs73894929	0.0006889	0	0.0003036	0.0009524	0.3186	0.02887-3.515	0.369187
p.S159S	rs16984872	0.008834	0	0.009715	0.0119	0.8142	0.4812-1.378	0.843633
p.G243S	rs202245813	0.00003542	0	0.0009107	0.001905	0.4777	0.1068-2.136	0.712623
p.N331S	rs138323389	0.00002325	0	0	0.001429	0	0	0.907626
p.E337E	rs142832802	0.0007279	0	0.001214	0.0004762	2.552	0.285-22.85	0.409379
p.E406K	rs144338465	0.002036	0.00005013	0.001821	0.0009524	1.914	0.386-9.493	0.967385
p.R448W	rs147979536	0.01958	0.001255	0.01913	0.01381	1.392	0.8939-2.169	0.227511
p.G508D	rs140719871	0.002588	0	0.002429	0.001905	1.276	0.3837-4.242	0.659792
p.V527E	rs61729226	0.0004252	0	0.0009107	0.0009524	0.9562	0.1596-5.728	0.508315
p.P564S	rs200674917	0.0005886	0	0.0006072	0.0004762	1.275	0.1156-14.07	0.923033
p.R665S	rs138807504	0.001883	0.0001504	0.003947	0.004762	0.8281	0.3625-1.892	0.693725
p.R111C	rs372250159	0.00000774	0.0002005	-	-	-	-	-
p.L517W	rs387907097	0	0.002011	-	-	-	-	-
p.V208M	rs61743614	0.00003098	0	-	-	-	-	-
p.V255M	rs751618387	0.00003878	0	-	-	-	-	-
p.G199S	rs182249285	0.00004396	0.0004893	-	-	-	-	-
p.V314M	rs202184911	0.000007744	0.005362	-	-	-	-	-
p.P347L	rs183670042	0	0.006716	-	-	-	-	-
p.A34G	rs566673079	0	0	-	-	-	-	-
p.R81W	NA	NA	NA	-	-	-	-	-
p.R540G	rs61740142	NA	0.0006383	-	-	-	-	-
p.V391M	rs116904482	0.0001006	0.009422	-	-	-	-	-

Odds ratio (OR), 95% confidence interval (CI 95) and *P*-value of single-variant score test are presented. For comparison, frequencies in the gnomAD database for European non-Finnish and East Asian cohorts are included. The gnomAD frequencies in the same cohorts for the variants described by [Chen et al. \(2020\)](#), and not detected in our analysis of AMP-PD, are also presented (bottom).

tute on Aging (NIA), and the National Institute of Environmental Health Sciences, both part of the National Institutes of Health, Department of Health and Human Services; project numbers 1ZIA-NS003154, Z01-AG000949-02 and Z01-ES101986. This work was also supported by the Department of Defense (award W81XWH-09-2-0128) and The Michael J Fox Foundation for Parkinson's Research. 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 want 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), the Parkinson's Disease Biomarkers Program (PDBP), and the Harvard Biomarker Study (HBS). Data used in the preparation of this article were obtained from the PPMI database. For up-to-date information on the study, visit <https://www.ppmi-info.org/data>. The PDBP is a consortium sup-

ported by NINDS. A full list of PDBP investigators can be found at <https://pdbp.ninds.nih.gov/policy>. 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. This work utilized the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>).

Conflict of Interest

The authors declare that they have no conflict of interest.

Submitted: August 31, 2020

Revised: October 19, 2020

Accepted: October 19, 2020

Published: December 30, 2020

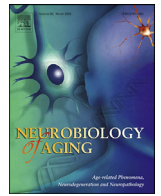
References

- Chen, K., Lu, Y., Peng, F., Yu, H. L., Wu, J. Y., Tan, Y. and Zhao, Y. X. (2020) TGM6 variants in Parkinson's disease: clinical findings and functional evidence. *Journal of Integrative Neuroscience* **19**, 51-64.
- Nalls, M. A., Blauwendraat, C., Vallerger, C. L., Heilbron, K., Bandres-Ciga, S., Chang, D., Tan, M., Kia, D. A., Noyce, A. J., Xue, A., Bras, J., Young, E., von Coelln, R., Simón-Sánchez, J., Schulte, C., Sharma, M., Krohn, L., Pihlstrøm, L., Siitonen, A., Iwaki, H., Leonard, H., Faghri, F., Gibbs, J. R., Hernandez, D. G., Scholz, S. W., Botia, J. A., Martinez, M., Corvol, J. C., Lesage, S., Jankovic, J., Shulman, L. M., Sutherland, M., Tienari, P., Majamaa, K., Toft, M., Andreassen, O. A., Bangale, T., Brice, A., Yang, J., Gan-Or, Z., Gasser, T., Heutink, P., Shulman, J. M., Wood, N. W., Hinds, D. A., Hardy, J. A., Morris, H. R., Gratten, J., Visscher, P. M., Graham, R. R. and Singleton, A. B. (2019) Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *The Lancet Neurology* **18**, 1091-1102.
- Zhan, X., Hu, Y., Li, B., Abecasis, G. R. and Liu, D. J. (2016) RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* **32**, 1423-1426.



Contents lists available at ScienceDirect

Neurobiology of Aging

journal homepage: www.elsevier.com/locate/neuaging.org

Assessment of ANG variants in Parkinson's disease

Francis P. Grenn^a, Anni Moore^a, Sara Bandres-Ciga^a, Lynne Krohn^{b,c},
Cornelis Blauwendraat^{a,*}, on behalf of the International Parkinson's Disease Genomics Consortium (IPDGC)

^a Molecular Genetics Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA

^b Department of Human Genetics, McGill University, Montreal, Quebec, Canada

^c Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada

ARTICLE INFO

Article history:

Received 10 September 2020

Revised 1 December 2020

Accepted 9 March 2021

Available online xxx

Keywords:

Parkinson's disease

Risk factor

ANG

ABSTRACT

Genetic risk factors are occasionally shared between different neurodegenerative diseases. Previous studies have linked ANG, a gene encoding angiogenin, to both Parkinson's disease (PD) and amyotrophic lateral sclerosis (ALS). Functional studies suggest ANG plays a neuroprotective role in both PD and amyotrophic lateral sclerosis by reducing cell death. We further explored the genetic association between ANG and PD by analyzing genotype data from the International Parkinson's Disease Genomics Consortium (14,671 cases and 17,667 controls) and whole genome sequencing data from the Accelerating Medicines Partnership - Parkinson's disease initiative (AMP-PD, <https://amp-pd.org/>) (1,647 cases and 1,050 controls). Our analysis did not replicate the findings of previous studies and identified no significant association between ANG variants and PD risk.

Published by Elsevier Inc.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

1. Introduction

Parkinson's disease (PD) is a neurodegenerative disease characterized by loss of dopaminergic neurons in the substantia nigra leading to symptoms of tremor, rigidity and slowed movement, and is the second most common neurodegenerative disease in the world. Both sporadic and familial forms of PD exist, and much work has been done to identify the environmental and genetic risk factors behind this disease. Over 20 genes have been associated with PD or parkinsonism in recent years, and the largest genome wide association studies for PD risk have identified 92 PD risk variants across 80 loci, explaining 16–36% of the heritable risk of PD (Blauwendraat, Nalls, and Singleton 2020; Nalls et al. 2019; Foo et al. 2020).

It is not uncommon to find genetic variations associated with multiple neurodegenerative disorders, suggesting shared pathways between diseases (Tan et al. 2019). For example, common variations in *MAPT* have been associated with PD (Zabetian et al. 2007), amyotrophic lateral sclerosis (ALS) (Karch et al. 2018) and Alzheimer's disease (AD) (Ferrari et al. 2017), and variations in *GBA* have been associated with PD (Sidransky et al. 2009) and Gaucher

disease (Riboldi and Di Fonzo 2019). Therefore, the interrogation of genes common to multiple neurodegenerative disorders is a logical next step in the identification of novel PD risk variants.

One such candidate is ANG, a relatively small protein coding gene on chromosome 14 with a 444 base pair coding region on two exons. This gene is thought to confer a large risk for both ALS and PD (Rayaprolu et al. 2012; van Es et al. 2011). However, studies in Asian populations have suggested there is no link between ANG variants and PD (Chen et al. 2014; Liu et al. 2013). ANG encodes angiogenin, a small protein that plays a role in the angiogenesis pathway, which forms new blood vessels. Angiogenin and its related pathway are thought to play a role in cancer and placental development (Amankwah, Sellers, and Park 2012; Pavlov et al. 2014). An in vitro study has shown that angiogenin has a neuroprotective effect on motor neurons (Subramanian, Crabtree, and Acharya 2008). ALS associated ANG variants are suggested to potentiate neuronal death through inhibition of the PI3K-Akt pathway (Kieran et al. 2008). A PD mouse model has also shown this gene has a neuroprotective effect on dopaminergic neurons (Steidinger, Standaert, and Yacoubian 2011). This neuroprotective effect is suggested to be lost when ANG is mutated, decreasing the viability of motor neurons (Wu et al. 2007). These findings are of relevance because PD is characterized by the loss of dopaminergic neurons and ALS is characterized by the loss of motor neurons. Angiogenin may play a larger role in areas such as the basal ganglia,

* Corresponding author at: Cornelis Blauwendraat, Porter Neuroscience Center, 35 Convent Drive, Bethesda, MD 20892, USA Phone: 301 435 1788

E-mail address: cornelis.blauwendraat@nih.gov (C. Blauwendraat).

a brain structure often associated with PD. This is supported by a study that identified elevated blood serum angiogenin levels in ALS patients, but not in PD patients (van Es et al. 2014). Structural work has shown ten *ANG* coding variants are associated with a decrease in angiogenin activity, and one coding variant, p.Arg145Cys, is associated with an increase in activity (Bradshaw et al. 2017).

Case control studies focusing on *ANG* have reported associations between *ANG* variants and PD. A study including 6,471 ALS cases, 3,146 PD cases and 7,668 controls from American and European populations has reported associations between *ANG* variants and both PD and ALS (van Es et al. 2011). A separate study including 630 PD cases and 676 controls from American populations has also reported association between *ANG* variants and PD (Rayaprolu et al. 2012). However, to date *ANG* variants have not been associated with either ALS or PD through genome wide association studies (GWAS) (Nicolas et al. 2018; Nalls et al. 2019). Here we scrutinize *ANG* variants in two large PD datasets to assess whether *ANG* variants contribute to PD risk in individuals of European ancestry.

2. Methods

2.1. Whole-genome sequencing and genotype data

We mined whole-genome sequencing (WGS) data from the Accelerating Medicines Partnership - Parkinson's disease initiative (AMP-PD, <https://amp-pd.org/>) which included 1,647 cases and 1,050 healthy controls from cohorts of European ancestry. The cohorts included were the Fox Investigation for New Discovery of Biomarkers (BioFIND), the Parkinson's Progression Markers Initiative (PPMI), the Harvard Biomarker Study (HBS), and the Parkinson's Disease Biomarkers Program (PDBP). We also looked at *ANG* variants in genotype data from the International Parkinson's Disease Genomics Consortium (IPDGC) which included 14,671 cases and 17,667 healthy controls. IPDGC variants were converted from hg19 to hg38 positions to match AMP-PD variant hg38 positions. Variants from both datasets were annotated using ANNOVAR (Wang, Li, and Hakonarson 2010). Variant frequencies in non-Finnish European populations were obtained from the hg38 gnomAD v3.0 dataset (Karczewski et al. 2020).

Variants identified by amino acid change from previous studies including Van Es et al. and Rayaprolu et al. did not initially match any variants identified in AMP-PD data due to differences in amino acid numbering. To resolve this we mapped each variant amino acid change to the angiogenin protein sequence. This sequence was obtained from Ensembl using the *ANG*-201 ENST00000336811.10 hg38 transcript (Yates et al. 2020). We learned that the reported amino acid changes from Van Es et al. and Rayaprolu et al. were offset by 25 or 24 amino acids due to numbering differences used for the signal peptide sequence that is cleaved from the mature angiogenin protein. This was accounted for in our analysis.

2.2. Statistical analyses

PLINK 1.9 was used to perform Fisher's exact test to identify significant variants (Purcell et al. 2007). Rare variant burden tests and single variant score tests were performed using RVTESTS (Zhan et al. 2016). We further analyzed existing summary statistics including the latest GWAS meta-analyses for PD risk and age of onset (Blauwendraat et al. 2019; Nalls et al. 2019) and additionally assessed public summary statistics from the most recent ALS GWAS (Nicolas et al. 2018).

We used the Genetic Association Study Power Calculator to calculate the statistical power of our study (Johnson and Abecasis, n.d.).

Principal component analysis was performed on AMP-PD samples to check for population stratification. Data from the International HapMap Project was merged with AMP-PD data for comparison (Supplementary Figure 3) (Consortium, the International HapMap and The International HapMap Consortium 2003).

2.3. Code

The code we used for analysis is available on the IPDGC github (<https://github.com/ipdgc/IPDGC-Trainees/blob/master/ANG.md>).

2.4. Expression Quantitative Trait Loci

We obtained expression quantitative trait loci (eQTL) data for *ANG* variants found in AMP-PD and IPDGC datasets. Both cis and trans eQTLs were obtained from the eQTLGen Consortium (Vösa et al., 2018). The Genotype-Tissue Expression portal was accessed on November 18, 2020 to obtain additional eQTL data.

3. Results

We identified a total of 168 *ANG* variants in the AMP-PD WGS data, with 9 of these identified as coding. We compared the 9 identified *ANG* coding variants to variants from two other studies (van Es et al. 2011; Rayaprolu et al. 2012) (Supplementary Table 2). Of the coding variants, two were synonymous and the other seven were nonsynonymous. All nonsynonymous variants were rare ($MAF < 0.01$), and allele frequencies did not differ significantly from gnomAD non-Finnish European allele frequencies although most variants were too rare to reliably test individually. The top variant after performing Fisher's exact test ($p = 0.017$) and single variant score test ($p = 0.016$) was not significant after Bonferroni correction for multiple tests ($p = 0.05/168 = 2.97E-4$) (Supplementary Table 1). Burden tests with a minor allele frequency less than 0.03 reported no significant results when using all variants (N variants=72; CMC $p = 0.493$, Fp $p = 0.509$, MB $p = 0.880$, Skat $p = 0.454$, SkatO $p = 0.523$, Zeggini $p = 0.395$). Likewise, there were no significant results when performing the same test on only coding variants (N variants=9; CMC $p = 0.866$, Fp $p = 0.510$, MB $p = 0.820$, Skat $p = 0.436$, SkatO $p = 0.556$, Zeggini $p = 0.868$).

After excluding the only 2 synonymous variants, rs11701 (p.G110=) and rs2228653 (p.T121=), from these 9, we observed an average frequency of 0.39% in PD cases and 0.48% in controls. Van Es et al. also removed 2 common variants, rs121909536 (K41I) and rs121909541 (p.I70V), from their analysis. After removing these same 2 variants from our data the average frequencies were 0.15% in PD cases and 0.19% in controls. This is in contrast with the 0.45% average frequency in PD cases and 0.05% in controls previously reported (van Es et al. 2011).

Twenty-six *ANG* variants were identified using the IPDGC imputed genotype data, all of which were non-coding (Supplementary Table 1). All variants were high quality imputed ($r^2 > 0.8$) or were directly genotyped. No significant association between *ANG* variants and PD risk (Fig. 1A) or onset (Supplementary Figure 1) was identified in data from the latest PD risk GWAS or in the PD age of onset GWAS (Nalls et al. 2019; Blauwendraat et al. 2019). No variants had a minor allele frequency less than 0.03, so the threshold was increased to 0.05 for burden tests. Only 2 variants were included at this threshold, which also reported no significant results ($N=2$; CMC $p = 0.893$, Fp $p = 0.960$, MB $p = 0.948$, Skat $p = 0.980$, SkatO $p = 1$, Zeggini $p = 0.842$). Additionally, no GWAS signal of interest is identified in the most recent ALS GWAS (Fig. 1B) (Nicolas et al. 2018).

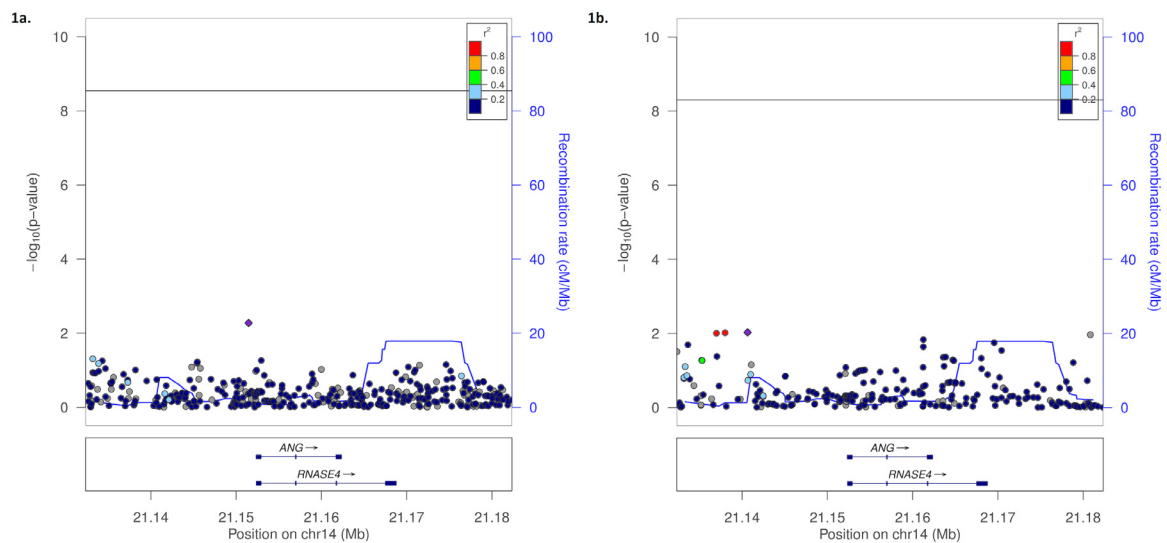


Fig. 1. Locus zoom plots for ANG and PD risk and ALS risk. The $-\log_{10}(p\text{-value})$ of variants on or near ANG are shown on the y-axis, and base-pair position of each variant is on the x-axis. p -values are taken from the PD risk GWAS (Fig. 1A) and the ALS risk GWAS (Fig. 1B). Variants are colored by their R^2 linkage disequilibrium value which is relative to the variant with the lowest p -value on these plots (colored purple). The genome-wide significance cutoff line for multiple test correction is included in black. Recombination rates are included in blue (Pruim et al. 2010).

4. Discussion

Rare coding variants in *ANG* have been reported to be associated with PD (van Es et al. 2011). Our goal was to further explore the role of *ANG* in PD by analyzing additional datasets from IPDGC and AMP-PD. Our study shows no significant enrichment of *ANG* single variants in PD cases or controls in either of these datasets. Rare variant burden tests also reported no significant results for *ANG*. Our analysis provides no evidence to support the hypothesis that genetic variation of *ANG* plays a role in PD risk or age at onset.

The 9 coding *ANG* variants we identified were from AMP-PD WGS data. This dataset included fewer samples (1,647 PD; 1,050 controls) than the Van Es et al. study (6,471 ALS; 3,146 PD; 7,668 non-ALS controls; 5,361 non-PD controls) which identified a total of 29 unique *ANG* coding variants. However, the frequency of *ANG* coding variants detected in AMP-PD data is 0.15% in PD cases and 0.19% in controls which is different from the previously reported 0.45% in PD cases (14 variants in 3,146 PD cases) and 0.05% in controls (3 variants in 5,361 non-PD controls) (van Es et al. 2011).

We queried eQTL datasets from eQTLGen and GTEx for *ANG* eQTLs to identify variants that affect gene expression levels (Võsa et al., 2018). Between both eQTLGen and GTEx data we identified a total of 725 unique *ANG* eQTLs (Supplementary Table 3). Twenty six of these 725 were trans eQTLs and the other 699 were cis eQTLs. Out of the 170 genetic variants identified in AMP-PD and IPDGC data, 74 were eQTLs for *ANG* or the nearby gene, *RNASE4*, ten of these eQTLs were located in the 5' untranslated region of *ANG*, one was exonic, and the other 63 were intronic. However, given that the genetic variants were not determined to be significant, it is likely that eQTLs also do not contribute to disease.

We used the Genetic Association Study Power Calculator to determine if our study was sufficiently powered to replicate the findings of previous studies (Johnson and Abecasis, n.d.). We used 1,647 cases and 1,050 controls as inputs, with a significance level of 0.05, and a disease prevalence of 0.01. We calculated the average allele frequency in AMP-PD data to be 0.0017 and included it along with a genotype relative risk of 6.7 which was taken from previous studies to allow for comparison with our study (van Es et al. 2011). We calculated a statistical power of 0.997, suggesting we are sufficiently powered to detect significant rare *ANG* variants in PD cases

and controls (Supplementary Figure 2). Additionally, AMP-PD samples used in our study were age matched which is in line with previous studies (Supplementary Table 4). However, the cumulative frequency of *ANG* variants identified in AMP-PD data was not significantly different as previously reported. A larger sample size may be needed to identify the missing coding variants so the role of *ANG* in PD can be assessed on an even larger scale.

Overall, despite some potentially interesting functional experiments supporting the neuroprotective effect of angiogenin, we cannot replicate the genetic association between *ANG* coding variants and PD. Therefore, we cannot conclude that *ANG* variants play a role in PD in European populations, which is in line with previous studies done in Asian populations (Chen et al. 2014; Liu et al. 2013).

Declaration of competing interest

The authors declare that they have no conflict of interest.

Acknowledgments

We would like to thank all of the subjects who donated their time and biological samples to be a part of this study. We also would like to thank all members of the International Parkinson's Disease Genomics Consortium (IPDGC). For a complete overview of members, acknowledgements and funding, please see <http://pdgenetics.org/partners>. This work was supported in part by the Intramural Research Programs of the National Institute of Neurological Disorders and Stroke (NINDS), the National Institute on Aging (NIA), and the National Institute of Environmental Health Sciences (NIEHS), all part of the National Institutes of Health, Department of Health and Human Services; project numbers 1Z1A-NS003154, Z01-AG000949-02 and Z01-ES101986. In addition, this work was supported by the Department of Defense (award W81XWH-09-2-0128), and The Michael J Fox Foundation for Parkinson's Research. 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. This work utilized the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:[10.1016/j.neurobiolaging.2021.03.006](https://doi.org/10.1016/j.neurobiolaging.2021.03.006).

References

- Amankwah, Ernest K., Sellers, Thomas A., Park, Jong Y., 2012. Gene variants in the angiogenesis pathway and prostate cancer. *Carcinogenesis* 33 (7), 1259–1269.
- Blauwendraat, Cornelis, Heilbron, Karl, Vallerger, Costanza L., Bandres-Ciga, Sara, von Coelln, Rainer, Pihlström, Lasse, Simón-Sánchez, Javier, et al., 2019. Parkinson's disease age at onset genome-wide association study: defining heritability, genetic loci, and α -synuclein mechanisms. *Movement Disorders: Off J Movement Disorder Soc* 34 (6), 866–875.
- Blauwendraat, Cornelis, Nalls, Mike A., Singleton, Andrew B., 2020. The genetic architecture of Parkinson's Disease. *Lancet Neurol* 19 (2), 170–178.
- Bradshaw, William J., Rehman, Saima, Pham, Tram T.K., Thiagarajan, Nethaji, Lee, Rebecca L., Subramanian, Vasanta, Ravi Acharya, K., 2017. Structural insights into human angiogenin variants implicated in Parkinson's Disease and Amyotrophic Lateral Sclerosis. *Scientific Reports* doi:[10.1038/srep41996](https://doi.org/10.1038/srep41996).
- Chen, Meng-Ling, Wu, Ruey-Meei, Tai, Chun-Hwei, Lin, Chin-Hsien, 2014. Mutational analysis of angiogenin gene in Parkinson's Disease. *PLoS One* 9 (11), e112661.
- Consortium, the International Hapmap, and The International HapMap Consortium, 2003. The International HapMap Project. *Nature* doi:[10.1038/nature02168](https://doi.org/10.1038/nature02168).
- Es, Michael A., van, Schelhaas, Helenius J., Vught, Paul W.J., van, Ticozzi, Nicola, Andersen, Peter M., Groen, Ewout J.N., Schulte, Claudia, et al., 2011. Angiogenin variants in parkinson disease and amyotrophic lateral sclerosis. *Ann Neurol* 70 (6), 964–973.
- Es, Michael A., van, Veldink, Jan H., Schelhaas, Helenius J., Bloem, Bastiaan R., Soedaar, Peter, Nuenen, Bart F.L., van, Verbeek, Marcel, Warrenburg, Bart P., van de Berg, Leonard H., van den, 2014. Serum Angiogenin Levels Are Elevated in ALS, but Not Parkinson's Disease: Table 1. *J Neurol, Neurosurg & Psychiatry* doi:[10.1136/jnnp-2013-307168](https://doi.org/10.1136/jnnp-2013-307168).
- Ferrari, Raffaele, Wang, Yunpeng, Vandrovicova, Jana, Gueffi, Sebastian, Witeolar, Aree, Karch, Celeste M., Schork, Andrew J., et al., 2017. Genetic architecture of sporadic frontotemporal dementia and overlap with alzheimer's and parkinson's diseases. *J Neurol, Neurosurg, and Psychiatry* 88 (2), 152–164.
- Foo, Jia, Nee, Elaine Guo, Chew, Yan, Chung, Sun Ju, Peng, Rong, Blauwendraat, Cornelis, Nalls, Mike A., Mok, Kin Y., et al., 2020. Identification of risk loci for parkinson disease in asians and comparison of risk between asians and europeans: a genome-wide association study. *JAMA Neurol* doi:[10.1001/jamaneurol.2020.0428](https://doi.org/10.1001/jamaneurol.2020.0428), April.
- Johnson, Jennifer Li, and Gonçalo R. Abecasis. 2021 n.d. "GAS power calculator: web-based power calculator for genetic association studies." Doi: [10.1101/164343](https://doi.org/10.1101/164343).
- Karch, Celeste M., Wen, Natalie, Fan, Chun C., Yokoyama, Jennifer S., Kouri, Naomi, Ross, Owen A., Höglinger, Gunter, et al., 2018. Selective genetic overlap between amyotrophic lateral sclerosis and diseases of the frontotemporal dementia spectrum. *JAMA Neurol* 75 (7), 860–875.
- Karczewski, Konrad J., Francioli, Laurent C., Tiao, Grace, Cummings, Beryl B., Alföldi, Jessica, Wang, Qingbo, Collins, Ryan L., et al., 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581 (7809), 434–443.
- Kieran, Dairín, Sebastia, Jordi, Greenway, Matthew J., King, Matthew A., Connaughton, Dervla, Concannon, Caoimhín G., Fenner, Beau, Hardiman, Orla, Prehn, Jochen H.M., 2008. Control of motoneuron survival by angiogenin. *The J Neuroscience: The Off J Soc for Neuroscience* 28 (52), 14056–14061.
- Liu, Bin, Zhang, Yu, Wang, Ying, Xiao, Qin, Yang, Qiong, Wang, Gang, Ma, Jianfang, et al., 2013. Angiogenin variants are not associated with parkinson's disease in the ethnic Chinese population. *Parkinsonism & Related Disorders* doi:[10.1016/j.parkreldis.2012.11.016](https://doi.org/10.1016/j.parkreldis.2012.11.016).
- Nalls, Mike A., Blauwendraat, Cornelis, Vallerger, Costanza L., Heilbron, Karl, Bandres-Ciga, Sara, Chang, Diana, Tan, Manuela, et al., 2019. Identification of novel risk loci, causal insights, and heritable risk for parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurology* 18 (12), 1091–1102.
- Nicolas, Aude, Kenna, Kevin P., Renton, Alan E., Ticozzi, Nicola, Faghri, Faraz, Chia, Ruth, Dominov, Janice A., et al., 2018. Genome-wide analyses identify KIF5A as a Novel ALS Gene. *Neuron* 97 (6), 1268–1283 e6.
- Pavlov, Nadine, Frendo, Jean-Louis, Guibourdenche, Jean, Degrelle, Séverine A., Evain-Brion, Danièle, Badet, Josette, 2014. Angiogenin expression during early human placental development; association with blood vessel formation. *BioMed Res Int*, 781632 2014 (July).
- Pruim, Randall J., Welch, Ryan P., Sanna, Serena, Teslovich, Tanya M., Chines, Peter S., Gliedt, Terry P., Boehnke, Michael, Abecasis, Gonçalo R., Willer, Cristen J., 2010. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 26 (18), 2336–2337.
- Purcell, Shaun, Neale, Benjamin, Todd-Brown, Kathe, Thomas, Lori, Ferreira, Manuel A.R., Bender, David, Maller, Julian, et al., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The Am J Human Genetics* doi:[10.1086/519795](https://doi.org/10.1086/519795).
- Rayaprolu, Sruti, Soto-Ortolaza, Alexandra, Rademakers, Rosa, Uitti, Ryan J., Wszolek, Zbigniew K., Ross, Owen A., 2012. Angiogenin variation and parkinson disease. *Ann Neurol*.
- Riboldi, Giulietta M., Di Fonzo, Alessio B., 2019. GBA, Gaucher Disease, and Parkinson's Disease: from genetic to clinic to new therapeutic approaches. *Cells* doi:[10.3390/cells8040364](https://doi.org/10.3390/cells8040364).
- Sidransky, E., Nalls, M.A., Aasly, J.O., Aharon-Peretz, J., Annesi, G., Barbosa, E.R., Bar-Shira, A., et al., 2009. Multicenter analysis of glucocerebrosidase mutations in Parkinson's Disease. *The New England J Med* 361 (17), 1651–1661.
- Steidinger, Trent U., Standaert, David G., Yacoubian, Talene A., 2011. A neuroprotective role for angiogenin in models of Parkinson's Disease. *Journal of Neurochemistry* doi:[10.1111/j.1471-4159.2010.07112.x](https://doi.org/10.1111/j.1471-4159.2010.07112.x).
- Subramanian, Vasanta, Crabtree, Benedict, Ravi Acharya, K., 2008. Human angiogenin is a neuroprotective factor and amyotrophic lateral sclerosis associated angiogenin variants affect neurite extension/pathfinding and survival of motor neurons. *Human Molecular Genetics* 17 (1), 130–149.
- Tan, Sean Hong, Karri, Venkatanaidu, Tay, Nicole Wuen Rong, Chang, Kuan Hui, Ah, Hui Yen, Ng, Phui Qi, Ho, Hui San, Keh, Hsiao Wai, Candasamy, Mayuren, 2019. Emerging Pathways to Neurodegeneration: Dissecting the Critical Molecular Mechanisms in Alzheimer's Disease, Parkinson's Disease. *Biomed & Pharmacotherapy = Biomed & Pharmacotherapie* 111 (March), 765–777.
- Vösa, Urmo, Claringbould, Annique, Westra, Harm-Jan, Bonder, Marc, Deelen, Patrick, Zeng, Biao, Kirsten, Holger, et al., 2018. "Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis." *bioRxiv*, October. doi:[10.1101/447367](https://doi.org/10.1101/447367).
- Wang, Kai, Li, Mingyao, Hakonarson, Hakon, 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38 (16), e164.
- Wu, David, Yu, Wenhao, Kishikawa, Hiroko, Folkert, Rebecca D., Iafrate, A. John, Shen, Yiping, Xin, Winnie, Sims, Katherine, Hu, Guo-Fu, 2007. Angiogenin loss-of-function mutations in amyotrophic lateral sclerosis. *Ann Neurol* 62 (6), 609–617.
- Yates, Andrew D., Achuthan, Premanand, Akanni, Wasiu, Allen, James, Allen, Jamie, Alvarez-Jarreta, Jorge, Ridwan Amodé, M., et al., 2020. Ensembl 2020. *Nucleic Acids Res* 48 (D1), D682–D688.
- Zabetian, C.P., Hutter, C.M., Factor, S.A., Nutt, J.G., Higgins, D.S., Griffith, A., Roberts, J.W., et al., 2007. Association analysis of MAPT H1 Haplotype and Sub-haplotypes in Parkinson's Disease. *Ann Neurol* 62 (2). doi:[10.1002/ana.21157](https://doi.org/10.1002/ana.21157).
- Zhan, Xiaowei, Hu, Youna, Li, Bingshan, Abecasis, Gonçalo R., Liu, Dajiang J., 2016. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* 32 (9), 1423–1426.

Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies



Mike A Nalls*, Cornelis Blauwendraat*, Costanza L Vallergera*, Karl Heilbron*, Sara Bandres-Ciga*, Diana Chang*, Manuela Tan, Demis A Kia, Alastair J Noyce, Angli Xue, Jose Bras, Emily Young, Rainer von Coelln, Javier Simón-Sánchez, Claudia Schulte, Manu Sharma, Lynne Krohn, Lasse Pihlström, Ari Siitonen, Hirotaka Iwaki, Hampton Leonard, Faraz Faghri, J Raphael Gibbs, Dena G Hernandez, Sonja W Scholz, Juan A Botia, Maria Martinez, Jean-Christophe Corvol, Suzanne Lesage, Joseph Jankovic, Lisa M Shulman, Margaret Sutherland, Pentti Tienari, Kari Majamaa, Mathias Toft, Ole A Andreassen, Tushar Bangale, Alexis Brice, Jian Yang, Ziv Gan-Or, Thomas Gasser, Peter Heutink, Joshua M Shulman, Nicholas W Wood, David A Hinds, John A Hardy, Huw R Morris, Jacob Gratten, Peter M Visscher, Robert R Graham, Andrew B Singleton, on behalf of the 23andMe Research Team†, System Genomics of Parkinson's Disease Consortium†, and the International Parkinson's Disease Genomics Consortium†

Summary

Background Genome-wide association studies (GWAS) in Parkinson's disease have increased the scope of biological knowledge about the disease over the past decade. We aimed to use the largest aggregate of GWAS data to identify novel risk loci and gain further insight into the causes of Parkinson's disease.

Methods We did a meta-analysis of 17 datasets from Parkinson's disease GWAS available from European ancestry samples to nominate novel loci for disease risk. These datasets incorporated all available data. We then used these data to estimate heritable risk and develop predictive models of this heritability. We also used large gene expression and methylation resources to examine possible functional consequences as well as tissue, cell type, and biological pathway enrichments for the identified risk factors. Additionally, we examined shared genetic risk between Parkinson's disease and other phenotypes of interest via genetic correlations followed by Mendelian randomisation.

Findings Between Oct 1, 2017, and Aug 9, 2018, we analysed 7·8 million single nucleotide polymorphisms in 37 688 cases, 18 618 UK Biobank proxy-cases (ie, individuals who do not have Parkinson's disease but have a first degree relative that does), and 1·4 million controls. We identified 90 independent genome-wide significant risk signals across 78 genomic regions, including 38 novel independent risk signals in 37 loci. These 90 variants explained 16–36% of the heritable risk of Parkinson's disease depending on prevalence. Integrating methylation and expression data within a Mendelian randomisation framework identified putatively associated genes at 70 risk signals underlying GWAS loci for follow-up functional studies. Tissue-specific expression enrichment analyses suggested Parkinson's disease loci were heavily brain-enriched, with specific neuronal cell types being implicated from single cell data. We found significant genetic correlations with brain volumes (false discovery rate-adjusted $p=0\cdot0035$ for intracranial volume, $p=0\cdot024$ for putamen volume), smoking status ($p=0\cdot024$), and educational attainment ($p=0\cdot038$). Mendelian randomisation between cognitive performance and Parkinson's disease risk showed a robust association ($p=8\cdot00\times10^{-7}$).

Interpretation These data provide the most comprehensive survey of genetic risk within Parkinson's disease to date, to the best of our knowledge, by revealing many additional Parkinson's disease risk loci, providing a biological context for these risk factors, and showing that a considerable genetic component of this disease remains unidentified. These associations derived from European ancestry datasets will need to be followed-up with more diverse data.

Funding The National Institute on Aging at the National Institutes of Health (USA), The Michael J Fox Foundation, and The Parkinson's Foundation (see appendix for full list of funding sources).

Copyright © 2019 Elsevier Ltd. All rights reserved.

Introduction

Parkinson's disease is a neurodegenerative disorder, affecting approximately 1 million individuals in the USA alone.¹ Patients with Parkinson's disease have a combination of progressive motor and non-motor symptoms affecting daily function and quality of life. The prevalence of Parkinson's disease is projected to double in some age

groups by 2030, creating a substantial burden on health-care systems.¹

Early investigations into the role of genetic factors in Parkinson's disease focused on the identification of rare mutations underlying familial disease;^{2,3} however, over the past decade there has been a growing appreciation for the contribution of genetics in sporadic disease.^{4,5} Genetic

Lancet Neurol 2019;
18: 1091–102

See [Comment](#) page 1069

*Joint first authors

†Full consortia membership is available in the appendix

Laboratory of Neurogenetics, National Institute on Aging (M A Nalls PhD, C Blauwendraat PhD, S Bandres-Ciga PhD, H Iwaki MD, H Leonard MS, F Faghri PhD, J R Gibbs PhD, D G Hernandez PhD, A B Singleton PhD) and National Institute of Neurological Disorders and Stroke (S W Scholz MD, M Sutherland PhD), National Institutes of Health, Bethesda, MD, USA; Data Tecnica International, Glen Echo, MD, USA (M A Nalls, H Iwaki, H Leonard); Queensland Brain Institute (A Xue BS, J Yang PhD, P M Visscher PhD), Institute for Molecular Bioscience (C L Vallergera MS, A Xue, J Yang, J Gratten PhD, P M Visscher), Mater Research Institute (J Gratten), The University of Queensland, Brisbane, QLD, Australia; 23andMe, Sunnyvale, CA, USA (K Heilbron PhD, D A Hinds PhD); Department of Human Genetics, Genentech, South San Francisco, CA, USA (D Chang PhD, T Bangale PhD, R R Graham PhD); Department of Molecular Neuroscience (M Tan MS, D A Kia BS, A J Noyce MD, J A Botia PhD, N W Wood MD, J A Hardy), Department of Neurodegenerative Diseases (J Bras), Department of Clinical and Movement Neuroscience and UCL Movement Disorders Centre, UCL Queen Square

Institute of Neurology, London, UK (H R Morris MD, D A Kia, M Tan, N W Wood); Preventive Neurology Unit, Wolfson Institute of Preventive Medicine, Queen Mary University of London, London, UK (A J Noyce); Center for Neurodegenerative Science, Van Andel Research Institute, Grand Rapids, MI, USA (J Bras PhD); Department of Neurology (E Young, J Jankovic MD, J M Shulman MD), Department of Molecular and Human Genetics (J M Shulman), and Department of Neuroscience (J M Shulman) Baylor College of Medicine, Houston, TX, USA; Department of Neurology, University of Maryland School of Medicine, Baltimore, MD, USA (L M Shulman PhD, R von Coelln MD); Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research (J Simón-Sánchez PhD, C Schulte PhD, T Gasser MD, P Heutink PhD), Centre for Genetic Epidemiology, Institute for Clinical Epidemiology and Applied Biometry (M Sharma PhD), University of Tübingen, Tübingen, Germany; German Center for Neurodegenerative Diseases, Tübingen, Germany (J Simón-Sánchez, C Schulte, T Gasser, P Heutink); Department of Human Genetics (L Krohn MS, Z Gan-Or MD), Montreal Neurological Institute (L Krohn, Z Gan-Or), and Department of Neurology and Neurosurgery (Z Gan-Or), McGill University, Montreal, QC, Canada; Department of Neurology (L Pihlstrøm MD, M Toft MD) and Division of Mental Health and Addiction (O A Andreassen), Oslo University Hospital, Oslo, Norway; Institute of Clinical Medicine, Department of Neurology, University of Oulu, Oulu, Finland (A Siitonen PhD, K Majamaa MD); Department of Neurology and Medical Research Center, Oulu University Hospital, Oulu, Finland (A Siitonen, K Majamaa); The Michael J Fox Foundation, New York, NY, USA (H Iwaki); Department of Computer Science, University of Illinois Urbana-Champaign, Champaign, IL, USA (F Faghiri); Department of Neurology,

Research in context

Evidence before this study

Previous studies have used genome-wide association study (GWAS) methods to discover 42 independent risk loci associated with Parkinson's disease. Some of these loci harbouring common risk variants also include rare variants implicated in familial Parkinson's disease risk such as *SNCA*, *LRRK2*, or *GBA*. Earlier studies have attempted to quantify how much heritable risk is captured by common variation that can be easily imputed using commercial genotyping arrays and estimate the amount of risk explained by GWAS. Since 2011, GWAS of Parkinson's disease have integrated expression and methylation datasets to evaluate possible candidate genes for follow-up at Parkinson's disease loci. Many epidemiological and observational studies have attempted to assess risk of Parkinson's disease and various exposures like smoking, caffeine, or occupational hazards, with a mixed track record of success at validating presumed associations.

Added value of this study

This study increased the count of independent common genetic risk factors for Parkinson's disease to 90. We added 38 novel risk variants not previously identified as genome-wide significant. We refined heritability estimates and genetic risk predictions suggesting that common genetic variants account for approximately 22% of Parkinson's disease risk on the liability scale, with a range of 16–36% of that risk being explained by GWAS loci in this study. These updated risk predictions also suggested that polygenic risk scoring can be used to achieve an area under the curve of near 70%, although this prediction uses many more variants than just the 90 independent risk factors identified in this report. Of the 90 risk variants we have characterised here, we have

nominated at least one possible candidate gene for follow-up functional studies in 70 of these genomic regions by mining recently available expression and methylation reference datasets on a scale not possible just a few years ago. We have additionally mined single cell RNA sequencing data from mice to identify tissue-specific signatures of enrichment relating to Parkinson's disease genetic risk, showing a major focus on neuronal cell types. We also used the massive amount of publicly available GWAS results to survey genetic correlations between Parkinson's disease and other phenotypes showing significant correlations with smoking, education, and brain morphology. Subsequent analyses using Mendelian randomisation methods showed probable causal links between increased cognitive performance and Parkinson's disease risk on a genetic level.

Implications of all the available evidence

Using updated heritability estimates and risk predictions, we took preliminary steps on a long path to early detection. In future studies, combining genetic and clinico-demographic risk factors could lead to earlier detection and refined diagnostics, which could help improve clinical trials. The generation of copious amounts of public summary statistics created by this effort relating to both the GWAS and subsequent analyses of gene expression and methylation patterns might be of use to investigators planning follow-up functional studies in stem cells or other cellular screens. This information would allow researchers to prioritise targets more efficiently, using our data as additional evidence. We hope our findings might have some downstream clinical impact in the future, such as improved patient stratification for clinical trials and genetically informed drug targets.

studies of sporadic Parkinson's disease have altered the foundational view of disease causes.

We aimed to undertake the largest-to-date GWAS for Parkinson's disease to identify novel mechanistic candidate genes for this disease. We will further assess the function of potential risk genes, estimate Parkinson's disease heritability, and develop a model to predict the proportion of this heritability. Our final goal is to identify potential Parkinson's disease biomarkers and risk factors.

Methods

Study design

The work flow and rationale of our study is shown in figure 1. Three sources of data were used for discovery analyses, these include three previously published GWAS studies,^{4,6} 13 new datasets (figure 1), and proxy-case data from the UK Biobank. Previous studies comprised summary statistics published in Nalls and colleagues,⁶ GWAS summary statistics from the 23andMe Web-Based Study of Parkinson's Disease by Chang and colleagues,⁴ and the publicly available NeuroX dataset from the International Parkinson's Disease Genomics Consortium previously

used as a replication sample.⁶ These cohorts have been reported in detail, but in brief represent all European ancestry Parkinson's disease case-control GWAS studies available for collaboration.^{4,6} We included 13 new case-control sample series for meta-analyses through either publicly available data or collaborations (appendix pp 1–3, 22). All samples from the 13 new datasets underwent similar standardised quality control for inclusion, mirroring that of previous studies. We attempted to generate summary statistics for GWAS meta-analyses as uniformly as possible. This analysis used fixed-effects meta-analyses as implemented in METAL⁷ to combine summary statistics across all sources.

Conditional joint analysis

To nominate variants of interest, we used a conditional and joint analysis strategy to algorithmically identify variants that best account for the heritable variation within and across loci.⁸ Additional analyses were used to further scrutinise putative associated variants and account for possible differential linkage disequilibrium (LD) signatures, including using the massive single site reference

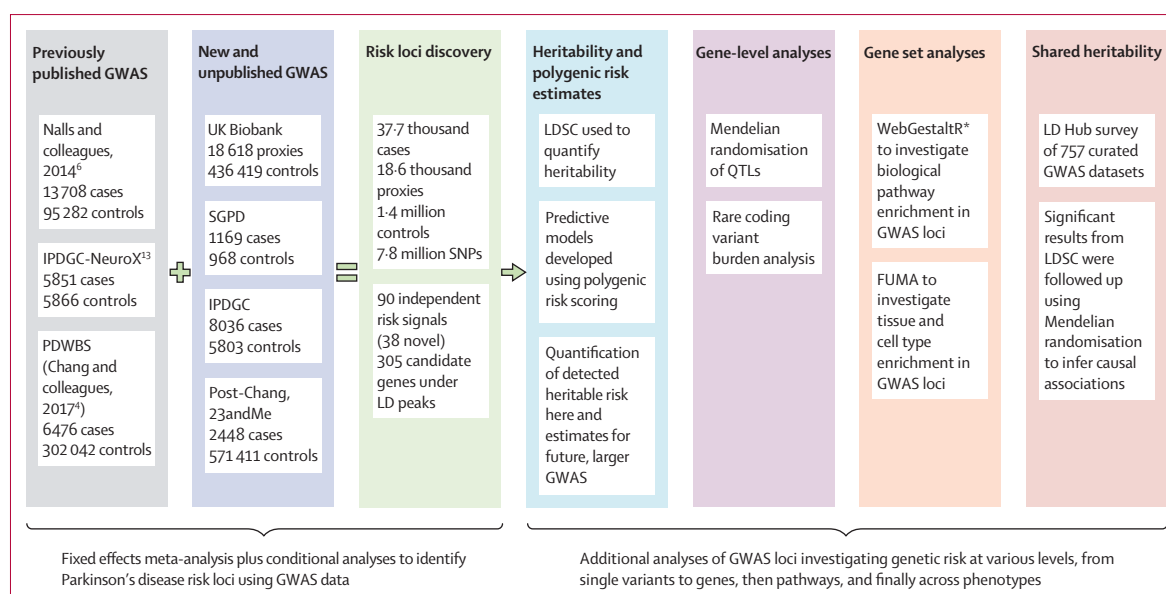


Figure 1: Workflow and rationale summary

GWAS=genome-wide association studies. SNPs=single nucleotide polymorphisms. IPDGC=International Parkinson's Disease Genomics Consortium. PDWBS=Parkinson's disease web based study. SGPD= Systems genomics of Parkinson's disease consortium. LD=linkage disequilibrium. LDSC=linkage disequilibrium score regression. QTLs=quantitative trait loci. FUMA=functional mapping and annotation of genetic associations platform. ^{*}WEB-based Gene Set Analysis Toolkit.

data from 23andMe in further conditional analyses. If a variant nominated during the conditional and joint analysis strategy phase of analysis was greater than 1 Mb from any of the genome-wide significant loci nominated in Chang and colleagues,⁴ we considered this to be a novel risk variant. We defined nominated risk variants as from a single locus if they were within 250 kb of each other. We instituted two filters after fixed-effects and conditional and joint analyses, excluding variants that had a random-effects *p* value across all datasets more than 4.67×10^{-4} and a conditional analysis *p* value more than 4.67×10^{-4} using participant level 23andMe genotype data. Please see appendix (pp 4–5, 22) summarising all variants nominated.

Heritability estimates and extant genetic risk

We used the R package PRSice2 for risk profiling,⁹ which calculates polygenic risk score (PRS) profiling in the standard weighted allele dose manner.^{4,6,10–12} In addition, PRSice incorporates permutation testing, in which case and control labels are swapped in the withheld samples to generate an empirical *p* value. This workflow identifies the best *p* value thresholds for variant inclusion while doing LD pruning. In many cases this best *p* value threshold for PRS construction does not meet what is commonly regarded as genome-wide significance.

A two-stage design was also used, training on the largest single array study (NeuroX-dbGaP¹³) and then tested on the second largest study (Harvard Biomarker Study) using the same array. These two targeted array studies were chosen for three reasons: precedent in the previous publications in which the NeuroX-dbGaP dataset was used in

PRS; direct genotyping of larger effect rare variants in *GBA* and *LRRK2*; and participant level genotypes for these datasets are publicly available.

To calculate heritability in clinically defined Parkinson's disease datasets, we used LD score regression employing the LD references for Europeans provided with the software.¹⁴ This workflow was also repeated on a per cohort level (appendix pp 10–11).

Functional causal inferences via quantitative trait loci

We used Mendelian randomisation to test whether changes in DNA methylation or RNA expression of genes physically proximal to significant Parkinson's disease risk loci were causally related to Parkinson's disease risk. To nominate genes of interest for Mendelian randomisation analyses, we took our putative 90 loci in the large LD reference used for the conditional and joint phase of analysis and identified SNPs in LD with our SNPs at an r^2 0.5 within 1 Mb (appendix pp 9–10). Mendelian randomisation was used by integrating discovery phase summary statistics with quantitative trait loci (QTL) association summary statistics across well curated methylation and expression datasets. We used the curated versions of Qi and colleagues brain methylation and expression summary statistics (multi-study and multi-tissue meta-analysis), with a specific focus on substantia nigra data (GTEx). We also made use of the blood expression data from Vösa and colleagues from 2018 (eQTLGen).^{15–19} For all QTL analyses, we used the multi-SNP summary-based Mendelian randomisation method as a framework to do Mendelian randomisation. All Mendelian randomisation effect estimates are reported on the scale of an SD increase in the exposure variable relating

Johns Hopkins University Medical Center, Baltimore, MD, USA (S W Scholz); Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, Spain (J A Botia); Institut national de la santé et de la recherche médicale Unité mixte de recherche 1220, Toulouse, France (M Martinez PhD); Paul Sabatier University, Toulouse, France (M Martinez PhD); Institut national de la santé et de la recherche médicale U1127, CNRS UMR 7225, Paris, France (J-C Corvol PhD, S Lesage PhD, A Brice MD); Sorbonne Université centre national de la recherche médicale, unité mixte de recherche 1127, Paris, France (J-C Corvol PhD, S Lesage PhD, A Brice MD); Assistance Publique Hôpitaux de Paris, Paris, France (J-C Corvol PhD, S Lesage PhD, A Brice MD); Institut du Cerveau et de la Moelle épinière, Paris, France (J-C Corvol, S Lesage, A Brice) Clinical Neurosciences, Neurology, University of Helsinki, Helsinki, Finland (P Tienari MD); Helsinki University Hospital, Helsinki, Finland (P Tienari); Institute of Clinical Medicine (M Toft) and Jebsen Centre for Psychosis Research (O A Andreassen MD), University of Oslo, Oslo, Norway; and Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX, USA (J M Shulman)

Correspondence to: Dr Mike A Nalls, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD 20892, USA mike@datatecnica.com
See Online for appendix
For WebGestaltR see <http://www.webgestalt.org/>
For the conditional and joint analysis strategy see <http://cnsgenomics.com/software/gcta/>
For the Harvard Biomarker Study see <https://neurodiscovery.harvard.edu/biomarkers-discovery>
For GTEx see <http://cnsgenomics.com/software/smr/#DataResource>
For eQTLGen see <http://www.eqtngen.org>

to a similar change in Parkinson's disease risk. Simply, these Mendelian randomisation analyses compare the local polygenic risk of an exposure (methylation or expression) to similar polygenic risk in an outcome (Parkinson's disease). This method infers causal associations under the assumption that there is no intermediate confounder associated with both parameters and that the association is not simply due to LD.

To further investigate expression enrichment across cell types in Parkinson's disease, we integrated GWAS summary statistics with expression and network data from the Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) webserver.²⁰

Rare coding variant burden tests

A uniformly quality controlled and imputed dataset from the International Parkinson's Disease Genomics Consortium was used to do burden tests for all rarer coding variants successfully imputed in a mean of 85% of the sample series (17188 cases and 22875 controls). These analyses include all variants at a hard call threshold of imputation quality more than 0.8. After annotation with ANNOVAR, we had 37503 exonic coding variants (non-synonymous, stop, or splicing) at minor allele frequency less than 5% and a subset of 29016 at minor allele frequency less than 1%.²¹ For inclusion in this phase, a gene had to contain at least two coding variants. After assembling this subset of 113 testable genes, we used the optimised sequence kernel association test to generate summary statistics at maximum minor allele frequencies of 1% and 5%.²²

LD score regression and causal inference

To investigate correlations of Parkinson's disease genetics with that of multiple traits and diseases, we used bivariate LD score regression.¹⁴ These analyses were done with data from the 757 GWAS available via LD Hub and biomarker GWAS summary statistics^{23–25} on c-reactive protein and cytokine measures; LD Hub was accessed on June 20, 2018, (version 1.2.0).^{23–25} The p values from the bivariate LD score regression were adjusted for false discovery rate to account for multiple testing. Traits showing significant genetic correlations with Parkinson's disease were analysed with Mendelian randomisation methods. We excluded the UK Biobank data when a nominated trait was from summary statistics derived from the UK Biobank or if the UK Biobank was included as part of a meta-analysis.

When complete GWAS summary statistics were available for traits of interest (relating to smoking and education), we used the more powerful bidirectional generalised summary-data-based Mendelian randomisation. We analysed GWAS summary statistics for smoking initiation (453693 records from a self-report survey with 208988 regular smokers and 244705 never regular smokers) and current smoking within the UK Biobank, current smoking contrasted 47419 current smokers versus 244705 never regular smokers. The same analysis was

done incorporating recent GWAS data regarding educational attainment (N=766345) from self-report in the UK and cognitive performance (N=257828) as measured by the g composite score.²⁶ These data were analysed using methods to mirror that of the UK Biobank Parkinson's disease GWAS dataset. Combined left and right putamen volume from a T2 weighted MRI GWAS was available from Oxford Brain Imaging Genetics server (accessed Dec 28, 2018).²⁷ All Mendelian randomisation analyses included GWAS on the scale of ten thousand samples and overcame the considerable power demands of the methods. For additional quality control, method details, and ancillary results, see the appendix.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all of the data and the final responsibility to submit for publication.

Results

Our study took place between Oct 1, 2017, and Aug 9, 2018. To maximise our power for locus discovery we used a single stage design, meta-analysing all available GWAS summary statistics. Supporting this design, we found strong genetic correlations using Parkinson's disease cases ascertained by clinicians compared with 23andMe self-reported cases (genetic correlation from LD score regression [rG] 0.85, SE 0.06) and UK Biobank proxy cases (rG 0.84, SE 0.134).

We identified 90 independent genome-wide significant association signals through our analyses of 37688 cases, 18618 UK Biobank proxy-cases, and 1417791 controls at 7784415 SNPs (figure 2, table 1, appendix pp 1–6). Of these, 38 signals are newly identified and more than 1 Mb from loci described previously (appendix p 4).⁴

We detected ten loci containing more than one independent risk signal (22 risk SNPs in total across these loci), of which nine had been identified by previous GWAS, including multi-signal loci in the vicinity of *GBA*, *NUCKS1* and *RAB29*, *GAK* and *TMEM175*, *SNCA*, and *LRRK2*. The novel multi-signal locus comprised independent risk variants rs2269906 (*UBTF* and *GRN*) and rs850738 (*FAM171A2*). Detailed summary statistics on all nominated loci can be found in the appendix (pp 4–6), including variants filtered out during additional quality control.

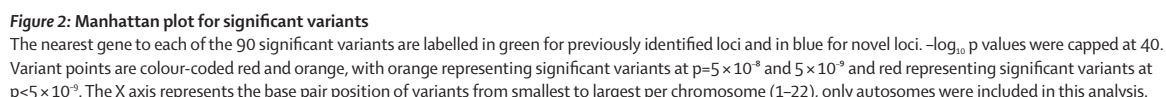
To quantify how much of the genetic liability we have explained and what direction to take with future Parkinson's disease GWAS, we generated updated heritability estimates and PRS. Using LD score regression on a meta-analysis of all 11 clinically ascertained datasets from our GWAS, we estimated the liability-scale heritability of Parkinson's disease as 0.22 (95% CI 0.18–0.26), only slightly lower than a previous estimate derived using genome-wide complex trait analysis (0.27, 0.17–0.38).^{14,28,29}

For the Oxford Brain Imaging Genetics server see <http://big.stats.ox.ac.uk/>

For the FUMA webserver version 1.3.1 see <https://fuma.ctglab.nl/>

For ANNOVAR see <http://annovar.openbioinformatics.org/en/latest/>

For LD Hub see <http://ldsc.broadinstitute.org/ldhub/>



Variants with p values in the range of 5×10^{-8} to 1.35×10^{-3} (used in the 1805 variant PRS) were rarer and had smaller effect estimates than variants reaching genome-wide significance. These sub-significant variants had a median minor allele frequency of 21.3% and a median effect estimate (absolute value of the log odds ratio of the SNP parameter from regression) of 0.047. Genome-wide significant risk variants were more common with a median minor allele frequency of 25.1%, and had a median effect estimate of 0.081. Here we assume that the lower minor

Of the 305 genes under LD peaks around our risk variants of interest, 237 were possibly associated with at least one QTL in public reference datasets and were therefore testable via summary-based Mendelian randomisation (appendix pp 8–9). The expression or methylation of 151 (64%) of these 237 genes was significantly associated with a possible causal change in Parkinson's disease risk.

Chromosome	Base pair position	Nearest gene	Effect allele	Other allele	Effect allele frequency	OR (95% CI)	Regression coefficient (β)	Standard error of β	p value, fixed-effects	p value, conditional joint analysis approach	p value, conditional effects	P, %
rs658353	1 161469054	FCGR2A	C	G	0.501	1.07 (1.05-1.09)	0.065	0.009	6.10 × 10 ⁻¹²	4.69 × 10 ⁻¹²	1.38 × 10 ⁻⁵	3.71 × 10 ⁻⁵ 40.2%
rs11578699	1 171719769	VAMP4	T	C	0.195	0.93 (0.91-0.95)	-0.070	0.012	4.47 × 10 ⁻⁹	4.45 × 10 ⁻⁹	2.63 × 10 ⁻³	1.09 × 10 ⁻⁷ 5.1%
rs76116224	2 18147848	KCN5	A	T	0.904	1.12 (1.08-1.16)	0.110	0.019	1.27 × 10 ⁻⁸	1.27 × 10 ⁻⁸	3.75 × 10 ⁻⁷	1.27 × 10 ⁻⁸ 0
rs2042477	2 96000943	KCNIP3	A	T	0.242	0.94 (0.92-0.96)	-0.066	0.012	1.38 × 10 ⁻⁸	1.48 × 10 ⁻⁸	3.49 × 10 ⁻⁵	1.38 × 10 ⁻⁸ 0
rs6808178	3 28705690	LINC00693	T	C	0.379	1.07 (1.05-1.09)	0.066	0.010	8.09 × 10 ⁻¹²	7.18 × 10 ⁻¹²	8.84 × 10 ⁻⁵	8.09 × 10 ⁻¹² 0
rs55961674	3 122196892	KPNA1	T	C	0.172	1.09 (1.06-1.12)	0.086	0.013	9.98 × 10 ⁻¹²	8.30 × 10 ⁻¹²	2.80 × 10 ⁻⁶	9.98 × 10 ⁻¹² 0
rs11707416	3 151108965	MED12L	A	T	0.367	0.94 (0.92-0.96)	-0.063	0.010	1.13 × 10 ⁻¹⁰	1.02 × 10 ⁻¹⁰	2.66 × 10 ⁻⁴	1.77 × 10 ⁻⁷ 10.9%
rs1450522	3 161077630	SPYSSB	A	G	0.674	0.94 (0.92-0.96)	-0.062	0.010	5.01 × 10 ⁻¹⁰	4.90 × 10 ⁻¹⁰	3.51 × 10 ⁻⁴	2.27 × 10 ⁻⁵ 24.6%
rs34025766	4 17968811	LCORL	A	T	0.159	0.92 (0.90-0.94)	-0.084	0.013	2.87 × 10 ⁻¹⁰	2.82 × 10 ⁻¹⁰	7.43 × 10 ⁻⁵	2.87 × 10 ⁻¹⁰ 0
rs62333164	4 170583157	CLCN3	A	G	0.326	0.94 (0.92-0.96)	-0.064	0.010	2.00 × 10 ⁻¹⁰	1.77 × 10 ⁻¹⁰	5.10 × 10 ⁻⁵	2.17 × 10 ⁻⁵ 21.3%
rs26431	5 102365794	PAM	C	G	0.703	1.06 (1.04-1.09)	0.062	0.010	1.57 × 10 ⁻⁹	1.65 × 10 ⁻⁹	6.00 × 10 ⁻³	2.36 × 10 ⁻⁷ 7.9%
rs11950533	5 134199105	C5orf24	A	C	0.102	0.91 (0.88-0.94)	-0.092	0.016	7.16 × 10 ⁻⁹	6.73 × 10 ⁻⁹	5.08 × 10 ⁻⁴	2.68 × 10 ⁻⁸ 1.9%
rs9761484	6 30108683	TRIM40	T	C	0.245	0.94 (0.92-0.96)	-0.064	0.011	1.62 × 10 ⁻⁸	1.43 × 10 ⁻⁸	1.26 × 10 ⁻⁶	1.62 × 10 ⁻⁸ 0
rs12528068	6 72487762	RIMS1	T	C	0.284	1.07 (1.05-1.09)	0.066	0.010	1.63 × 10 ⁻¹⁰	1.79 × 10 ⁻¹⁰	9.80 × 10 ⁻⁶	1.63 × 10 ⁻¹⁰ 0
rs997368	6 112243291	FYN	A	G	0.805	1.07 (1.05-1.10)	0.071	0.012	1.84 × 10 ⁻⁹	1.97 × 10 ⁻⁹	2.61 × 10 ⁻⁵	1.84 × 10 ⁻⁹ 0
rs75859381	6 133210361	RP512	T	C	0.967	0.80 (0.75-0.86)	-0.221	0.034	1.04 × 10 ⁻¹⁰	9.67 × 10 ⁻¹¹	1.09 × 10 ⁻⁶	1.04 × 10 ⁻¹⁰ 0
rs76949143	7 66009851	GS1-124K5-11	A	T	0.051	0.87 (0.82-0.91)	-0.143	0.025	1.43 × 10 ⁻⁸	1.51 × 10 ⁻⁸	5.47 × 10 ⁻⁹	2.04 × 10 ⁻⁶ 12.3%
rs2086641	8 130901909	FAM49B	T	C	0.723	0.94 (0.92-0.96)	-0.061	0.011	1.81 × 10 ⁻⁸	1.57 × 10 ⁻⁸	6.07 × 10 ⁻⁶	1.81 × 10 ⁻⁸ 0
rs6476434	9 34046391	UBAP2	T	C	0.734	0.94 (0.92-0.96)	-0.062	0.011	6.58 × 10 ⁻⁹	6.56 × 10 ⁻⁹	2.74 × 10 ⁻⁴	6.58 × 10 ⁻⁹ 0
rs10748818	10 104015279	GBF1	A	G	0.851	0.92 (0.90-0.95)	-0.079	0.013	1.05 × 10 ⁻⁹	1.23 × 10 ⁻⁹	7.47 × 10 ⁻⁶	1.05 × 10 ⁻⁹ 0
rs7938782	11 10558777	RNF141	A	G	0.878	1.09 (1.06-1.12)	0.087	0.015	2.12 × 10 ⁻⁹	1.97 × 10 ⁻⁹	2.17 × 10 ⁻⁷	2.12 × 10 ⁻⁹ 0
rs7134559	12 46419086	SCAF1	T	C	0.404	0.95 (0.93-0.97)	-0.054	0.010	3.96 × 10 ⁻⁸	3.80 × 10 ⁻⁸	1.69 × 10 ⁻²	1.84 × 10 ⁻⁵ 25.2%
rs11610045	12 133063768	FBRSL1	A	G	0.490	1.06 (1.04-1.08)	0.060	0.009	1.77 × 10 ⁻¹⁰	1.62 × 10 ⁻¹⁰	3.57 × 10 ⁻⁵	8.79 × 10 ⁻⁷ 19.5%
rs9568188	13 49927732	CAB39L	T	C	0.740	1.06 (1.04-1.09)	0.062	0.011	1.15 × 10 ⁻⁸	1.11 × 10 ⁻⁸	4.29 × 10 ⁻⁶	2.46 × 10 ⁻⁴ 21.4%
rs4771268	13 97865021	MBNL2	T	C	0.230	1.07 (1.05-1.09)	0.068	0.011	1.45 × 10 ⁻⁹	1.67 × 10 ⁻⁹	1.41 × 10 ⁻⁴	1.45 × 10 ⁻⁹ 0
rs12147950	14 37989270	MIPOL1	T	C	0.438	0.95 (0.93-0.97)	-0.053	0.010	3.54 × 10 ⁻⁸	3.58 × 10 ⁻⁸	1.06 × 10 ⁻³	3.54 × 10 ⁻⁸ 0
rs3742785	14 75373034	RP56KL1	A	C	0.787	1.07 (1.05-1.10)	0.071	0.012	1.92 × 10 ⁻⁹	2.08 × 10 ⁻⁹	2.22 × 10 ⁻⁶	8.18 × 10 ⁻⁶ 24.8%
rs2904880	16 28944396	CD19	C	G	0.309	0.94 (0.92-0.96)	-0.065	0.011	7.87 × 10 ⁻¹⁰	8.68 × 10 ⁻¹⁰	1.39 × 10 ⁻⁵	7.87 × 10 ⁻¹⁰ 0
rs6500328	16 50736656	NOD2	A	G	0.599	1.06 (1.04-1.08)	0.059	0.010	1.82 × 10 ⁻⁹	1.53 × 10 ⁻⁹	1.43 × 10 ⁻³	1.82 × 10 ⁻⁹ 0
rs12600861	17 7355621	CHRNA1	A	C	0.648	0.95 (0.93-0.96)	-0.057	0.010	1.01 × 10 ⁻⁸	1.15 × 10 ⁻⁸	5.10 × 10 ⁻³	1.01 × 10 ⁻⁸ 0
rs2269906	17 42294337	UBTF	A	C	0.653	1.07 (1.04-1.09)	0.063	0.010	6.24 × 10 ⁻¹⁰	8.63 × 10 ⁻⁹	1.17 × 10 ⁻⁵	6.24 × 10 ⁻¹⁰ 0
rs850738	17 42434630	FAM171A2	A	G	0.606	0.93 (0.91-0.95)	-0.071	0.011	1.29 × 10 ⁻¹¹	3.55 × 10 ⁻¹⁰	4.18 × 10 ⁻⁴	2.17 × 10 ⁻⁷ 17.0%
rs61169879	17 59917366	BRP1	T	C	0.164	1.09 (1.06-1.11)	0.082	0.013	9.28 × 10 ⁻¹⁰	9.40 × 10 ⁻¹⁰	9.07 × 10 ⁻⁷	6.21 × 10 ⁻⁶ 16.4%
rs666463	17 76425480	DNAH17	A	T	0.833	1.08 (1.05-1.11)	0.076	0.013	3.20 × 10 ⁻⁹	2.90 × 10 ⁻⁹	1.62 × 10 ⁻⁵	4.17 × 10 ⁻⁴ 41.0%
rs1941685	18 31304318	ASXL3	T	G	0.498	1.05 (1.04-1.07)	0.053	0.009	1.69 × 10 ⁻⁸	1.61 × 10 ⁻⁸	1.64 × 10 ⁻⁸	1.69 × 10 ⁻⁸ 0
rs8087969	18 48683589	MEX3C	T	G	0.550	0.94 (0.93-0.96)	-0.058	0.010	1.41 × 10 ⁻⁸	1.46 × 10 ⁻⁸	1.09 × 10 ⁻⁴	1.41 × 10 ⁻⁸ 0
rs77351827	20 6006041	CRLS1	T	C	0.128	1.08 (1.05-1.11)	0.080	0.014	8.87 × 10 ⁻⁹	7.94 × 10 ⁻⁹	1.84 × 10 ⁻⁵	4.38 × 10 ⁻⁷ 11.2%
rs2248244	21 38852361	DYRK1A	A	G	0.283	1.07 (1.05-1.10)	0.071	0.011	2.74 × 10 ⁻¹¹	2.51 × 10 ⁻¹¹	6.31 × 10 ⁻⁵	8.78 × 10 ⁻⁶ 34.3%

Summary statistics for 38 novel genome-wide significant Parkinson's disease variants using data from all available genome-wide association studies.

Table 1. Novel loci associated with Parkinson's disease

	Max p value threshold	Pseudo r^2 from PRS*	Regression coefficient (β)	SE of β	p value	OR, highest quartile PRS	95% CI, highest quartile PRS	SNPs (N)	Samples (N)	Area under the curve	95% CI (Delong)	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Balanced accuracy
Training dataset: IPDGC, Neurox	1.35×10^{-3}	0.029	0.553	0.022	8.99×10^{-18}	3.74	3.35–4.18	1809	11 243	0.640	0.630–0.650	0.569	0.632	0.591	0.611	0.601
Validation dataset: HBS	4.00×10^{-2}	0.054	0.709	0.072	8.28×10^{-23}	6.25	4.26–9.28	1805	999	0.692	0.660–0.725	0.628	0.686	0.691	0.623	0.657

Estimates of performance for predictive models including single study estimates, estimates from meta-analyses across studies, as well as a two stage design. Here the best p value threshold column denotes the filtering value for SNP inclusion to achieve the maximal pseudo (Nagelkerke's) r^2 . The OR column is the exponent of the regression coefficient (β) from logistic regression of the PRS on case status, with the SE representing the precision of these estimates. These same metrics are derived across array types and datasets using random-effects meta-analyses. The area under the curve is included as the most common metric for predictive model performance. PRS=polygenic risk score. OR=odds ratio. IPDGC=International Parkinson's Disease Genomics Consortium. HBS=Harvard Biomarker Study. * r^2 approximation adjusted for an estimated prevalence of 0.5%, equivalent to roughly half of the unadjusted r^2 estimates for the PRS. All calculations and reported statistics include only the PRS and no other parameters after adjusting for principal components 1–5, age, and sex at variant selection in the NeuroX-dbGap dataset.¹³ SNP=single nucleotide polymorphism.

Table 2: Summary of genetic predictive model performance

Of the 90 Parkinson's disease GWAS risk variants, 70 were in loci containing at least one of these putatively causal genes after multiple test correction (table 3). For 53 (76%) of these 70 Parkinson's disease GWAS hits, the gene nearest to the most significant SNP was a putatively causal gene (appendix pp 4–6). Most loci tested contained multiple putatively causal genes. The nearest putatively causal gene to the rs850738 and *FAM171A2* GWAS risk signal is *GRN*, a gene known to be associated with frontotemporal dementia.³² Mutations in *GRN* have also been shown to be connected with another lysosomal storage disorder, neuronal ceroid lipofuscinosis.³³

As an orthogonal approach for nominating genes under GWAS peaks, we did rare coding variant burden analyses. We did kernel-based burden tests on the 113 genes of the 305 under our GWAS peaks that contained two or more rare coding variants (minor allele frequency <5% or <1%). After Bonferroni correction for 113 genes, we identified seven significant genes: *LRRK2*, *GBA*, *CATSPER3* (rs11950533 and *C5orf24* locus), *LAMB2* (rs12497850 and *IP6K2* locus), *LOC442028* (rs2042477 and *KCNIP3* locus), *NFKB2* (rs10748818 and *GBF1* locus), and *SCARB2* (rs6825004 locus). These results suggest that some of the risk associated with these loci might be due to rare coding variants or that these are pleomorphic risk loci. The *LRRK2* and *NFKB2* associations at minor allele frequency less than 1% remained significant after correcting for the approximate 20 000 genes in the human genome ($p=2.15 \times 10^{-12}$ for *LRRK2* and $p=4.02 \times 10^{-7}$ for *NFKB2*, appendix pp 8–9).

We tested whether genes of interest were enriched in 10651 biological pathways (from gene ontology annotations) using FUMA.^{20,34} We found ten significantly enriched pathways (false discovery rate-adjusted $p<0.05$, appendix pp 9–10), including four related to vacuolar function and three related to known drug targets (calcium transporters, ikeda_mir1_targets_dn and ikeda_mir30_targets_up; kinase signalling, kim_pten_targets_dn³⁵). At least three candidate genes within novel loci are involved in lysosomal storage disorders (*GUSB*, *GRN*, and *NEU1*), a pathway of keen interest in Parkinson's disease.³⁶ Our GWAS results also include candidate genes *VAMP4* and *NOD2* from the endocytic pathway.³⁷

To establish the tissues and cell types most relevant to the causes of Parkinson's disease using FUMA,^{20,34} we tested whether the genes highlighted by our Parkinson's disease GWAS were enriched for expression in 53 tissues. We found 13 significant tissues, all of which were brain-derived (appendix p 17), in contrast to what has been seen in Alzheimer's disease, which shows a strong bias towards blood, spleen, lung, and microglial enrichments.³⁸ To further disentangle the enrichment in brain tissues, we tested whether our Parkinson's disease GWAS genes were enriched for expression in 88 brain cell types using single cell RNA sequencing reference data from mouse brains using DropViz.³⁹ After false discovery rate correction we found seven significant brain cell types, all of which were

For DropViz see <http://dropviz.org>

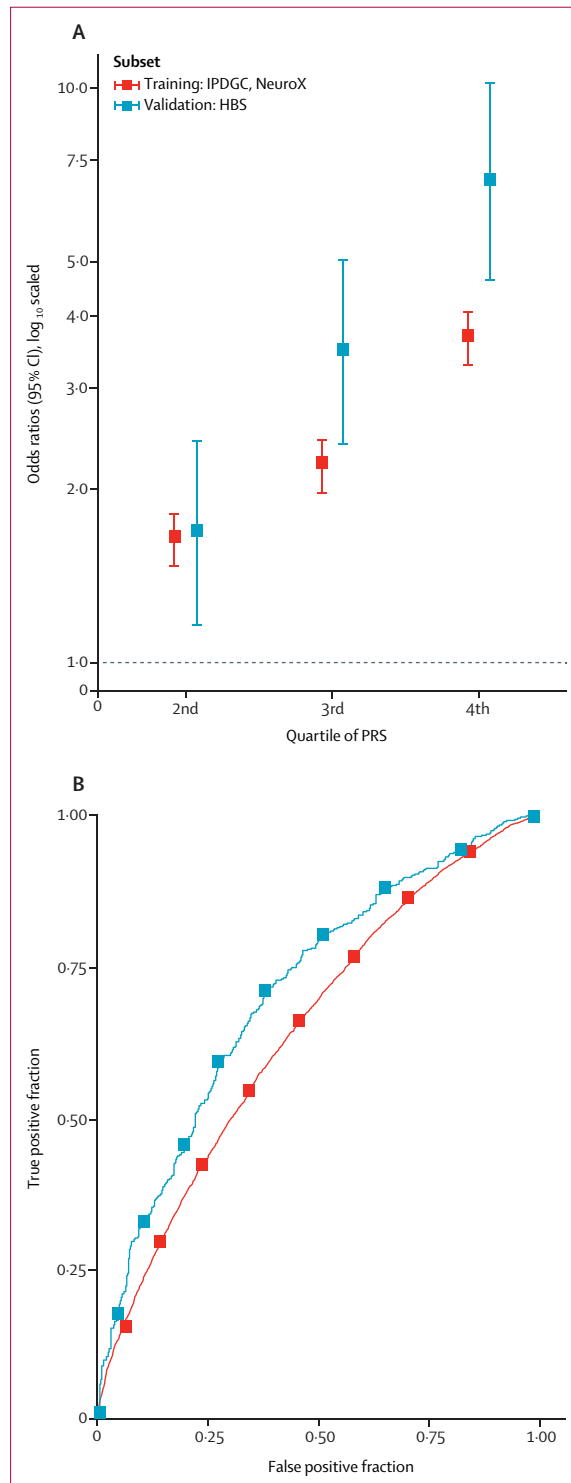


Figure 3: Predictive model

The odds ratio of developing Parkinson's disease for each quartile of PRS compared with the lowest quartile of genetic risk (A). PRS receiver-operator curves for the more inclusive 1805 variant PRS in the validation dataset and in the corresponding training dataset that was used for PRS thresholding and single nucleotide polymorphism selection (B). PRS=polygenic risk score. IPDGC=International Parkinson's Disease Genomics Consortium. HBS=Harvard Biomarker Study.

neuronal (appendix p 17). The strongest enrichment was for neurons in the substantia nigra at $p=1.0 \times 10^{-6}$, with additional significant results at $p < 5.0 \times 10^{-4}$ for the globus pallidus, thalamus, posterior cortex, frontal cortex, hippocampus, and entopeduncular nucleus.

Next, we used cross-trait genetic correlation and Mendelian randomisation to identify possible Parkinson's disease biomarkers and risk factors by comparing with 757 other GWAS datasets curated by LD Hub.²⁵ We found four significant genetic correlations (false discovery rate-adjusted $p < 0.05$, appendix pp 10–11) including positive correlations with intracranial volume ($p=0.0035$) and putamen volume ($p=0.024$),⁴⁰ and negative correlations with current tobacco use ($p=0.024$) and academic qualifications ($p=0.038$; eg, National Vocational Qualifications, Higher National Diploma, Higher National Certificate, or equivalent).⁴¹ The negative association with an individual's academic qualifications suggests that individuals without a college education might be at less risk of Parkinson's disease. The correlation between Parkinson's disease and smoking status might not be independent from the correlation between Parkinson's disease and education as smoking status and years of education were significantly correlated.⁴²

We used Mendelian randomisation to assess whether there was evidence of a causal relationship between Parkinson's disease and five phenotypes related to academic qualifications, smoking, and brain volumes described above (appendix pp 19–21). Cognitive performance had a large, significant causal effect on Parkinson's disease risk (Mendelian randomisation effect 0.213, SE 0.041; Bonferroni-adjusted $p=8.00 \times 10^{-7}$), whereas Parkinson's disease risk did not have a significant causal effect on cognitive performance (Bonferroni-adjusted $p=0.125$). Educational attainment also had a significant causal effect on Parkinson's disease risk (Mendelian randomisation effect 0.162, SE 0.040, Bonferroni-adjusted $p=2.06 \times 10^{-4}$), and Parkinson's disease risk also had a weak but significant causal effect on educational attainment (Mendelian randomisation effect 0.007, SE 0.002, Bonferroni-adjusted $p=7.45 \times 10^{-3}$). There was no significant causal relationship between Parkinson's disease and current smoking status in forward analysis (Mendelian randomisation effect -0.069 , SE 0.031; Bonferroni-adjusted $p=0.125$) or reverse analysis (Mendelian randomisation effect 0.004, SE 0.010, Bonferroni-adjusted $p=1.00$). Smoking initiation (the act of ever starting smoking) did not have a causal effect on Parkinson's disease risk (Mendelian randomisation effect -0.063 , SE 0.034, Bonferroni-adjusted $p=0.32$), whereas Parkinson's disease had a small, but significantly positive causal effect on smoking initiation (Mendelian randomisation effect 0.027, SE 0.006, Bonferroni-adjusted $p=1.62 \times 10^{-5}$). Intracranial volume could not be tested because the GWAS data (available from Oxford Brain Imaging Genetics server) did not contain any genome-wide significant risk variants. No significant causal relationship was observed between Parkinson's disease and

	Probe	Chromosome	Probe, base pair	Top SNP, base pair	Top SNP	SNPs (N)	Effect	SE	p value	Benferroni adjusted p value
VAMP4 ¹⁹	ENSG000000117533	1	171 690 343	171 717 417	rs10913587	98	-0.272	0.05	5.67 × 10 ⁻⁷	1.19 × 10 ⁻⁴
KCNIP3 ¹⁶	ENSG000000115041	2	96 007 438	95 989 766	rs3772034	14	-0.161	0.04	1.12 × 10 ⁻⁵	1.15 × 10 ⁻³
MAP4K4 ¹⁹	ENSG000000071054	2	102 410 880	102 338 377	rs6733355	3	1.119	0.24	2.32 × 10 ⁻⁶	4.87 × 10 ⁻⁴
TMEM163 ¹⁶	ENSG000000152128	2	135 344 950	135 248 544	rs598668	28	0.074	0.02	3.55 × 10 ⁻⁷	3.65 × 10 ⁻⁵
KPNA1 ¹⁹	ENSG000000114030	3	122 187 294	122 201 610	rs73190142	110	0.310	0.05	1.56 × 10 ⁻⁶	3.28 × 10 ⁻⁴
GAK ⁶	ENSG00000178950	4	884 612	906 131	rs11248057	1	0.508	0.10	7.47 × 10 ⁻⁷	7.69 × 10 ⁻⁵
CAMK2D ¹⁹	ENSG000000145349	4	114 527 635	114 730 260	rs115671064	146	-0.006	0.05	5.74 × 10 ⁻⁶	1.21 × 10 ⁻³
PAM ¹⁹	ENSG000000145730	5	102 228 247	102 118 633	rs2432162	679	-0.031	0.01	2.08 × 10 ⁻⁶	4.36 × 10 ⁻⁴
LOC100131289 ¹⁶	cg21339923	6	27 636 378	27 636 378	rs78149975	2	-0.094	0.02	1.53 × 10 ⁻⁶	3.06 × 10 ⁻⁴
TRIM40 ¹⁶	cg01641092	6	30 094 300	30 094 315	rs9261443	8	0.072	0.01	6.15 × 10 ⁻⁶	1.23 × 10 ⁻³
HLA-DRB5 ¹⁶	cg26036029	6	32 552 443	32 570 311	rs34039593	8	-0.153	0.02	7.53 × 10 ⁻¹⁰	1.51 × 10 ⁻⁷
GPNMB ¹⁶	ENSG000000136235	7	23 295 156	23 294 668	rs858274	74	0.090	0.01	2.73 × 10 ⁻²¹	2.81 × 10 ⁻¹⁹
CTSB ¹⁶	ENSG000000164733	8	11 713 495	11 699 279	rs4631423	33	-0.150	0.04	4.37 × 10 ⁻⁹	4.50 × 10 ⁻⁷
BIN3 ¹⁶	ENSG000000147439	8	22 502 296	22 456 517	rs71513892	32	0.046	0.01	1.43 × 10 ⁻⁶	1.48 × 10 ⁻⁴
SH3GL2 ¹⁶	ENSG000000107295	9	17 688 103	17 684 784	rs10756899	15	0.252	0.05	5.83 × 10 ⁻⁸	6.00 × 10 ⁻⁶
ITGA8 ¹⁶	ENSG000000077943	10	15 659 036	15 548 925	rs7910668	6	-0.201	0.05	6.13 × 10 ⁻⁵	6.32 × 10 ⁻³
RNF141 ¹⁹	ENSG000000110315	11	10 548 001	10 553 355	rs4910153	120	-0.054	0.05	6.25 × 10 ⁻⁷	1.31 × 10 ⁻⁴
IGSF9B ¹⁶	cg25790212	11	133 800 774	133 800 477	rs11223626	1	-0.172	0.04	3.24 × 10 ⁻⁶	6.48 × 10 ⁻⁴
FBRSL1 ¹⁶	cg03621470	12	133 137 479	133 138 334	rs10781619	16	-0.057	0.01	6.35 × 10 ⁻⁵	1.27 × 10 ⁻²
CAB39L ¹⁶	ENSG000000102547	13	49 950 524	49 918 175	rs35214871	30	0.097	0.02	3.51 × 10 ⁻⁸	3.62 × 10 ⁻⁶
GCH1 ¹⁶	ENSG000000131979	14	55 339 148	55 348 837	rs3825611	6	0.113	0.03	2.76 × 10 ⁻⁴	2.85 × 10 ⁻²
SVT17 ¹⁶	ENSG000000103528	16	19 229 472	19 273 554	rs727747	4	0.177	0.05	1.54 × 10 ⁻⁴	1.58 × 10 ⁻²
SETD1A ¹⁹	ENSG000000099381	16	30 982 526	30 950 352	rs7206511	34	-0.710	0.09	2.75 × 10 ⁻¹³	5.77 × 10 ⁻¹¹
CHRN81 ¹⁶	ENSG000000170175	17	7 354 703	7 373 595	rs60488855	18	0.115	0.03	1.67 × 10 ⁻⁵	1.72 × 10 ⁻³
UBTF ¹⁹	ENSG000000108312	17	42 290 697	42 297 631	rs113844752	34	-0.466	0.09	5.68 × 10 ⁻⁶	1.19 × 10 ⁻³
MAPT ¹⁶	ENSG000000186868	17	44 038 724	44 862 347	rs199502	6	0.265	0.03	7.13 × 10 ⁻¹⁴	7.35 × 10 ⁻¹²
WNT3 (GTEXv7)	ENSG000000108379.5	17	44 875 148	44 908 263	rs9904865	2	-0.082	0.02	4.01 × 10 ⁻⁶	4.81 × 10 ⁻⁵
DNAH17 ¹⁶	cg09006072	17	76 425 972	76 427 732	rs589582	3	0.100	0.02	2.44 × 10 ⁻⁵	4.88 × 10 ⁻³
MEX3C ¹⁹	ENSG000000176624	18	48 722 797	48 731 131	rs12458916	40	-0.291	0.05	5.28 × 10 ⁻⁵	1.11 × 10 ⁻²

Multi-SNP eQTL Mendelian randomisation results focusing only on the most significant association per nearest gene to the Parkinson's disease risk variant of interest after Bonferroni correction. If a locus was significantly associated with both brain and blood QTLs after multiple test correction, we opted to show the most significant brain tissue derived association here after filtering for possible polygenicity (HEIDI p > 0.01). Effects with significant HEIDI p values might indicate a possible effect complicated by linkage disequilibrium and are less likely to be a true causal association. All tested QTL summary statistics can be found in the appendix pp 8-9. Effect estimates represent the change in Parkinson's disease odds ratio per one SD increase in gene expression or methylation. SNP= single nucleotide polymorphism. QTL=quantitative trait locus.

Table 3: Summary of significant functional inferences from QTL associations via Mendelian randomisation for nominated genes of interest

putamen volume ($p > 0.05$ in both the forward and reverse directions).

Discussion

This meta-analysis of GWAS marks a crucial step forward in our understanding of the genetic architecture of Parkinson's disease and provides a genetic reference set for the broader research community. We identified 90 independent common genetic risk factors for Parkinson's disease, nearly doubling the number of known Parkinson's disease risk variants. We re-evaluated the cumulative contribution of genetic risk variants, both of genome-wide significance and not yet discovered, to refine our estimates of heritable Parkinson's disease risk. We also nominated probable genes at each locus for further follow-up using QTL analyses and rare variant burden analyses. Our work has highlighted the pathways, tissues, and cell types involved in Parkinson's disease causes. Finally, we identified intracranial and putamenal volume as potential future Parkinson's disease biomarkers, and cognitive performance as a Parkinson's disease risk factor. Altogether, the data presented here has substantially expanded the resources available for future investigations into potential Parkinson's disease interventions.

We were able to explain 16–36% of Parkinson's disease heritability, the range being directly related to varying prevalence estimates (0.5–2.0%). Power estimates suggest that expansions of case numbers to 99 000 cases will continue to reveal additional insights into Parkinson's disease genetics. Although these risk variants will have small effects or be quite rare, they will help to further expand our knowledge of the genes and pathways that drive Parkinson's disease risk.

Population-wide screening for individuals who are likely to develop Parkinson's disease is currently not feasible using our 1805 variant PRS alone. There would be roughly 14 false positives per true positive assuming a prevalence of 0.5%. Although large-scale genome sequencing and non-linear machine learning methods will probably improve these predictive models, we have previously shown that we will need to incorporate other data sources (eg, smell tests, family history, age, sex) to generate algorithms that have more value in population-wide screening.⁴³

Evaluating these results in the larger context of pathway, tissue, and cellular functionality revealed that genes near Parkinson's disease risk variants showed enrichment for expression in the brain, contrasting with previous findings in Alzheimer's disease. Strikingly, we showed that the expression enrichment of genes at Parkinson's disease loci occurred exclusively in neuronal cell types. We also found that Parkinson's disease genes were enriched in chemical signalling pathways and pathways involving the response to a stressor. We believe that this contrast, in which the pathway enrichment analyses suggest at least some immune component to Parkinson's disease and the expression enrichment analysis does not suggest any significant immune related tissue component should be

viewed with caution. In particular, the marginal p values of most immune-related pathways in our analyses after multiple test correction reinforce this caution. These observations could be informative for disease modelling efforts, highlighting the importance of disease modelling in neurons and possibly incorporating a cellular stress component. This information can help inform and focus stem-cell derived therapeutic development efforts that are underway.⁴⁴

We found four phenotypes that were genetically correlated with Parkinson's disease. Putamen and intracranial volumes might prove to be valuable in future Parkinson's disease biomarker studies. Our bidirectional generalised summary-data-based Mendelian randomisation results suggest a complex causative connection between smoking initiation and Parkinson's disease that will require further follow-up. One of the implications of this work is that Parkinson's disease trials of nicotine or other smoking-related compounds might be less likely to succeed due to the marginal strength of associations shown in this report affecting study power. The strong causal effect of cognitive performance on Parkinson's disease is supported by observational studies.⁴⁵

Although this study marks major progress in assessing genetic risk factors for Parkinson's disease, much remains to be established. No defined external validation dataset was used, which could be seen as a limitation. However, due to the size of the dataset, it is considered infeasible to build a sufficiently large replication series. Also, external replication of the novel associations we present will be difficult simply because of the sample sizes needed. Simulations have suggested that, without replication, variants with p values between 5×10^{-8} and 5×10^{-9} should be interpreted with greater caution.^{46,47} We found 16 risk variants in this range, including two known variants near *WNT3* (proximal to the *MAPT* locus) and *BIN3*. To a degree, the fact that we filtered our variants with a secondary random-effects meta-analysis could make our 90 Parkinson's disease GWAS hits somewhat more robust because of the conservative nature of random-effects.

This study focused on Parkinson's disease risk in individuals of European ancestry. Adding datasets from non-European populations would be helpful to further improve our granularity in association testing and ability to fine-map loci through integration of more variable LD signatures while also evaluating population-specific associations. Also, risk predictions might not generalise across populations in some cases and ancestry specific PRS should be investigated. Additionally, large ancestry-specific Parkinson's disease LD reference panels, such as those for patients who are Ashkenazi Jews, will help us further unravel the genetic risk of loci such as *GBA* and *LRRK2*. This clarification might be particularly crucial at these loci where LD patterns could be variable within European populations, accentuating the possible influence of LD reference series on conditional analyses in some cases.⁴⁸ Finally, our work used state-of-the-art QTL

datasets to nominate candidate genes, but many QTL associations are hampered by both small sample size and low *cis*-SNP density. Larger QTL studies and Parkinson's disease-specific network data from large scale cellular screens would allow us to build a more robust functional inference framework.

As the field moves forward there are some crucial next steps that should be prioritised. First, allowing researchers to share participant-level data in a secure environment would facilitate inclusiveness and uniformity in analyses while maintaining the confidentiality of study participants. Our work suggests that GWAS of increasing size will continue to provide useful biological insights into Parkinson's disease. In addition to studies of the genetics of Parkinson's disease risk, studies of disease onset, progression, and subtype will be important and will require large series of well characterised patients.⁴⁹ We also believe that work across diverse populations is important, not only to be able to best serve these populations but also to aid in fine mapping of loci. Notably, the use of genome sequencing technologies could further improve discovery by capturing rare variants and structural variants, but with the caveat that very large sample sizes will be required. Although much is left to do, we believe that our study represents a substantial step forward and that the results and data will serve as a foundational resource for the community to pursue this next phase of Parkinson's disease research.

Contributors

MAN, CB, CLV, KH, SB-C, DC, MT, DK, LR, JS-S, LK, LP, and ABS were responsible for study-level analysis. MAN, CB, SB-C, AJN, AX, JY, JG, PMV, and ABS were responsible for additional analysis and data management. MAN, CB, CLV, KH, SB-C, LP, MS, KM, MT, AB, JY, ZG-O, TG, PH, JMS, NW, DAH, JH, HRM, JG, PMV, RRG, and ABS were responsible for design and funding. MAN, CB, CLV, KH, SB-C, DC, MT, DAK, AJN, AX, JB, EY, RvC, JS-S, CS, MS, LK, LP, AS, HI, HL, FF, JRG, DGH, SWS, JAB, MM, OAA, J-CC, SL, JJ, LMS, MS, PT, KM, MT, AB, JY, ZG-O, TG, PH, JMS, NW, DAH, JH, HRM, JG, PMV, RRG, and ABS were responsible for critical review and writing of the manuscript.

Declaration of interests

MAN reports that this work was done under a consulting contract with National Institutes of Health, he also consults for Lysosomal Therapeutics Inc, Neuron23, and Illumina. KH reports other support from 23andMe, during the conduct of the study outside the submitted work. DC reports other support from Genentech outside the submitted work. MT reports grants from Parkinson's UK during the conduct of the study. AJN reports grants from Parkinson's UK and Virginia Kieley benefaction; grants and non-financial support from GE Healthcare; and personal fees from Profile, Bial, and Britannia, outside the submitted work. RvC reports grants from American Brain Foundation and Michael and Eugenia Brin Foundation, during the conduct of the study. LP reports grants from Southeastern Regional Health Authority, Norway, during the conduct of the study. AS reports grants from Sigrid Juselius Foundation, during the conduct of the study. SWS is a scientific advisory council member for the Lewy Body Dementia Association. J-CC reports grants from French Ministry of Health, during the conduct of the study; grants from Sanofi, personal fees from Ever Pharma, Denali, Biogen, Air Liquide, BrainEver, and TheraNexus, outside the submitted work. LMS reports grants from National Institutes of Health, outside the submitted work. PT has a patent c9orf72 in the diagnosis and treatment of neurodegenerative disease pending. MT reports grants from Research Council of Norway, Regional Health Authority South-Eastern Norway, and Michael J Fox Foundation, and personal fees from Roche, outside the submitted work. OAA reports grants from Research Council of

Norway, and KG Jebsen Stiftelsen during the conduct of the study; and personal fees from Lundbeck, outside the submitted work; OAA also has a patent (PCT/US2014/011014) pending. TB reports other from Genentech, outside the submitted work. AB reports grants from France Parkinson and Fondation pour la Recherche sur le Cerveau, grants from Agence nationale de recherche (ANR) EPIG, grants from ANR joint program in neurodegenerative disease, grants from Roger de Spoelberch Foundation, France Alzheimer, Ecole des neurosciences Paris, Institut de France, CHU de Nîmes, European Rare Disease Network, ANR EPIG, and APHP, outside the submitted work. ZG-O reports personal fees from Lysosomal Therapeutics, Idorsia, Inception Sciences, Denali, and Prevail Therapeutics, outside the submitted work. TG reports grants from The Michael J Fox Foundation for Parkinson's Research, personal fees from UCB Pharma, other from Joint Programming for Neurodegenerative Diseases program, funded by the European Commission, personal fees from Novartis, Teva, and MedUpdate, outside the submitted work; and has a patent issued (EP1802749 [A2] KASPP [LRRK2] gene), for the detection and treatment of neurodegenerative disorders issued PH is a consultant for NEURON23. JMS reports grants from Burroughs Wellcome Fund during the conduct of the study, grants from National Institutes of Health, and personal fees from Helis Medical Foundation, outside the submitted work. DAH reports other from 23andMe, outside the submitted work. HRM reports grants from NIHR Biomedical Research Centre, Parkinson's UK, Medical Research Council, and The Cure Parkinson's Trust, during the conduct of the study; grants from PSP Association and CBD Solutions, personal fees from Teva, Boehringer Ingelheim, GlaxoSmithKline, grants from Drake Foundation, personal fees from Biogen, UCB, and Biohaven, outside the submitted work. RRG reports other from Genentech, during the conduct of the study and other from Genentech, outside the submitted work. All other authors declare no competing interests.

Data sharing

GWAS summary statistics for the post-Chang 23andMe dataset and 23andMe summary statistics included in the studies of Chang and colleagues⁴ and Nalls and colleagues⁶ will be made available through 23andMe to qualified researchers under an agreement with 23andMe that protects the privacy of the 23andMe participants. Interested investigators should visit <http://research.23andme.com/dataset-access> to submit a request. An immediately accessible version of the summary statistics is available online, excluding Nalls and colleagues⁶, 23andMe post-Chang and colleagues⁴ and Web-Based Study of Parkinson's Disease but including all analysed SNPs. After approval from 23andMe, the full summary statistics including all analysed SNPs and samples in this GWAS meta-analysis will be accessible to approved researchers. Underlying participant level International Parkinson's Disease Genomics Consortium data are available to potential collaborators, please contact ipdgc.contact@gmail.com.

Acknowledgments

For details of individual funding see appendix p 16.

References

- 1 Dorsey ER, Constantinescu R, Thompson JP, et al. Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. *Neurology* 2007; **68**: 384–86.
- 2 Polymeropoulos MH, Higgins JJ, Golbe LI, et al. Mapping of a gene for Parkinson's disease to chromosome 4q21-q23. *Science* 1996; **274**: 1197–99.
- 3 Singleton AB, Farrer M, Johnson J, et al. α -Synuclein locus triplication causes Parkinson's disease. *Science* 2003; **302**: 841.
- 4 Chang D, Nalls MA, Hallgr msd ttir IB, et al. A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat Genet* 2017; **49**: 1511–16.
- 5 Fung H-C, Scholz S, Matarin M, et al. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol* 2006; **5**: 911–16.
- 6 Nalls MA, Pankratz N, Lill CM, et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat Genet* 2014; **46**: 989–93.
- 7 Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; **26**: 2190–91.

For 23andMe publication dataset access and for more information see <https://research.23andme.com/collaborate/#publication>

For summary statistics see <https://bit.ly/2ofzGrk>

- 8 Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* 2012; **44**: 369–75, S1–3.
- 9 Euesden J, Lewis CM, O'Reilly PF. PRSice: polygenic risk score software. *Bioinformatics* 2014; **31**: 1466–68.
- 10 International Parkinson Disease Genomics Consortium, et al. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet* 2011; **377**: 641–49.
- 11 International Parkinson's Disease Genomics Consortium (IPDGC), Wellcome Trust Case Control Consortium 2 (WTCCC2). A two-stage meta-analysis identifies several new loci for Parkinson's disease. *PLoS Genet* 2011; **7**: e1002142.
- 12 Nalls MA, Escott-Price V, Williams NM, et al. Genetic risk and age in Parkinson's disease: continuum not stratum. *Mov Disord* 2015; **30**: 850–54.
- 13 Nalls MA, Bras J, Hernandez DG, et al. NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases. *Neurobiol Aging* 2015; **36**: e7–12.
- 14 Bulik-Sullivan BK, Loh P-R, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015; **47**: 291–95.
- 15 GTEx Consortium. The genotype-tissue expression (GTEx) project. *Nat Genet* 2013; **45**: 580–85.
- 16 Qi T, Wu Y, Zeng J, et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat Commun* 2018; **9**: 2282.
- 17 Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* 2016; **48**: 481–87.
- 18 Wu Y, Zeng J, Zhang F, et al. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat Commun* 2018; **9**: 918.
- 19 Vösa U, Claringbould A, Westra H-J, et al. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv* 2018; published online Oct 19. DOI:10.1101/447367 (preprint).
- 20 Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* 2017; **8**: 1826.
- 21 Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat Protoc* 2015; **10**: 1556–66.
- 22 Lee S, Emond MJ, Bamshad MJ, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 2012; **91**: 224–37.
- 23 LD Hub. <http://ldsc.broadinstitute.org/ldhub/> (accessed Jun 20, 2018).
- 24 Prins BP, Abbasi A, Wong A, et al. Investigating the causal relationship of c-reactive protein with 32 complex somatic and psychiatric outcomes: a large-scale cross-consortium Mendelian randomization study. *PLoS Med* 2016; **13**: e1001976.
- 25 Ahola-Olli AV, Würtz P, Havulinna AS, et al. Genome-wide association study identifies 27 loci influencing concentrations of circulating cytokines and growth factors. *Am J Hum Genet* 2017; **100**: 40–50.
- 26 Lee JJ, Wedow R, Okbay A, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet* 2018; published online Jul 23. DOI:10.1038/s41588-018-0147-3.
- 27 Elliott LT, Sharp K, Alfaro-Almagro F, et al. Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* 2018; **562**: 210–16.
- 28 Keller MF, Saad M, Bras J, et al. Using genome-wide complex trait analysis to quantify “missing heritability” in Parkinson's disease. *Hum Mol Genet* 2012; **21**: 4996–5009.
- 29 Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011; **88**: 76–82.
- 30 Wray NR, Yang J, Goddard ME, Visscher PM. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet* 2010; **6**: e1000864.
- 31 Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006; **38**: 209–13.
- 32 Cruts M, Gijssels I, van der Zee J, et al. Null mutations in progranulin cause ubiquitin-positive frontotemporal dementia linked to chromosome 17q21. *Nature* 2006; **442**: 920–24.
- 33 Smith KR E. Strikingly different clinicopathological phenotypes determined by progranulin-mutation dosage. *Am J Hum Genet* 2012; **90**: 1102–07.
- 34 Wang J, Vasaiakar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res* 2017; **45**: W130–37.
- 35 Gene set enrichment analysis. 2019. <http://software.broadinstitute.org/gsea/index.jsp> (accessed June 20, 2018).
- 36 Robak LA, Jansen IE, van Rooij J, et al. Excessive burden of lysosomal storage disorder gene variants in Parkinson's disease. *Brain* 2017; **140**: 3191–203.
- 37 Bandres-Ciga S, Saez-Atienzar S, Bonet-Ponce L, et al. The endocytic membrane trafficking pathway plays a major role in the risk of Parkinson's disease. *Mov Disord* 2019; **34**: 460–68.
- 38 Jansen IE, Savage JE, Watanabe K, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet* 2019; **51**: 404–13.
- 39 Saunders A, Macosko EZ, Wysoker A, et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* 2018; **174**: 1015–30.
- 40 Hibar DP, Stein JL, Renteria ME, et al. Common genetic variants influence human subcortical brain structures. *Nature* 2015; **520**: 224–29.
- 41 Neale lab. Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK Biobank. 2017. <http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank> (accessed Jun 24, 2018).
- 42 Bulik-Sullivan B, ReproGen Consortium, Finucane HK, Anttila V, Gusev A, Day FR. An atlas of genetic correlations across human diseases and traits. *Nat Genet* 2015; **47**: 1236–41.
- 43 Nalls MA, McLean CY, Rick J, et al. Diagnosis of Parkinson's disease on the basis of clinical and genetic classification: a population-based modelling study. *Lancet Neurol* 2015; **14**: 1002–09.
- 44 Parkinson's progression Markers Initiative. 2019. <http://www.ppmi-info.org/access-data-specimens/> (accessed June 20, 2018).
- 45 Valdés EG, Andel R, Sieurin J, et al. Occupational complexity and risk of Parkinson's disease. *PLoS One* 2014; **9**: e106676.
- 46 Wu Y, Zheng Z, Visscher PM, Yang J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. *Genome Biol* 2017; **18**: 86.
- 47 Pulit SL, de With SAJ, de Bakker PIW. Resetting the bar: statistical significance in whole-genome sequencing-based association studies of global populations. *Genet Epidemiol* 2017; **41**: 145–51.
- 48 Rivas MA, Avila BE, Koskela J, et al. Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population. *PLoS Genet* 2018; **14**: e1007329.
- 49 Iwaki H, Blauwendraat C, Leonard HL, et al. Genome-wide association study of Parkinson's disease progression biomarkers in 12 longitudinal patients' cohorts. *bioRxiv* 2019; published online March 27. DOI:10.1101/585836 (preprint).

Analysis of PTRHD1 common and rare variants in European patients with Parkinson's disease

Yuri L. Sosero* MD^{1,2}, Sara Bandres-Ciga PhD³, Ziv Gan-Or MD, PhD^{1,2,4}, Lynne Krohn MSc^{1,2},
on behalf of the IPDGC

Affiliations:

1. Department of Human Genetics, McGill University, Montréal, QC, H3A 1A1, Canada
2. Montreal Neurological Institute, McGill University, Montréal, QC, H3A 1A1, Canada
3. Molecular Genetics Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD 20892, USA
4. Department of Neurology and Neurosurgery, McGill University, Montréal, QC, H3A 1A1, Canada

***Corresponding author:**

Yuri Ludwig Sosero
Montreal Neurological Institute,
McGill University
1033 Pine Avenue, West,
Ludmer Pavilion, room 308-309
Montreal, QC, H3A 1A1,
Phone: +1-438-979-5052
e-mail: yuri.sosero@mail.mcgill.ca

Keywords: PTRHD1, Parkinson's disease, familial parkinsonism, burden test, SKATO, GWAS

Abstract

Three family studies identified three different variants in the peptidyl-tRNA hydrolase domain containing 1 gene (*PTRHD1*) in patients affected by syndromic parkinsonism. In the current study, our objective was to investigate whether *PTRHD1* variants are associated with Parkinson's disease (PD) risk and age at onset (AAO). To evaluate the association between *PTRHD1* and PD risk, we analyzed whole genome sequencing (WGS) data of 1,647 PD cases and 1,050 healthy controls, as well as genome-wide imputed genotyping data on 14,671 PD cases and 17,667 controls, all of European ancestry. Furthermore, we examined the association of *PTRHD1* with PD risk and AAO using summary statistics data from the most recent PD genome-wide association study (GWAS) meta-analyses. Our results show no association between *PTRHD1* and PD risk or AAO. We conclude that *PTRHD1* does not play a major role in PD in the European population. Further large-scale studies including subjects with different ancestry and family trios might further clarify the relationship of this gene with PD and atypical parkinsonism.

1. Introduction

Parkinson's disease is a complex neurodegenerative disorder resulting from a variety of genetic and non-genetic risk factors. Three family studies nominated peptidyl-tRNA hydrolase domain containing 1 gene (*PTRHD1*) as a possible disease-causing gene in atypical parkinsonism (Jaberi et al., 2016; Khodadadi et al., 2017; Kuipers et al., 2018). In these studies, two conducted in Iranian families (Jaberi et al., 2016; Khodadadi et al., 2017) and one in a specific sub-Saharan African community (Kuipers et al., 2018), the carriers showed some level of early onset cognitive impairment, and later developed symptoms of parkinsonism. In particular, homozygous *PTRHD1* c.157C>T (p.His53Tyr) variants were detected in two Iranian siblings with intellectual disability and motor abnormalities. At their second/third decade of age, the siblings presented with muscle stiffness, anxiety, resting and postural tremor, along with hypersexuality and hypersomnia (Khodadadi et al., 2017). Similar symptoms, including intellectual disability/attention deficits and, successively, gait disturbance, bradykinesia, tremor and falls, have also been reported in two male siblings from another Iranian family, homozygous carriers of a *PTRHD1* c.155G>A (p.Cys52Tyr) variant (Jaberi et al., 2016). Likewise, the study on the Sub-Saharan Xhosa-speaking community showed that, within the same family, three homozygous carriers of a *PTRHD1* deletion (c.169_196del, p.Ala57Argfs*26) had mental disability and symptoms of parkinsonism, with a wider clinical variability compared to the two previous studies (one member showed just tremor, one just bradykinesia and hypokinesia, one both) (Kuipers et al., 2018).

Despite these findings, no large-scale study has been performed before in the European population to investigate the relationship between *PTRHD1* and PD. To test this potential association, we examined the *PTRHD1* region in Europeans using whole genome sequencing (WGS) data from the Accelerating Medicines Partnership - Parkinson's disease initiative (AMP-PD; www.amp-pd.org) and large-scale genome-wide imputed genotyping data from the International Parkinson's Disease Genomics Consortium (IPDGC).

2. Methods

2.1 Population

To investigate the association of *PTRHD1* common and rare variants with PD, we analyzed AMP-PD WGS data including 1,647 PD cases and 1,050 healthy controls of European descent from three different cohorts: Parkinson's Disease Biomarker Program (PDBP; <https://pdbp.ninds.nih.gov/>), Biofind (<https://biofind.loni.usc.edu/>), and Parkinson's Progression Markers Initiative (PPMI; <https://www.ppmi-info.org/>). In addition, we examined the association between *PTRHD1* and PD on a large European cohort of individual-level GWAS data including 14,671 PD cases and 17,667 controls from the International Parkinson's Disease Genomics Consortium (IPDGC). Finally, we assessed summary statistics from the most recent GWAS meta-analyses on PD risk (Nalls et al., 2019) and age at onset (AAO) (Blauwendraat et al., 2020) excluding 23anMe data. In particular, the meta-analysis on PD risk consists of 15,056 PD cases, 18,618 UK Biobank proxy-cases (i.e., subjects with a first degree relative with PD) and 449,056 controls, whereas the meta-analysis on PD AAO includes 17,996 PD patients.

2.2 Genetic data

WGS data, processing and quality control pipelines can be found at www.amp-pd.org. Quality control for GWAS data was performed as previously described (Nalls et al., 2019). For both genotyping and sequencing datasets, *PTRHD1* region was annotated using ANNOVAR. To specifically analyze rare variants from WGS data, we selected variants with a minor allele frequency (MAF) <0.01. Allele frequency was calculated using PLINK v1.9 (Chang et al., 2015).

2.3 Statistical analysis

To test the association between variants in the *PTRHD1* region and PD, we performed Fisher exact test and logistic regression in WGS and imputed individual-level genotyping data adjusted for age,

sex and principal components to account for population substructure, including variants within *PTRHD1* region +/- 100 Kb. Furthermore, summary statistics from the meta-analysis on PD AAO (Blauwendraat et al., 2020) and PD risk (Nalls et al., 2019) was investigated. Finally, as the *PTRHD1* variants analyzed in the reported family studies were all rare, we evaluated the enrichment of *PTRHD1* rare variants in WGS data using burden and kernel tests. The analyses on rare variants were first performed on all *PTRHD1* rare variants that passed quality control, and then repeated only on the nonsynonymous rare variants. Bonferroni correction for multiple comparisons was applied as needed. Association analyses were performed using PLINK v1.9 and burden analyses using software package RVTESTS v.2.1.0 (Zhan et al., 2016).

3. Results

In the WGS data we identified 1829 variants in the *PTRHD1* region +/-100 Kb, of which 29 were found within the *PTRHD1* gene. Among these, two were coding variants and 13 were identified as rare variants (MAF < 0.01). After Bonferroni correction, both logistic regression and Fisher test showed no association between *PTRHD1* variants and PD (Supplementary Tables S1-S2). We repeated the analyses using GWAS individual-level data and detected 511 variants in the *PTRHD1* region +/-100 Kb, with 14 variants in *PTRHD1* gene, none of which were coding. Similarly, the results showed no significant association between *PTRHD1* and PD (Supplementary Tables S3-S4). In addition, the analyses using the PD meta-analysis summary statistics confirmed such lack of association of *PTRHD1* with PD risk as well as with PD AAO (Figure 1). Finally, burden and kernel tests in WGS data did not show an enrichment of *PTRHD1* rare variants in PD, compared to controls (Supplementary Table S5). As the previously identified *PTRHD1* variants were all homozygous, we also assessed the number of homozygous/compound heterozygous carriers of at least one *PTRHD1* coding variant, showing similar proportions of such carriers between PD patients and controls (Supplementary Table S6). Of note, the three *PTRHD1* variants reported in the

previous studies on Iranian and African families (Jaberi et al., 2016; Khodadadi et al., 2017; Kuipers et al., 2018) were not found in our dataset.

4. Discussion

Herein, we did not find evidence for an association between *PTRHD1* and PD in the European population. There are several reasons that might explain the discrepancy between our results and the previous findings on *PTRHD1* in parkinsonian families (Jaberi et al., 2016; Khodadadi et al., 2017; Kuipers et al., 2018): 1) The findings from the family studies might not replicate on a large scale, suggesting that *PTRHD1* variants might be either extremely rare in PD or not associated at all; 2) *PTRHD1* carriers in the reported family studies manifested atypical parkinsonism with early-onset intellectual impairment, whereas our study was performed on patients with typical PD. Therefore, *PTRHD1* variants might be associated with rare forms of atypical parkinsonism, rather than typical PD; 3) The family studies were performed in Iranian and African subjects, whereas our study was performed in European patients. These ethnic differences might as well explain the discrepancy between our and the three previous family studies on *PTRHD1*. Our findings are further supported by a recent study conducted on PD families and early onset PD patients in a Taiwanese population (Chen et al., 2020). The study demonstrated an absence of *PTRHD1* pathogenic coding variants in PD patients.

In conclusion, our results do not provide evidence for an association between *PTRHD1* and PD. Larger studies on non-European ancestry populations and family trios are required to clarify the role of *PTRHD1* in PD and atypical parkinsonism.

5. Acknowledgements

We would like to thank all of the subjects who donated their time and biological samples to be a part of this study. We also would like to thank all members of the International Parkinson's Disease Genomics Consortium (IPDGC). For a complete overview of members, acknowledgements and

funding, please see <http://pdgenetics.org/partners>. This work was supported in part by the Intramural Research Programs of the National Institute of Neurological Disorders and Stroke (NINDS), the National Institute on Aging, and the National Institute of Environmental Health Sciences both part of the National Institutes of Health, Department of Health and Human Services; project numbers 1ZIA-NS003154, Z01-AG000949-02 and Z01-ES101986. In addition, this work was supported by the Department of Defense (award W81XWH-09-2-0128), and The Michael J Fox Foundation for Parkinson's Research. 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). This work utilized the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>). ZGO is supported by the Fonds de recherche du Québec - Santé (FRQS) Chercheurs-boursiers award, in collaboration with Parkinson Quebec, and by the Young Investigator Award by Parkinson Canada.

6. Conflicts of Interest

ZGO has received consulting fees from Lysosomal Therapeutics Inc., Idorsia, Prevail Therapeutics, Denali, Ono Therapeutics, Neuron23, Handl Therapeutics, Deerfield and Inception Sciences (now Ventus). None of these companies were involved in any parts of preparing, drafting and publishing this study. Other authors have no additional disclosures to report.

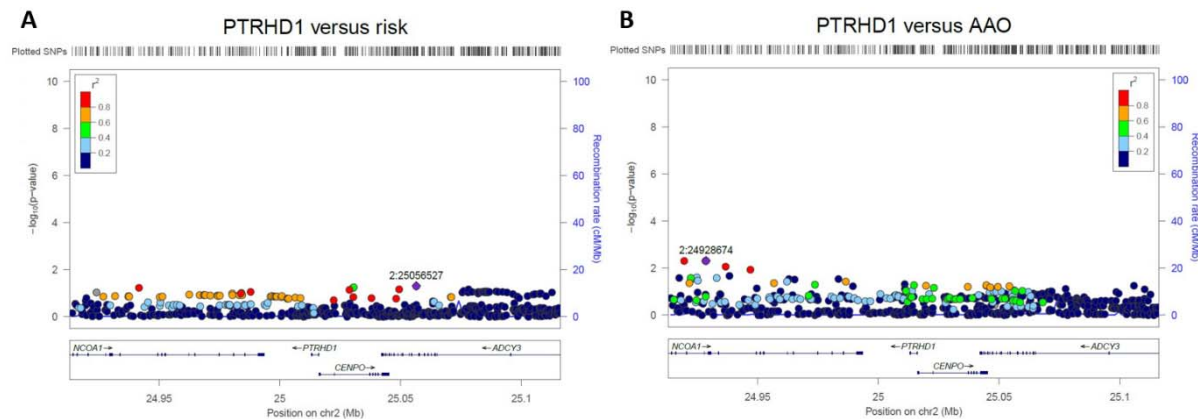


Fig. 1 (A) Locus zoom plot of *PTRHD1* variants versus PD risk. (B) Locus zoom plot of *PTRHD1* variants versus PD AAO. The position of the variants on the chromosome 2 (x axis) is plotted against the log10-scaled p-values (left y axis). The most strongly associated SNP is indicated by a purple diamond. Pairwise linkage disequilibrium scores are defined by different colors, explained by the legend on the upper right corner. The right vertical axis indicates the regional recombination rate (cM/Mb).

Abbreviations: chr, chromosome; cM, centimorgan; Mb, Megabase; PD, Parkinson's disease; SNP, single nucleotide polymorphism.

7. References

- Blauwendraat, C., Nalls, M.A., Singleton, A.B., 2020. The genetic architecture of Parkinson's disease. *Lancet Neurol* 19(2), 170-178.
- Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7.
- Chen, S.J., Ho, C.H., Lin, H.Y., Lin, C.H., Wu, R.M., 2020. Lack of PTRHD1 mutation in patients with young-onset and familial Parkinson's disease in a Taiwanese population. *Neurobiol Aging*.
- Jaberi, E., Rohani, M., Shahidi, G.A., Nafissi, S., Arefian, E., Soleimani, M., Moghadam, A., Arzenani, M.K., Keramatian, F., Klotzle, B., Fan, J.B., Turk, C., Steemers, F., Elahi, E., 2016. Mutation in ADORA1 identified as likely cause of early-onset parkinsonism and cognitive dysfunction. *Mov Disord* 31(7), 1004-1011.
- Khodadadi, H., Azcona, L.J., Aghamollai, V., Omrani, M.D., Garshasbi, M., Taghavi, S., Tafakhori, A., Shahidi, G.A., Jamshidi, J., Darvish, H., Paisan-Ruiz, C., 2017. PTRHD1 (C2orf79) mutations lead to autosomal-recessive intellectual disability and parkinsonism. *Mov Disord* 32(2), 287-291.
- Kuipers, D.J.S., Carr, J., Bardien, S., Thomas, P., Sebaste, B., Breedveld, G.J., van Minkelen, R., Brouwer, R.W.W., van Ijcken, W.F.J., van Slegtenhorst, M.A., Bonifati, V., Quadri, M., 2018. PTRHD1 Loss-of-function mutation in an african family with juvenile-onset Parkinsonism and intellectual disability. *Mov Disord* 33(11), 1814-1819.
- Nalls, M.A., Blauwendraat, C., Vallerga, C.L., Heilbron, K., Bandres-Ciga, S., Chang, D., Tan, M., Kia, D.A., Noyce, A.J., Xue, A., Bras, J., Young, E., von Coelln, R., Simon-Sanchez, J., Schulte, C., Sharma, M., Krohn, L., Pihlstrom, L., Siitonen, A., Iwaki, H., Leonard, H., Faghri, F., Gibbs, J.R., Hernandez, D.G., Scholz, S.W., Botia, J.A., Martinez, M., Corvol, J.C., Lesage, S., Jankovic, J., Shulman, L.M., Sutherland, M., Tienari, P., Majamaa, K., Toft, M., Andreassen, O.A., Bangale, T., Brice, A., Yang, J., Gan-Or, Z., Gasser, T., Heutink, P., Shulman, J.M., Wood, N.W., Hinds, D.A., Hardy, J.A., Morris, H.R., Gratten, J., Visscher, P.M., Graham, R.R., Singleton, A.B., and Me Research, T., System Genomics of Parkinson's Disease, C., International Parkinson's Disease

Genomics, C., 2019. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol* 18(12), 1091-1102.

Zhan, X., Hu, Y., Li, B., Abecasis, G.R., Liu, D.J., 2016. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* 32(9), 1423-1426.

Common and rare variants in HFE are not associated with Parkinson's disease in Europeans

Prabhjyot Saini,^{1,2} Sara Bandres-Ciga,³ Jose Luis Alcantud,⁴ Clara Ruz,⁴ Ziv Gan-Or,^{1,2,5} for the International Parkinson's disease Genomics Consortium

Affiliations:

¹Montreal Neurological Institute, McGill University, Montréal, QC, Canada

²Department of Human Genetics, McGill University, Montréal, QC, Canada

³Molecular Genetics Section, Laboratory of Neurogenetics, NIA, NIH, Bethesda, MD, USA

⁴Instituto de Neurociencias Federico Olóriz, Centro de Investigación Biomédica, 18016 Granada, Spain.

⁵Department of Neurology and neurosurgery, McGill University, Montréal, QC, Canada

Corresponding author:

Ziv Gan-Or

Department of Neurology and Neurosurgery

McGill University

1033 Pine Avenue, West, Ludmer Pavilion, room 312

Montreal, QC, H3A 1A1, Canada

Email: ziv.gan-or@mcgill.ca

Telephone: +1-514-398-5845

Abstract

A recent study suggested that the p.H63D variant in *HFE*, a gene involved in iron homeostasis, may modify α -synuclein pathology, the pathological hallmark of Parkinson's disease (PD). If indeed this gene and specific variant are involved in PD, we expect to find differential distribution of *HFE* variants when comparing PD patients and controls. We analyzed genome-wide association study (GWAS) data from 14,671 PD patients and 17,667 controls and full sequencing data from additional 1,647 PD patients and 1,050 controls, using logistic regression models, and burden and Kernel tests. The *HFE* p.H63D variant was not associated with PD, nor did all the other common variants in the *HFE* locus. We did not find association of rare *HFE* variants with PD as well in all types of burden and Kernel tests. Our results do not support a role for *HFE* in PD.

Introduction

Parkinson's disease (PD) is a complex neurodegenerative disorder, caused in most cases by multiple genetic and environmental factors, and the normal process of aging. Previous studies have suggested that iron homeostasis may have a role in PD; pathological and imaging studies showed that iron accumulation in the substantia nigra may contribute to PD pathogenesis (Ward et al., 2014). Multiple genes are involved in iron homeostasis, one of the most important of them being *HFE*, encoding the homeostatic iron regulator protein (also called hereditary hemochromatosis protein). This protein is a membrane protein thought to regulate iron absorption, and bi-allelic *HFE* mutations may lead to hereditary haemochromatosis, an iron storage disorder.

Despite several negative genetic studies (Biasiotto et al., 2008; Duan et al., 2016; Guerreiro et al., 2006; Rhodes et al., 2014; Xia et al., 2015), a recent study suggested that the p.H63D variant in *HFE* may contribute to Parkinson's disease pathogenesis by modifying α -synuclein pathology in cell models (Kim et al., 2020). The same group has previously suggested, using a mouse model with the homolog of the same variant (H67D), that *HFE* is a disease modifier in PD (Nixon et al., 2018). In order to examine whether this and other *HFE* variants are associated with PD, we performed a comprehensive analysis of common and rare *HFE* variants in large case-control cohorts of PD patients and controls of European origin.

Methods

Two cohorts were included in the current study: 1) a total of 14,671 PD patients and 17,667 controls collected through collaborators from the International PD Genomics Consortium (IPDGC), and 2) 1,647 PD patients and 1,050 controls from AMP-PD (<https://amp-pd.org/>). Demographic details on these cohorts can be found in Supplementary Table S1A and S1B. We

performed standard genome-wide association study (GWAS) quality control (QC) on the individual and variant level data as previously described (Nalls et al., 2019). We performed a similar QC process for the AMP-PD whole genome sequencing data, as detailed in the AMP-PD portal (<https://amp-pd.org/whole-genome-data>). Genotype data was extracted from both datasets for all variants within 100kb upstream and downstream of *HFE*, and was annotated using ANNOVAR (Wang et al., 2010). To examine whether common variants (allele frequency >0.01) in *HFE* are associated with PD, we performed logistic regression adjusted for age, sex, 10 principal components and sites of cohorts. To determine whether rare variants (allele frequency <0.01) we performed a set of burden and Kernel association tests included in the RVtests R package (Zhan et al., 2016). Bonferroni correction for multiple comparisons was applied, and all code used in the current study is available at the IPDGC GitHub at <https://github.com/ipdgc/IPDGC-Trainees/blob/master/>. The institutional review board (McGill University Health Center Research Ethics Board - MUHC REB) approved the study protocols.

Results

None of the common variants in the *HFE* region was associated with risk of PD (Figure 1). The specific *HFE* variant recently reported to be associated with decreased aggregation of α -synuclein, p.H63D, had similar allele frequencies in patients and controls (0.15 and 0.14, respectively), and was not associated with risk of PD ($p=1$ after correction for multiple comparisons). Another *HFE* variant previously reported to be associated with PD, p.C282Y, had similar frequencies in patients and controls as well (0.049 in both) and was not associated with PD (corrected $p=1$). Similarly, none of the burden and Kernel association tests we used to examine the association of rare *HFE* variants with PD yielded statistically significant results

(Table 1). The list of all common and rare variants found in *HFE* is available in Supplementary Table S2 and S3, respectively.

Discussion

In the current study we performed a thorough genetic analysis of *HFE* in large case-control cohorts of PD patients and controls. Our results do not support a role for common or rare *HFE* variants in PD. Previous smaller genetic association studies focusing on *HFE* reported contradicting results, as some studies reported associations of different *HFE* variants with PD, while other studies reported lack of association (Biasiotto et al., 2008; Borie et al., 2002; Buchanan et al., 2002; Duan et al., 2016; Guerreiro et al., 2006; Halling et al., 2008; Rhodes et al., 2014; Xia et al., 2015). Specifically for p.H63D, all the previous studies, including two meta-analyses (Duan et al., 2016; Xia et al., 2015), reported lack of association. Together with our study, there is no support from human genetic studies in PD patients for the involvement of *HFE* p.H63D in PD. Furthermore, there is no convincing evidence to suggest an association between other *HFE* variants and risk of PD. We cannot rule out that a very small effect does exist, but since it was not detected by the recent GWAS with over 50,000 PD patients and proxy-cases (Nalls et al., 2019), it is likely to be too small to be clinically meaningful, if at all.

The main limitation of the current study is that it includes only participants of European descent, and since PD is a complex disorder, it is possible that in other populations *HFE* variants may still contribute to the disease when found on another genetic background. In a recent GWAS of Asian population, however, no GWAS-significant association has been reported for *HFE* (Foo et al., 2020).

Despite the lack of association in the current study, the role of iron and iron homeostasis in PD should be further studied, whether by genetic studies or by biochemical, functional,

clinical and pathological studies. We recommend that genetic data from the IPDGC and other available studies should be consulted before studying specific genetic variants in different models.

Acknowledgements

We thank the participants for contributing to the study. We would like to also thank all members of the International Parkinson Disease Genomics Consortium (IPDGC). For a complete overview of members, acknowledgements and funding, please see <http://pdgenetics.org/partners>. This research was supported in part by the Intramural Research Program of the NIH, National institute on Aging. This work was also supported in part by grants from the Michael J. Fox Foundation, the Canadian Consortium on Neurodegeneration in Aging (CCNA), and the Canada First Research Excellence Fund (CFREF), awarded to McGill University for the Healthy Brains for Healthy Lives initiative (HBHL). ZGO is supported by the Fonds de recherche du Québec - Santé (FRQS) Chercheurs-boursiers award, in collaboration with Parkinson Quebec, and by the Young Investigator Award by Parkinson Canada.

Conflict of interests

ZGO has received consulting fees from Lysosomal Therapeutics Inc., Idorsia, Prevail Therapeutics, Denali, Ono Therapeutics, Neuron23, Handl Therapeutics, Deerfield and Inception Sciences (now Ventus). None of these companies were involved in any parts of preparing, drafting and publishing this study. Other authors have no additional disclosures to report.

References

- Biasiotto, G., Goldwurm, S., Finazzi, D., Tunesi, S., Zecchinelli, A., Sironi, F., Pezzoli, G., Arosio, P., 2008. HFE gene mutations in a population of Italian Parkinson's disease patients. *Parkinsonism Relat Disord* 14(5), 426-430.
- Borie, C., Gasparini, F., Verpillat, P., Bonnet, A.M., Agid, Y., Hetet, G., Brice, A., Durr, A., Grandchamp, B., French Parkinson's disease genetic study, g., 2002. Association study between iron-related genes polymorphisms and Parkinson's disease. *J Neurol* 249(7), 801-804.
- Buchanan, D.D., Silburn, P.A., Chalk, J.B., Le Couteur, D.G., Mellick, G.D., 2002. The Cys282Tyr polymorphism in the HFE gene in Australian Parkinson's disease patients. *Neurosci Lett* 327(2), 91-94.
- Duan, C., Wang, M., Zhang, Y., Wei, X., Huang, Y., Zhang, H., Cheng, L., Gai, Z., 2016. C282Y and H63D Polymorphisms in Hemochromatosis Gene and Risk of Parkinson's Disease: A Meta-Analysis. *Am J Alzheimers Dis Other Dement* 31(3), 201-207.
- Foo, J.N., Chew, E.G.Y., Chung, S.J., Peng, R., Blauwendraat, C., Nalls, M.A., Mok, K.Y., Satake, W., Toda, T., Chao, Y., Tan, L.C.S., Tandiono, M., Lian, M.M., Ng, E.Y., Prakash, K.M., Au, W.L., Meah, W.Y., Mok, S.Q., Annuar, A.A., Chan, A.Y.Y., Chen, L., Chen, Y., Jeon, B.S., Jiang, L., Lim, J.L., Lin, J.J., Liu, C., Mao, C., Mok, V., Pei, Z., Shang, H.F., Shi, C.H., Song, K., Tan, A.H., Wu, Y.R., Xu, Y.M., Xu, R., Yan, Y., Yang, J., Zhang, B., Koh, W.P., Lim, S.Y., Khor, C.C., Liu, J., Tan, E.K., 2020. Identification of Risk Loci for Parkinson Disease in Asians and Comparison of Risk Between Asians and Europeans: A Genome-Wide Association Study. *JAMA Neurol* 77(6), 746-754.
- Guerreiro, R.J., Bras, J.M., Santana, I., Januario, C., Santiago, B., Morgadinho, A.S., Ribeiro, M.H., Hardy, J., Singleton, A., Oliveira, C., 2006. Association of HFE common mutations with Parkinson's disease, Alzheimer's disease and mild cognitive impairment in a Portuguese cohort. *BMC Neurol* 6, 24.
- Halling, J., Petersen, M.S., Grandjean, P., Weihe, P., Brosen, K., 2008. Genetic predisposition to Parkinson's disease: CYP2D6 and HFE in the Faroe Islands. *Pharmacogenet Genomics* 18(3), 209-212.
- Kim, Y., Stahl, M.C., Huang, X., Connor, J.R., 2020. H63D variant of the homeostatic iron regulator (HFE) gene alters alpha-synuclein expression, aggregation, and toxicity. *J Neurochem*.
- Nalls, M.A., Blauwendraat, C., Vallerga, C.L., Heilbron, K., Bandres-Ciga, S., Chang, D., Tan, M., Kia, D.A., Noyce, A.J., Xue, A., Bras, J., Young, E., von Coelln, R., Simon-Sanchez, J., Schulte, C., Sharma, M., Krohn, L., Pihlstrom, L., Siitonen, A., Iwaki, H., Leonard, H., Faghri, F., Gibbs, J.R., Hernandez, D.G., Scholz, S.W., Botia, J.A., Martinez, M., Corvol, J.C., Lesage, S., Jankovic, J., Shulman, L.M., Sutherland, M., Tienari, P., Majamaa, K., Toft, M., Andreassen, O.A., Bangale, T., Brice, A., Yang, J., Gan-Or, Z., Gasser, T., Heutink, P., Shulman, J.M., Wood, N.W., Hinds, D.A., Hardy, J.A., Morris, H.R., Gratten, J., Visscher, P.M., Graham, R.R., Singleton, A.B., and Me Research, T., System Genomics of Parkinson's Disease, C., International Parkinson's Disease Genomics, C., 2019. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol* 18(12), 1091-1102.
- Nixon, A.M., Meadowcroft, M.D., Neely, E.B., Snyder, A.M., Purnell, C.J., Wright, J., Lamendella, R., Nandar, W., Huang, X., Connor, J.R., 2018. HFE Genotype Restricts the Response to Paraquat in a Mouse Model of Neurotoxicity. *J Neurochem* 145(4), 299-311.

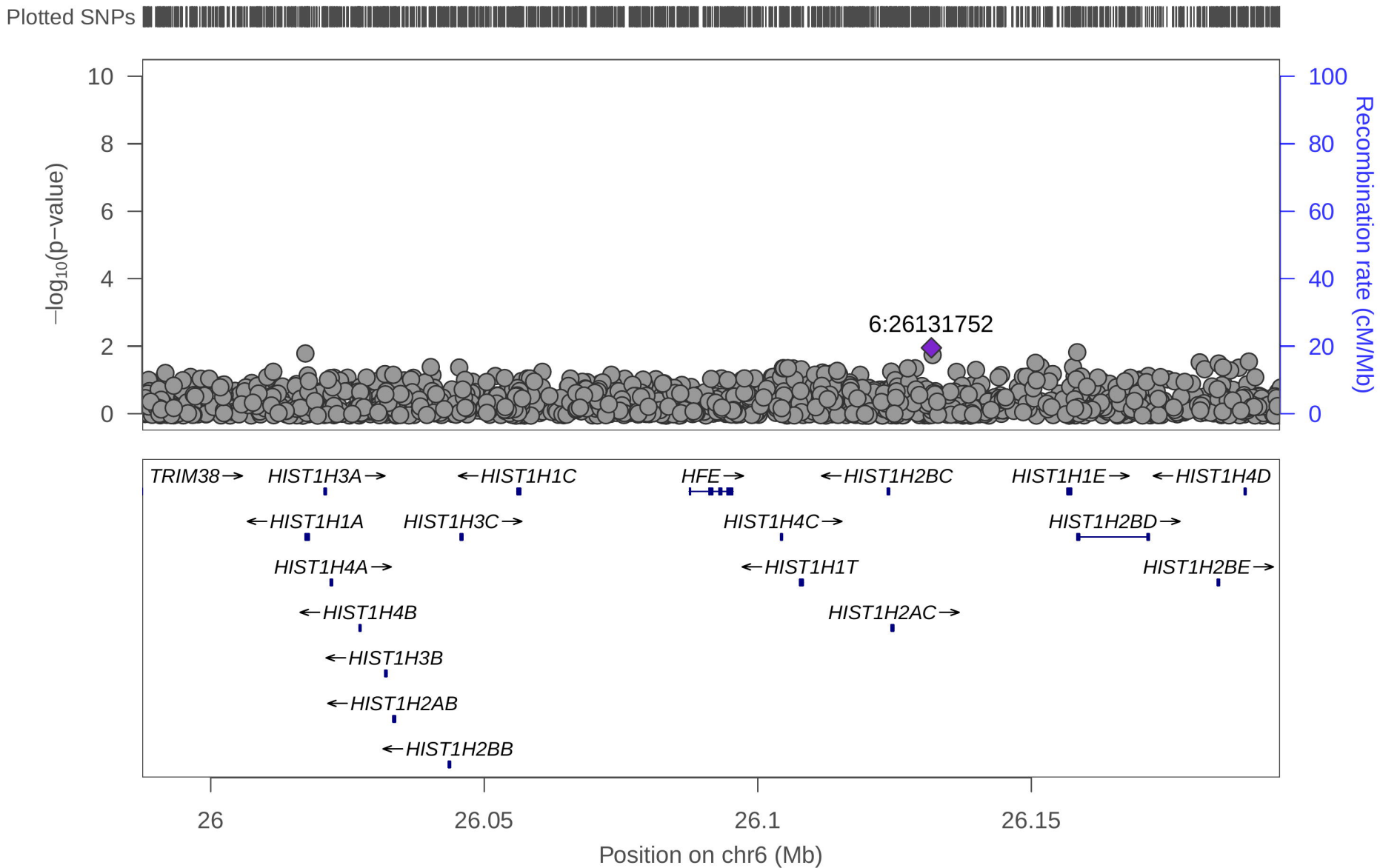
- Rhodes, S.L., Buchanan, D.D., Ahmed, I., Taylor, K.D., Lorient, M.A., Sinsheimer, J.S., Bronstein, J.M., Elbaz, A., Mellick, G.D., Rotter, J.I., Ritz, B., 2014. Pooled analysis of iron-related genes in Parkinson's disease: association with transferrin. *Neurobiol Dis* 62, 172-178.
- Wang, K., Li, M., Hakonarson, H., 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38(16), e164.
- Ward, R.J., Zucca, F.A., Duyn, J.H., Crichton, R.R., Zecca, L., 2014. The role of iron in brain ageing and neurodegenerative disorders. *Lancet Neurol* 13(10), 1045-1060.
- Xia, J., Xu, H., Jiang, H., Xie, J., 2015. The association between the C282Y and H63D polymorphisms of HFE gene and the risk of Parkinson's disease: A meta-analysis. *Neurosci Lett* 595, 99-103.
- Zhan, X., Hu, Y., Li, B., Abecasis, G.R., Liu, D.J., 2016. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* 32(9), 1423-1426.

Table 1. Burden tests for *HFE* in the AMP-PD cohorts

Cohort	Burden test	All variants			Coding variants		
		Number of Variants	Number of Polygenic Variants	P-value	Number of Variants	Number of Polygenic Variants	P-value
AMP-PD	Zeggini	54	47	0.5167	12	9	0.3644
	Fp	54	47	0.4510	12	9	0.1134
	CMC	54	47	0.2539	12	9	0.3644
	Madson-Browning	54	47	0.2940	12	9	0.2959
	SKAT	54	47	0.8299	12	9	0.8266
	SKAT-O	54	47	0.6310	12	9	0.3425

Abbreviations: SKAT: Sequence kernel association test, SKAT-O: Sequence kernel association test optimized, CMC: combined multivariate and collapsing

HFE versus PD risk



Multi-Modality Machine Learning Predicting Parkinson's Disease

Integrating clinico-demographic, genetic, and transcriptomic data within an automated machine learning open science framework to predict Parkinson's disease and identify potential novel therapeutic targets for drug development.

AUTHORS AND AFFILIATIONS

Mary B. Makarious^{1,2,3}, Hampton L. Leonard^{1,4,5,6}, Dan Vitale⁵, Hirotaka Iwaki^{1,4,5}, Lana Sargent^{1,4,7,8}, Anant Dadu⁹, Ivo Violich¹⁰, Elizabeth Hutchins¹¹, David Saffo¹², Sara Bandres-Ciga¹, Jonggeol Jeff Kim^{1,13}, Yeajin Song⁵, Matt Bookman¹⁴, Willy Nojopranoto¹⁴, Roy H. Campbell⁹, Sayed Hadi Hashemi⁹, Juan A. Botia^{15,16}, John F. Carter¹⁷, Melina Maleknia¹⁸, David W. Craig¹⁰, Kendall Van Keuren-Jensen¹¹, Huw R. Morris^{2,3}, John A. Hardy^{2,3,19,20}, Cornelis Blauwendraat¹, Andrew B. Singleton^{1,4}, Faraz Faghri^{1,4,5}, Mike A. Nalls^{1,4,5} on behalf of the Accelerating Medicines Program - Parkinson's Disease (AMP PD) and the Global Parkinson's Genetics Program (GP2).

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

2 - Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, London, UK

3 - UCL Movement Disorders Centre, University College London, London, UK

4 - Center for Alzheimer's and Related Dementias, National Institutes of Health, Bethesda, MD, USA

5 - Data Tecnica International LLC, Glen Echo, MD, USA

6 - German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany

7 - School of Nursing, Virginia Commonwealth University, Richmond, VA, USA

8 - Geriatric Pharmacotherapy Program, School of Pharmacy, Virginia Commonwealth University, VA, USA

9 - Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA

10 - Institute of Translational Genomics, University of Southern California, Los Angeles, CA, USA

11- Neurogenomics Division, Translational Genomics Research Institute (TGen), Phoenix, AZ, USA

12 - Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA

13 - Preventive Neurology Unit, Wolfson Institute of Preventive Medicine, Queen Mary University of London

14 - Verily Life Sciences, South San Francisco, CA, USA

15 - Department of Molecular Neuroscience, UCL Queen Square Institute of Neurology, London, UK

16 - Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, Spain

17 - ModelOp, Chicago, IL, USA

18 - Georgia Institute of Technology, Atlanta, GA, USA

19 - UK Dementia Research Institute and Department of Neurodegenerative Disease and Reta Lila Weston Institute, London, UK

20 - Institute for Advanced Study, The Hong Kong University of Science and Technology, Hong Kong SAR, China

SUMMARY

Background

Personalized medicine promises individualized disease prediction and treatment. The convergence of machine learning (ML) and available multi-modal data is key moving forward. We build upon previous work to deliver multi-modal predictions of Parkinson's Disease (PD).

Methods

We performed automated ML on multi-modal data from the Parkinson's Progression Marker Initiative (PPMI). After selecting the best performing algorithm, all PPMI data was used to tune the selected model. The model was validated in the Parkinson's Disease Biomarker Program (PDBP) dataset. Finally, networks were built to identify gene communities specific to PD.

Findings

Our initial model showed an area under the curve (AUC) of 89.72% for the diagnosis of PD. The tuned model was then tested for validation on external data (PDBP, AUC 85.03%). Optimizing thresholds for classification, increased the diagnosis prediction accuracy (balanced accuracy) and other metrics. Combining data modalities outperforms the single biomarker paradigm. UPSIT was the largest contributing predictor for the classification of PD. The transcriptomic data was used to construct a network of disease-relevant transcripts.

Interpretation

We have built a model using an automated ML pipeline to make improved multi-omic predictions of PD. The model developed improves disease risk prediction, a critical step for better assessment of PD risk. We constructed gene expression networks for the next generation of genomics-derived interventions. Our automated ML approach allows complex predictive models to be reproducible and accessible to the community.

Funding

National Institute on Aging, National Institute of Neurological Disorders and Stroke, the Michael J. Fox Foundation, and the Global Parkinson's Genetics Program.

RESEARCH IN CONTEXT

Evidence before this study

Prior research into predictors of Parkinson's disease (PD) has either used basic statistical methods to make predictions across data modalities, or they have focused on a single data type or biomarker model. We have done this using an open-source automated machine learning (ML) framework on extensive multi-modal data, which we believe yields robust and reproducible results. We consider this the first true multi-modality ML study of PD risk classification.

Added value of this study

We used a variety of linear, non-linear, kernel, neural networks, and ensemble ML algorithms to generate an accurate classification of both cases and controls in independent datasets using data that is not involved in PD diagnosis itself at study recruitment. The model built in this paper significantly improves upon our previous models that used the entire training dataset in previous work¹. Building on this earlier work, we showed that the PD diagnosis can be refined using improved algorithmic classification tools that may yield potential biological insights. We have taken careful consideration to develop and validate this model using public controlled-access datasets and an open-source ML framework to allow for reproducible and transparent results.

Implications of all available evidence

Training, validating, and tuning a diagnostic algorithm for PD will allow us to augment clinical diagnoses or risk assessments with less need for complex and expensive exams. Going forward, these models can be built on remote or asynchronously collected data which may be important in a growing telemedicine paradigm. More refined diagnostics will also increase clinical trial efficiency by potentially refining phenotyping and predicting onset, allowing providers to identify potential cases earlier. Early detection could lead to improved treatment response and higher efficacy. Finally, as part of our workflow, we built new networks representing communities of genes correlated in PD cases in a hypothesis-free manner, showing how new and existing genes may be connected and highlighting therapeutic opportunities.

INTRODUCTION

For progressive neurodegenerative diseases, early and accurate diagnosis is key to effectively developing and using new interventions. This early detection paradigm aims to identify, analyze, and prevent or manage the disease before the patient recognizes signs and symptoms while the disease process is most amenable to intervention. Here we describe work that facilitates accurate and early diagnosis using cost-effective methods in a data-driven manner¹.

Biomedical researchers are currently at the convergence of scientific advances that will facilitate progress in early detection and remote identification of potentially high-risk individuals: first, the availability of substantial clinical, demographic, and genetic/genomic datasets. Second, advances in the automation of machine learning (ML) pipelines and artificial intelligence, to

maximize the value of this massive amount of readily available data ². Previous biomarker studies, particularly in neurodegenerative diseases, have focused on widely known statistical approaches and linear models, using a single metric or handful of metrics for predictions. We aim to contribute to the field with non-linear and ML-based approaches and leverage rapidly growing publicly available data to build these models. We also extend these models, providing not just disease prediction but also biological insight.

We have used publicly available multi-modal Parkinson's disease (PD) data to build an accurate peri-diagnostic model to predict disease risk. We also used the features nominated by our workflow to build novel, unbiased networks of genes related to the onset of PD that highlight biological pathways of interest and therapeutic targets. This work leveraged clinico-demographic and multi-omic data produced and curated to build models that can impact both trial recruitment and drug development. The models we have developed here improved performance over previous related efforts, with performance metrics at current cross-validation in withheld samples being equivalent, or in some cases, better than the training phase of earlier work¹. The data came from the Accelerating Medicines Program - Parkinson's Disease (AMP PD) program [<https://amp-pd.org/>] and the code used to carry out analyses comes from open-source automated ML software, all of which have been made publicly available to support reproducibility, transparency, and open science ^{3,4}.

METHODS

Online Appendix

Full methods and links to code can be found in the **Online Appendix**. A brief overview of the process can be found below. All of the data and code used for this project is also available at publication via a linked Terra workspace for full reproducibility [[link pending publication](#)]. GenoML is an open-source Python package automating machine learning workflows for genomics⁵. Source code and documentation is available at [<https://genoml.com/>] and on GitHub [<https://github.com/GenoML/genoml2>]. AMP PD data and quality control notebooks are access-controlled [<https://amp-pd.org/>]. Additionally, we have developed an interactive website [<https://share.streamlit.io/anant-dadu/shapleypdpredictiongenetics/main>] where researchers can investigate components of the predictive model.

Procedures and statistical analysis overview

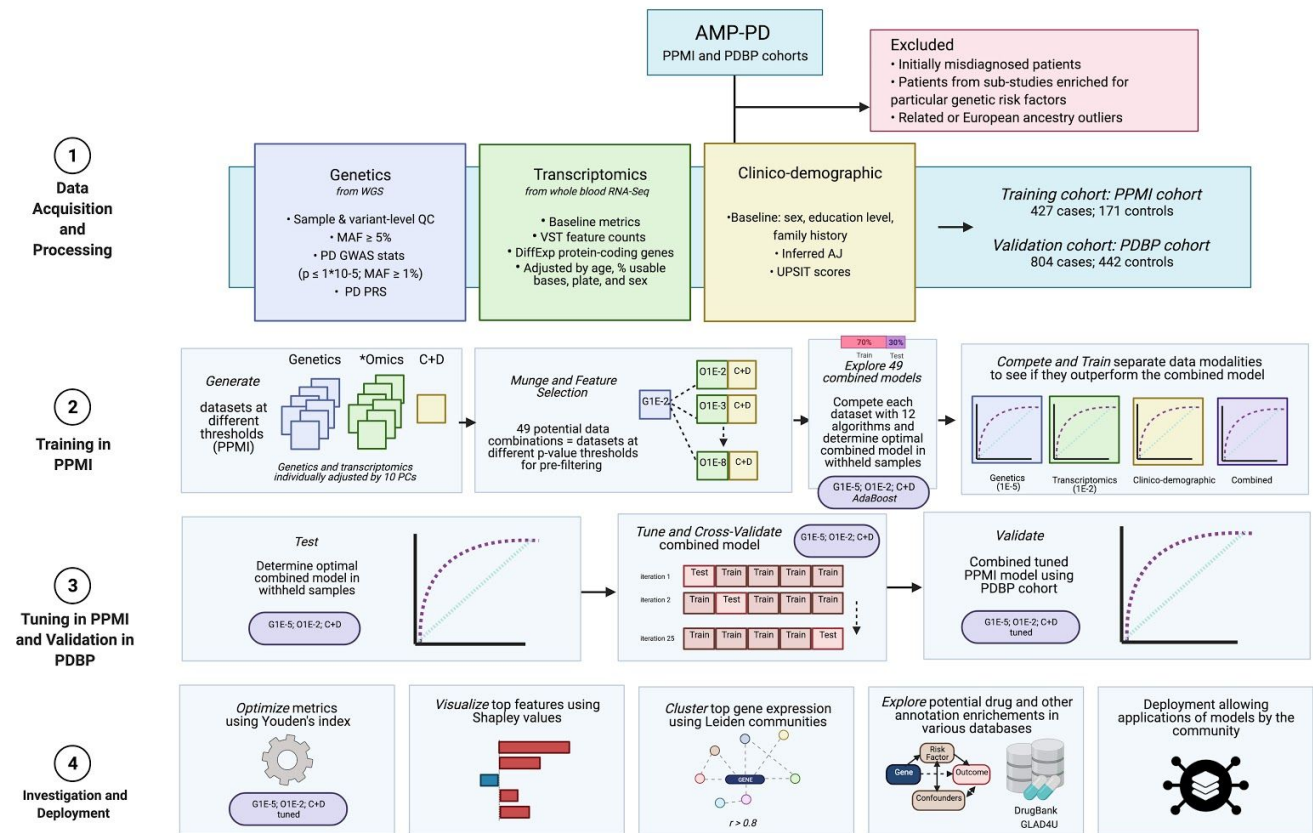
Figure 1 summarizes the workflow and data used in this project. Our workflow began with data munging that includes feature selection, adjustment, and normalization. Then we moved on to algorithm competition and feature selection based on a 70:30 (training:testing) split in the PPMI dataset. We then compared how each algorithm performed on identical training and testing data. Once the best performing algorithm had been selected, a thorough hyperparameter tuning of the algorithm with 5 fold cross-validation (also in the entire PPMI) was performed. Model was exported to enable external validation and transfer learning in the readers' own data. This hyperparameter tuning and cross-validation phase was carried out to both improve performance but also reduce bias ⁶. We validated the models built by taking the trained and tuned models

from PPMI and fit them to the external validation dataset, PDBP. Cohort summaries can be found below in **Table 1** with additional information in the **Online Appendix**.

Table 1: Descriptive statistics of studies included from AMP PD.

Study	Status	Age at baseline mean (SD)	UPSIT score mean (SD)	Male (%)	Positive family history of PD (%)	Inferred Ashkenazi ancestry (%)
<i>PPMI</i>	<i>Case</i>	61.75 (9.69)	23.48 (8.35)	65.57	25.53	6.09
	<i>Control</i>	60.61 (10.43)	34.18 (4.71)	63.74	5.85	11.11
<i>PDBP</i>	<i>Case</i>	64.59 (8.99)	19.65 (8.01)	64.18	24.88	3.61
	<i>Control</i>	62.87 (10.96)	32.52 (5.98)	45.25	8.14	4.07

Figure 1: Workflow and Data Summary. Scientific notation in the workflow diagram denotes minimum p-values from reference GWAS or differential expression studies as a pre-screen for feature inclusion.



Role of the funding source

The study's funders had no role in the study design, data collection, data analysis, data interpretation, or writing of the report. All authors and the public can access all data and statistical programming code used in this project for the analyses and results generation. MAN takes final responsibility for the decision to submit the paper for publication.

RESULTS

We have shown that integrating multiple modalities improved model performance in predicting PD diagnosis in a mixed population of cases and controls. Additional interpretation for ML metrics and models is included in the **Online Appendix**. Our multi-modality model showed a higher area under the curve (AUC; 89.72%) than just the clinico-demographic data available prior to neurological assessment (87.52%), the genetics-only model from genome sequencing

data and polygenic risk score (PRS; 70.66%), or the transcriptomics-only model from genome-wide whole blood RNA sequencing (RNAseq) data (79.73%) in withheld PPMI samples (see **Table 2** and **Figure 2** for summaries). This model's performance improved in both PPMI and PDBP after tuning, described below and in **Table 3**. Similar results can be seen when this model is validated in the PDBP data set (AUC from the combined modality model at 83.84% before tuning) detailed in **Table 4** and **Figure 3**. Additionally, the multi-modal model also had the lowest false positive and false negative rates compared to other models, only focusing on a single modality, in both the withheld test set in PPMI and in the PDBP validation set. Thus, moving from single to multiple data modalities yielded better results in not only AUC but across all performance metrics. A strength of using multi-modal approaches is that some modalities may better predict case or control status than others (**Table 2**). Here, we leveraged data diversity to increase overall sensitivity and specificity. Our final multi-modal model had higher accuracy and balanced accuracy at 85.56% and 82.41%, respectively, sensitivity at 89.31%, and specificity at 75.51%, when compared to models built only on a single data modality. Notably, this improved balanced accuracy is of particular importance in binary classifiers where one of the predicted classes is much rarer than the other, like PD, which is relatively infrequent in the general population. Special attention was given to validate the model, interpreting and visualizing the top features aiding in the prediction of classification, and further investigation into optimizing the model, developing hypothesis-free transcriptomic communities, and exploring potential drug-gene interactions.

Figure 2: Receiver operating characteristic curves and case probability density plots in withheld training samples at default thresholds comparing performance metrics in different data modalities from the PPMI dataset. P-values mentioned indicate the threshold of significance used per data type, except for the inclusion of all clinico-demographic features.

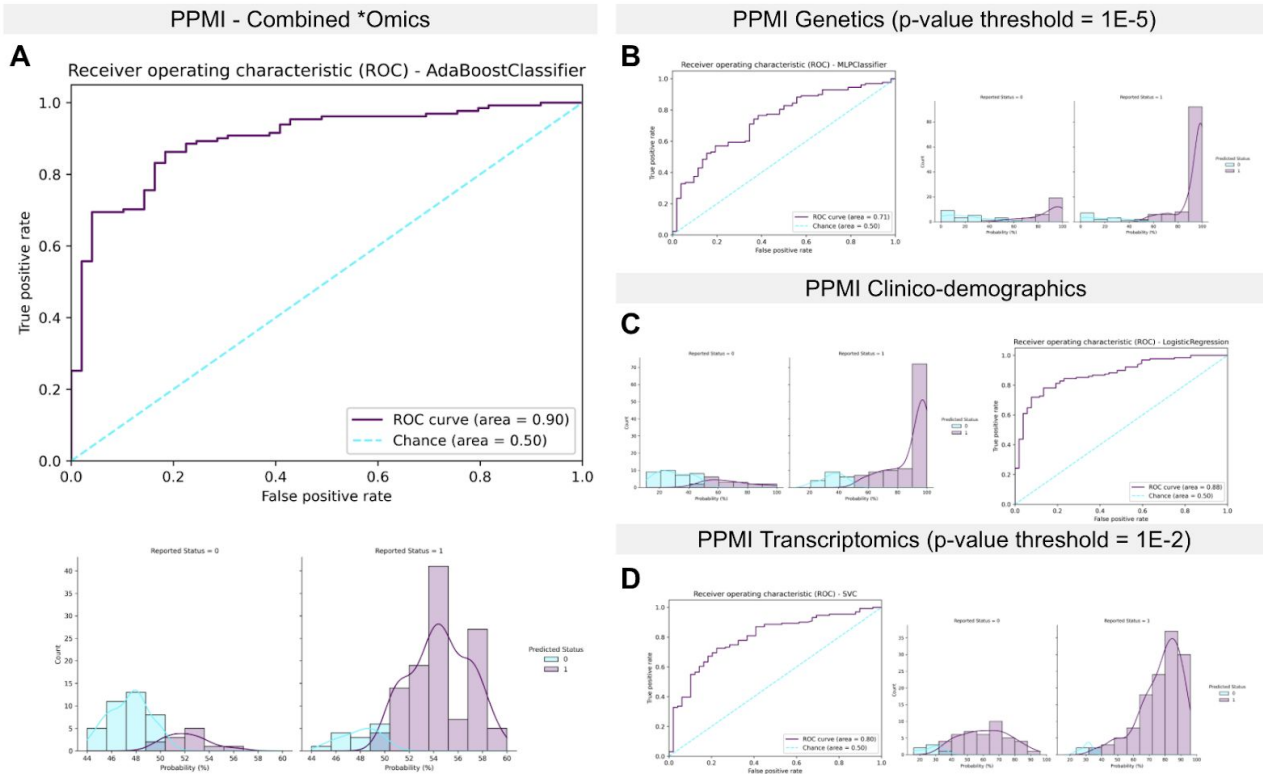


Table 2: Performance metric summaries comparing training in withheld samples in PPMI.

Data Modality	Stage	Algorithm	AUC (%)	Accuracy (%)	Balanced accuracy (%)	Log Loss	Sensitivity	Specificity	PPV	NPV
Genetics (P<1E-5)	Training in PPMI (70:30)	MLPClassifier	70.66	70.00	60.64	0.83	0.83	0.38	0.77	0.48
Clinico-demographic	Training in PPMI (70:30)	LogisticRegression	87.52	79.44	75.27	0.39	0.85	0.65	0.86	0.64
Transcriptomics (P<1E-2)	Training in PPMI (70:30)	SVC	79.73	73.89	54.60	0.48	0.97	0.12	0.75	0.60
Combined	Training in PPMI (70:30)	AdaBoostClassifier	89.72	85.56	82.41	0.63	0.89	0.76	0.91	0.73

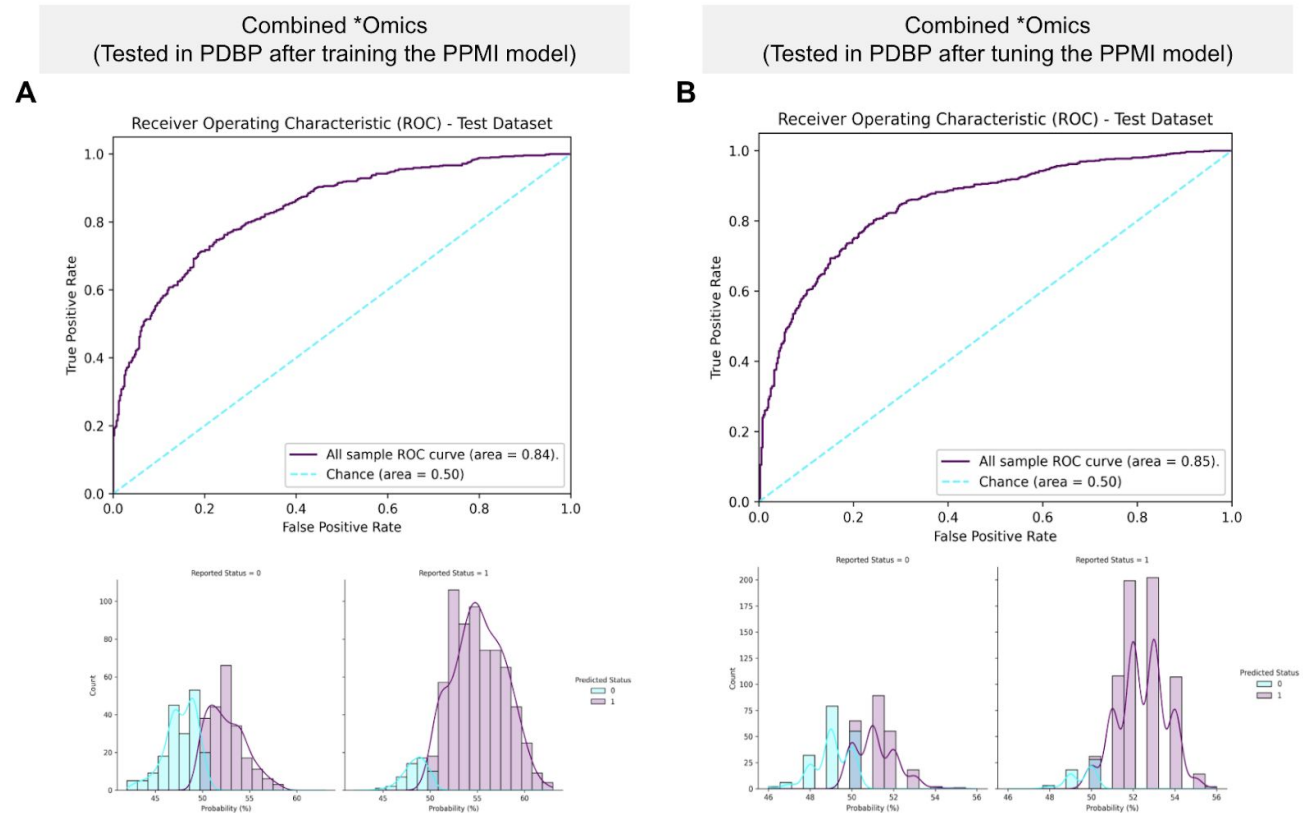
Table 3: Performance metric summaries comparing tuning in PPMI for different subsets of data modalities.

Data Modality	Stage	Algorithm	AUC at training (%)	Mean, AUC during CV for baseline model (%)	Standard deviation, AUC during CV for baseline model (%)	Min, AUC during CV for baseline model (%)	Max, AUC during CV for baseline model (%)	Mean, AUC during CV for tuned model (%)	Standard deviation, AUC during CV for tuned model (%)	Min, AUC during CV for tuned model (%)	Max, AUC during CV for tuned model (%)
Genetics (P<1E-5)	Tuning in PPMI	MLPClassifier	70.66	69.44	4.46	62.45	75.73	70.93	5.39	61.29	76.71
Clinico-demographic	Tuning in PPMI	LogisticRegression	87.52	88.51	2.17	86.19	91.98	88.55	2.20	86.33	92.15
Transcriptomics (P<1E-2)	Tuning in PPMI	SVC	79.73	78.05	4.27	71.49	82.62	79.01	4.71	70.88	84.01
Combined	Tuning in PPMI	AdaBoostClassifier	89.72	86.99	2.30	84.27	90.70	90.17	1.64	88.06	92.73

Table 4: Performance metric summaries comparing combined tuned and untuned model performance on the PDBP validation dataset.

Data Modality	Stage	Algorithm	AUC (%)	Accuracy (%)	Balanced Accuracy (%)	Log Loss	Sensitivity	Specificity	PPV	NPV
Combined	Untuned in PPMI as reference	AdaBoostClassifier	89.72	85.56	82.41	0.63	0.89	0.76	0.91	0.73
Combined; Untuned	Validation in PDBP	AdaBoostClassifier	83.84	75.81	69.31	0.64	0.93	0.46	0.75	0.78
Combined; Tuned	Validation in PDBP	AdaBoostClassifier	85.03	75.00	68.09	0.67	0.93	0.43	0.74	0.78

Figure 3: Receiver operating characteristic and case probability density plots in the external dataset (PDBP) at validation for the trained and then tuned models at default thresholds. Probabilities are predicted case status (r_1), so controls (a status of 0) skews towards more samples on the left, and positive PD cases (a status of 1) skews more samples on the right.



One benefit of the ML approach we have used is its ability to tune model parameters. The best performing tuned model that included all data showed an AUC distribution of 88.06% to 92.70% at 5-fold cross-validation with a mean of 90.20% and a standard deviation of 2.3% in PPMI (see **Table 3**). When validated in the PDBP data, we saw an AUC of 85.03%, sensitivity at 93.12%, and specificity at 43.07% for the tuned multi-modal model.

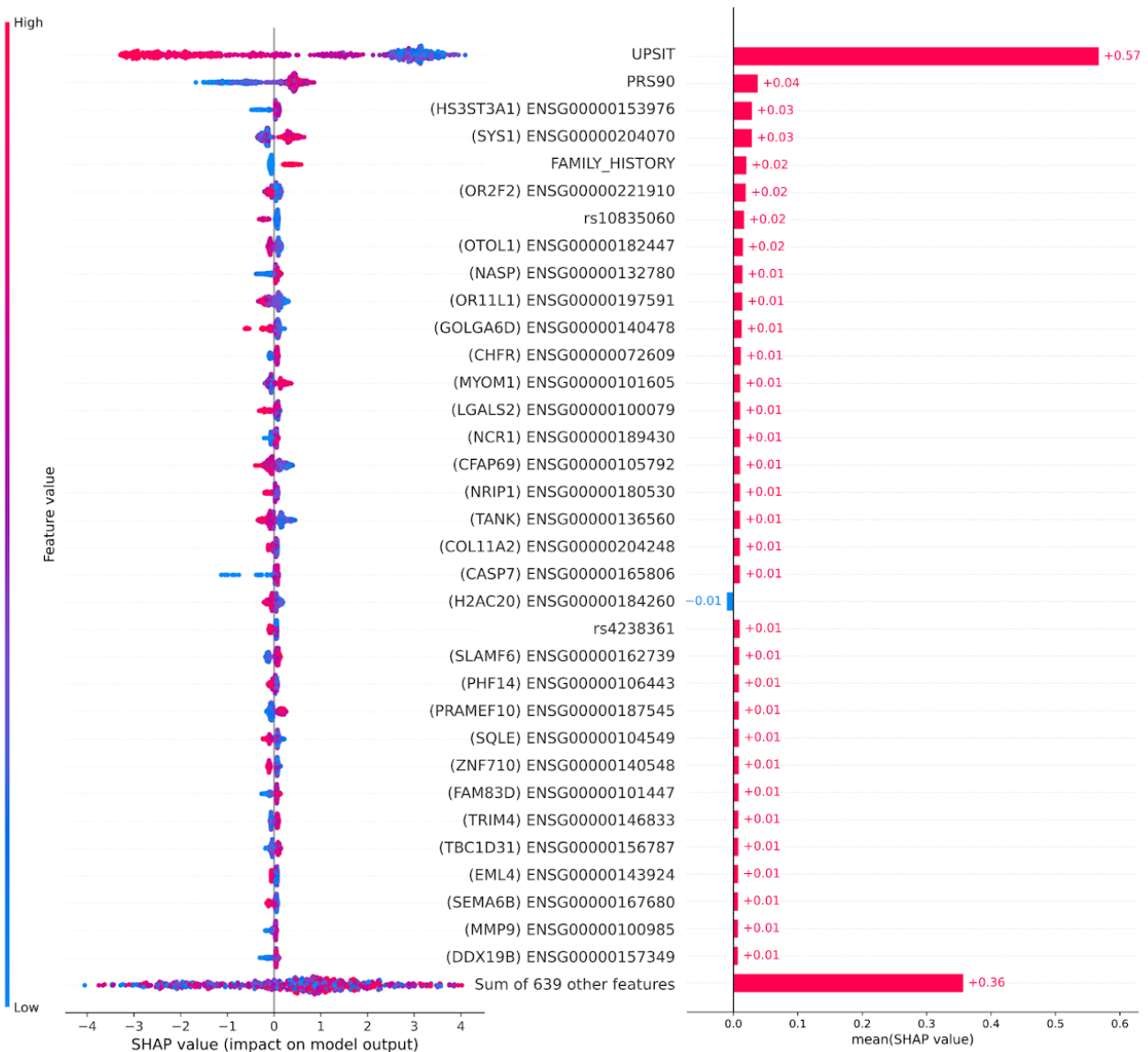
These models then improved further when *post-hoc* optimization of case probability thresholds was carried out. We considered the optimized version of the tuned model (including all data modalities) as our gold standard. When applied to withheld PPMI samples, the training phase model increased its balanced accuracy quantified performance to 83.95%. This optimization also led to improved balanced accuracy of 77.97% when fitting the tuned model referenced above to the PDBP validation data. See **Table 5** for details on other related metrics and a summary of optimized versus default thresholds. In general, our threshold optimization allowed general increases in classifier performance at a minimal computational cost.

Table 5: Optimizing the AUC threshold in withheld training samples and in the validation data.

Dataset	Model	Optimization	Case Probability Threshold (%)	Accuracy (%)	Balanced accuracy (%)	Log loss	Sensitivity	Specificity	PPV	NPV
PPMI, withheld samples	Training phase	optimized	51	85	83.95	0.05	0.86	0.82	0.93	0.69
PPMI, withheld samples	Training phase	default	50	85.56	82.41	0.05	0.89	0.76	0.91	0.73
PDBP, external test samples	Tuned model	optimized	51	78.58	77.97	0.07	0.80	0.76	0.85	0.68
PDBP, external test samples	Tuned model	default	50	75	68.09	0.09	0.93	0.43	0.74	0.78

Our model build included 71 SNPs and 596 protein-coding transcripts in addition to expected features like the demographics, family history, olfactory function, and previous genome-wide significant polygenic risk estimates in the form of PRS⁷. The Shapley Additive exPlanation (SHAP) plots in **Figure 4** show the relative importance of the features in the model approximated using withheld training data. When investigating the SHAP values for both the training and testing samples, the UPSIT score, as well as PRS, contributed most to the predictive power of the model, but the accuracy of these are supplemented by many smaller effect transcripts and risk SNPs. It also indicates that the lower UPSIT score (designated by the blue color on the left-hand side) value corresponds to a higher probability of PD, as most of the blue-colored points lie on the right side of the baseline risk estimate. Looking closer at these features, we can also observe that the directionality of different genetic features is not uniform. This signifies that over-expression of some genes corresponds to healthy controls while for some features it is in the opposite direction. We have also created a website that allows readers to further explore feature contributions to model accuracy in various scenarios [<https://share.streamlit.io/anant-dadu/shapleypdpredictiongenetics/main>]. The addition of SNPs outside the PRS could suggest potential compensatory or risk modification effects interacting with the PRS. For more information on pairwise interactions between the PRS and individual SNPs, please see the **Online Appendix**.

Figure 4: Feature importance plots for top 5% of features in withheld training data.



Gene expression network communities were constructed using RNAseq data extracted from positive PD cases. These genes were nominated by the feature selection process. These communities of genes represent part of a potentially novel PD-specific network derived from whole blood RNAseq. Consider the network itself to be conceptually similar to a pathway, composed of genes whose expression was strongly correlated in the case-only transcriptomics dataset, and the communities being subgroups of closely related genes within the larger set. We identified 13 network communities consisting of 300 genes with an Erdos-Renyi modularity score of 0.794 (a modularity score closer to 1 indicates better model fit). A link to the full community annotations and a graphical summary can be found in the **Online Appendix**.

For genes in our network communities, we evaluated the potential over representation of known drug target genes across the identified communities. When comparing the genes defined as part of the network communities (N = 300) to those selected for inclusion as features in the case:control model build (N = 598), we noted enrichments of genes connected to fostamatinib (FDR adjusted p-value 2.21×10^{-4} , for genes *MYLK*, *EPHA8*, *HCK*, *DYRK1B*, and *BUB1B-PAK6*) and copper (FDR adjusted p-value 0.0286, for genes *HSP90AA*, *CBX5* and *HSPD1*) from the DrugBank annotations. The same query in the GLAD4U database resulted in a significant over representation of l-lysine annotated genes (FDR adjusted p-value 0.0057, for genes *DDX50*, *UBA2*, *ESCO1*, *CDC34*, *ANKIB1*, *PCMT1*, *DNAJA1*, *PRMT3*, *ASPSCR1*, *BRDT*, *LOXL4*, *CBX5*, *HAT1*, *MARCH1*, *HSP90AA1*, *KPNB1*, *KMT5B*, *PSIP1*, *XPOT*, *SLC7A9*, *ZNF131*, *DDX18*, *RBBP5*, and *MSL1*). All other variations of our drug target enrichment analyses yielded no significant drugs overrepresented after multiple test correction, although the top ranked result was consistently gamma hydroxybutyric acid (unadjusted p-values 0.0056-0.0001 for genes *SLC16A7*, *SLC16A3*, and *GABBR1*).

DISCUSSION

In an era where genetics and genomics combined with clinico-demographic data are increasingly available to researchers, we can now build multi-modal models at a scale that use multiple data modalities for increased performance.

This study illustrates that integrating diverse data modalities into modeling efforts can improve the quality of predictions. This paper is a proof of concept for future directions in this area of predictive modeling in large healthcare data and indicative of its relevance to other areas such as clinical trial enrollment and stratification. In our modeling process, we have succeeded in building robust model(s) of peri-diagnostic PD while also generating *de novo* network communities of genes correlated in PD cases, providing further data for potential therapeutic development. A strength of using this multi-modal approach is that the different modalities compensate for one another, with some modalities better at predicting case status while others are better at classifying controls. All of this was accomplished in a completely transparent and open science framework, from the underlying data to code and resulting models.

While we do not suggest this as a replacement for current diagnostic screening methods, it can be a potential adjunct screening that could aid in identifying high-risk individuals, especially on a large biobank or study recruitment scale. Additional studies will need to be conducted to ascertain the model's ability to distinguish very early PD cases from other diseases within high risk cohort studies. The estimated prevalence of PD in an aging population is about 2%⁸. At this prevalence for our optimized PPMI model described in **Table 5**, the positive predictive value (PPV) and negative predictive values (NPV) were calculated to be 8.75% and 99.66%, respectively. The false discovery rate (FDR) and false omission rates (FOR) at this prevalence are 91.26% and 0.34%, respectively. The low FOR indicates that for every 1000 individuals who are classified as healthy controls, there are likely 3 to 4 missed positive PD cases. Using this optimized combined model and accounting for the estimated prevalence of PD, about ten times the number of individuals would be flagged as high risk or a potential PD case for every real

case, indicating that this model is best suited to identifying large groups of individuals to monitor within a health registry or biobank to prioritize for further testing.

The strength of our work is its high balanced accuracy in delineating cases and controls. Its other strength is its applicability and utility across datasets (further details on this in the **Online Appendix** for transfer learning). This model has the potential to be used in large healthcare system settings to identify at-risk individuals for potential monitoring as well as nominating future candidates for various preventative interventions or clinical trial enrollments. Our diagnostic model includes both time-varying (age, UPSIT, and RNAseq data) and time static (biological sex, family history, and genetics) features that likely peak accuracy at the time of diagnosis due to model training on the PPMI dataset, which includes only newly diagnosed imaging-confirmed and unmedicated cases. Additionally, the ability to refine the phenotype of the participant group based on a combination of clinician and algorithmic insight will benefit trial recruitment and could only increase the efficacy of a trial. Finally, since our model is diagnostic in nature and designed to target PD early in the disease course, it may be beneficial in helping get treatments or interventions to patients before irreparable damage has been done as large pools of at risk individuals can be flagged for follow-up and closer monitoring for potential symptom onset ⁹.

We have created an interactive web-based application for others to investigate the driving factors in our best model that incorporates all data modalities; it also gives users the flexibility to explore variations like transcriptomics-only models or a new model with none of the clinico-demographic features present (see **Online Appendix** for details). For the combined model that we have focused on describing in this report; decision plots are provided, these are useful as the web application is capable of letting users explore how and why individuals that were difficult to classify were labeled as cases or controls. The web application (as well as **Figure 4**) shows that the UPSIT score, in general, was the strongest factor in deciding if an individual was classified to be a positive PD case or healthy control by the model. As an example, a decision plot shows that a sample that was clinically diagnosed to be a PD case, we see that most of the features seemed to indicate that the individual was about to be classified as a PD case by the model, but ultimately an unexpectedly high UPSIT score misclassified the individual as a healthy control (decision plots work to visualize the path a model takes before arriving at a classification; see the additional figure in the **Online Appendix** for a graphical representation of misclassification). In general, UPSIT itself accounts for roughly half of the decision making process in our model, and in some instances, is a blessing and a curse with regard to model performance.

Genes and variants affecting the model's performance shown in **Figure 4** may have some impact on PD biology. Many of the top features we nominated that are shown in **Figure 4** are transcriptomic in nature; this enrichment of transcriptomic data in the top of the feature importance plots may be due to the PRS accounting for a substantial part of the strongest purely genetic aspects of PD risk. Some interesting biologically plausible connections can be drawn from these highly ranked features. For example, the most impactful feature from the transcriptomics data, the expression of gene *HS3ST3A1*, has been implicated in α -Synuclein aggregation in PD cellular models, as well as having been recently part of a novel GWAS

finding associated with white matter hyperintensity burden in elderly populations (along with some aspects of cognitive decline)^{10,11}. Another top ranking transcriptomic feature, *OTOL1*, has been suggested before as a putative genetic modifier of familial PD age at onset¹². *CHFR* has been associated in previous studies with rotenone related PD risk¹³. *CASP7* is potentially biologically interesting due to its expression being implicated in apoptosis and neuroprotection as well as rare missense mutations in the gene being associated with late onset familial Alzheimer's^{14,15}. The genetic variant, rs4238361 is a potential PD risk modifying variant whose nearest coding gene is *VPS13C*, a gene that harbors both rare and common PD variants of interest^{16,17}. The gene *PHF14* has been suggested to be downregulated in neurodegenerative diseases, this potential effect mirrors that suggested in **Figure 4**¹⁸. Recent single-cell sequencing work has provided evidence for a connection between *SQLE* and dopamine stress responses in neurons relating to PD risk¹⁹. Additionally, *MMP9* overexpression has been suggested to be associated with neuronal cell death in neurodegeneration²⁰.

Another strength of this work is that feature selection from model building easily segues into network community analyses to build relatively low bias networks, compared to those with potential bias taken from literature and text mining^{21,22}. Here we can push therapeutic and biomarker research by identifying communities of connected genes in the blood transcriptome of PD patients. Nodes in these networks suggest shared effects in genetically targeted drugs, informing development cycles and benefitting developers as drugs connected to genetic or genomic data often have a higher level of success in trials compared to those without similar evidence^{23,24}.

Modeling exercises like these not only have the potential to build useful classifiers, they may also identify drug targets. This can happen at the feature selection, and network build phases. In our network community build based on case-only expression data, only two quantitative trait loci in blood from Mendelian randomization in the previously published PD GWAS were included. These two genes are *ZBTB4* and *FCF1*. What may be more interesting is the enriched drug targets within the nominated genes from our analyses of the transcriptomic data. This result is not truly surprising as our network communities are based on genes highly correlated in cases only, and aimed to build clusters of genes connected by similar expression patterns among cases. More interestingly are the findings from the network communities recognizing the overrepresentation of genes targeted by known drugs. For instance, in this study, we uncovered an interaction between gamma hydroxybutyric acid and *SLC16A7*, *SLC16A3*, and *GABBR1* genes. The *SLC16A7* and *SLC16A3* genes are a part of a family of drug transporter genes known as monocarboxylate transporters²⁵. Drug transporters have a role in almost every part of the therapeutic process, from absorption, distribution, and elimination of drug molecules. The *GABBR1* gene encodes a receptor for gamma-aminobutyric acid (GABA) expressed throughout the brain; defects in this gene underlie several neurobehavioral diseases^{25,26}. Gamma hydroxybutyric acid acts as an agonist, activating GABA-B receptors to exert its sedative effects. Identification of drug metabolism and receptor gene/drug interactions may lead to and drug discovery, thereby helping us optimize drug therapy.

Our main weakness in this research is the lack of diversity in available sample series. Current research suggests that genetic predictive models have mixed results when being applied across genetic ancestry groups²⁷. With subsequent iterations of this work being facilitated by the Global Parkinson's Genetics Program (GP2) program over the next five years^{28,29}, we hope to expand this modeling effort into a diverse set of genetic ancestry groups and generally in larger sample series. We also acknowledge that no optimal dataset to validate the findings from PPMI exists because of the inherent study design of PPMI focusing on unmedicated recently diagnosed PD cases. Clearly, the ongoing extension of the PPMI study will facilitate further work.

Overall, we believe this work represents a significant conceptual and scientific advance past previous efforts. This classifier has improved performance, is more broadly applicable, and is highly reproducible. Further, the transparency of this approach and the contributions of data types move the field away from black-box predictors of disease. A further strength of this work is the use of open-source automated ML software thoughtfully designed for scientists and scientific data, developing models and validating them on public controlled-access datasets, visualizing the top contributing features, and providing all the code and software publicly.

This work has helped to push past the previous paradigm of focusing on a single biomarker or class of biomarker in biomedical research to maximize data value for clinical and computational scientists by leveraging ML algorithms that explore complex relationships between features. We have provided a model(s) to improve risk prediction in PD to help with interventional and prospective studies as well as healthcare resource prioritization. We have also integrated additional analyses and data resources that may aid in developing and/or refining future interventions.

CONTRIBUTORS

MBM, FF, and MAN contributed to the concept and design of the study. MBM, HL, DV, HI, IV, EH, DS, JJK, YS, MB, DWC, KVJ, CB, ABS, and MAN were involved in the acquisition of data, data generation, and data cleaning. MBM, HL, DV, HI, FF, AD, LS and MAN did the analysis and interpretation of data. MBM, HL, DV, HI, LS, AD, IV, EH, DS, SBC, JJK, YS, MB, WN, SHH, RHC, MM, JAB, DWC, KVJ, HRM, JAH, CB, ABS, and MAN contributed to the drafting of the article and revising it critically.

DECLARATION OF INTERESTS

HL, HI, FF, DV, YS, 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. HRM is employed by UCL. 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, 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).

ACKNOWLEDGEMENTS

This work was supported in part by the Intramural Research Program of the National Institute on Aging and National Institute of Neurological Disorders and Stroke (project number Z01-AG000949-02). Data used in the preparation of this article were obtained from the AMP PD Knowledge Platform. For up-to-date information on the study, <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, Sanofi, and Verily. Clinical data and biosamples used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI), and the Parkinson's Disease Biomarkers Program (PDBP). PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including full names of all of the PPMI funding partners found at www.ppmi-info.org/fundingpartners. The PPMI Investigators have not participated in reviewing the data analysis or content of the manuscript. 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 manuscript. PDBP sample and clinical data collection is supported under grants by NINDS: U01NS082134, U01NS082157, U01NS082151, U01NS082137, U01NS082148, U01NS082133. A portion of the resources used in the preparation of this article were obtained from Global Parkinson's Genetics Program (GP2). GP2 is funded by the Aligning Science Against Parkinson's (ASAP) initiative and implemented by The Michael J. Fox Foundation for Parkinson's Research (<https://parkinsonsroadmap.org/gp2/>). For a complete list of GP2 members, see <https://parkinsonsroadmap.org/gp2/>. Workflow diagram was created with BioRender.com.

REFERENCES

- 1 Nalls MA, McLean CY, Rick J, *et al.* Diagnosis of Parkinson's disease on the basis of clinical and genetic classification: a population-based modelling study. *Lancet Neurol* 2015; **14**: 1002–9.
- 2 Sudlow C, Gallacher J, Allen N, *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015; **12**: e1001779.
- 3 GenoML · Automated Machine Learning (AutoML) for Genomics. <https://genoml.github.io/index.html> (accessed Nov 11, 2020).
- 4 Home. <https://amp-pd.org/> (accessed Nov 11, 2020).
- 5 Makarios MB, Leonard HL, Vitale D, *et al.* GenoML: automated machine learning for genomics. arXiv:210303221 [cs, q-bio] 2021; published online March 4. (accessed March 5, 2021).

- 6 Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. 2019; published online Aug 23. <http://arxiv.org/abs/1908.09635> (accessed Feb 21, 2021).
- 7 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**. DOI:10.1016/S1474-4422(19)30320-5.
- 8 Alves G, Forsaa EB, Pedersen KF, Dreetz Gjerstad M, Larsen JP. Epidemiology of Parkinson's disease. *J Neurol* 2008; **255 Suppl 5**: 18–32.
- 9 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**: 331–8.
- 10 Lehri-Boufala S, Ouidja M-O, Barbier-Chassefière V, *et al.* New Roles of Glycosaminoglycans in α -Synuclein Aggregation in a Cellular Model of Parkinson Disease. *PLoS One* 2015; **10**: e0116641.
- 11 Genome-wide association study of white matter hyperintensity volume in elderly persons without dementia. *NeuroImage: Clinical* 2020; **26**: 102209.
- 12 Hill-Burns EM, Ross OA, Wissemann WT, *et al.* Identification of genetic modifiers of age-at-onset for familial Parkinson's disease. *Hum Mol Genet* 2016; **25**: 3849–62.
- 13 Cabeza-Arvelaiz Y, Schiestl RH. Transcriptome Analysis of a Rotenone Model of Parkinsonism Reveals Complex I-Tied and -Untied Toxicity Mechanisms Common to Neurodegenerative Diseases. *PLoS One* 2012; **7**: e44700.
- 14 Zhang X, Zhu C, Beecham G, *et al.* A rare missense variant of CASP7 is associated with familial late-onset Alzheimer's disease. *Alzheimers Dement* 2019; **15**. DOI:10.1016/j.jalz.2018.10.005.
- 15 Magalingam KB, Radhakrishnan A, Ramdas P, Haleagrahara N. Quercetin glycosides induced neuroprotection by changes in the gene expression in a cellular model of Parkinson's disease. *J Mol Neurosci* 2015; **55**. DOI:10.1007/s12031-014-0400-x.
- 16 Lesage S, Drouet V, Majounie E, *et al.* Loss of VPS13C Function in Autosomal-Recessive Parkinsonism Causes Mitochondrial Dysfunction and Increases PINK1/Parkin-Dependent Mitophagy. *Am J Hum Genet* 2016; **98**: 500–13.
- 17 Rudakou U, Ruskey JA, Krohn L, *et al.* Analysis of common and rare variants in late-onset Parkinson disease. *Neurol Genet* 2020; **6**: 385.
- 18 Ibáñez K, Boulossa C, Tabarés-Seisdedos R, Baudot A, Valencia A. Molecular Evidence for the Inverse Comorbidity between Central Nervous System Disorders and Cancers Detected by Transcriptomic Meta-analyses. *PLoS Genet* 2014; **10**. DOI:10.1371/journal.pgen.1004173.
- 19 Single-Cell Transcriptomics of Parkinson's Disease Human In Vitro Models Reveals Dopamine Neuron-Specific Stress Responses. *Cell Rep* 2020; **33**: 108263.
- 20 He X, Zhang L, Yao X, *et al.* Association Studies of MMP-9 in Parkinson's Disease and Amyotrophic Lateral Sclerosis. *PLoS One* 2013; **8**: e73777.

- 21 Gillis J, Ballouz S, Pavlidis P. Bias tradeoffs in the creation and analysis of protein-protein interaction networks. *J Proteomics* 2014; **100**: 44–54.
- 22 Haynes WA, Tomczak A, Khatri P. Gene annotation bias impedes biomedical research. *Sci Rep* 2018; **8**: 1362.
- 23 Nelson MR, Tipney H, Painter JL, *et al.* The support of human genetic evidence for approved drug indications. *Nat Genet* 2015; **47**: 856–60.
- 24 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.
- 25 Halestrap AP. The SLC16 gene family - structure, role and regulation in health and disease. *Mol Aspects Med* 2013; **34**: 337–49.
- 26 Ngo D-H, Vo TS. An Updated Review on Pharmaceutical Properties of Gamma-Aminobutyric Acid. *Molecules* 2019; **24**. DOI:10.3390/molecules24152678.
- 27 Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Current clinical use of polygenic scores will risk exacerbating health disparities. DOI:10.1101/441261.
- 28 GP2: The Global Parkinson's Genetics Program. *Mov Disord* 2021; published online Jan 29. DOI:10.1002/mds.28494.
- 29 Riley EAU, AU Riley E, Schekman R. Open science takes on Parkinson's disease. *eLife*. 2021; **10**. DOI:10.7554/elife.66546.

TABLE OF CONTENTS

Online repository of code and models for transfer learning	2
Online Repository	2
Network Communities	2
Online Methods	2
Study design and participants	2
Table 1: Descriptive statistics of studies included from AMP PD	3
Data munging	4
Feature and model selection	5
Post-hoc optimization for class imbalance	6
Network communities	6
ML metrics and model interpretation	6
Details on data from the Accelerating Medicines Partnership - Parkinson's Disease (AMP-PD) and the Global Parkinson's Genetic Project (GP2)	7
Cohort Summaries - PPMI	7
Study Overview for PPMI	7
Study Inclusion Criteria for PPMI - PD Cases	8
Study Exclusion Criteria for PPMI - PD Cases	8
Study Inclusion Criteria for PPMI - Healthy Controls	8
Study Exclusion Criteria for PPMI - Healthy Controls	9
Cohort Summaries - PDBP	9
Study Overview for PDBP	9
Study Inclusion Criteria for PDBP - PD Cases	9
Study Exclusion Criteria for PDBP - PD Cases	9
Study Inclusion Criteria for PDBP - Healthy Controls	10
Study Exclusion Criteria for PDBP - Healthy Controls	10
Age distribution per cohort	11
<i>Supplemental Figure 1: Age distributions of cohorts, broken down by cases and controls</i>	11
Network Graphical Summary of Nominated Genes	11
<i>Supplemental Figure 2: Network plot of nominated genes</i>	12
Misclassified cases by the best performing model	12
Supplemental Figure 3: Misclassified case as a healthy control using the best model	13
Additional References	13

Online repository of code and models for transfer learning

Online Repository

The [link to the GitHub repository \(https://github.com/GenoML/GenoML_multimodal_PD/\)](https://github.com/GenoML/GenoML_multimodal_PD/) includes the following:

- Figures referenced in the manuscript
- Code used for manuscript
- Tables referenced in the manuscript
- Additional supplementary figures and tables
- Links and references to the main software used (GenoML)
- Trained and tuned models and their associated performance metrics for each model referenced in the manuscript for the community to be able to deploy and use
- An example of how to run these trained and tuned models for transfer learning
- Link to the interactive web application for the community to investigate further

Network Communities

The network community code can be found on the **Online Repository** [here](#). Annotated community members plus additional sparse annotations from Nalls et al. 2019 can be found [here](#). Page ranks and similar annotations for genes comprising the network communities can be found [here](#).

Online Methods

Study design and participants

Clinical, demographic plus genome-wide DNA and RNA sequencing data were taken at baseline visits from the Parkinson's Progression Marker Initiative (PPMI) and the Parkinson's Disease Biomarkers Program (PDBP) in cases with PD and controls unaffected by neurologic diseases. Since our model is retrospective, we aimed only to analyze refined Parkinson's disease diagnosis, by excluding any samples with conflicting diagnostic data within a decade of post-enrollment follow-up. We excluded any case whose medical history included an additional neurological disease diagnosis or retraction of their PD diagnosis during follow-up. We also excluded controls developing PD or another neurodegenerative disease(s) after enrollment. Additionally, a subset of Parkinson's disease cases and controls from the PPMI study were excluded as they came from a targeted study recruitment design purposely enriching for known genetic risk mutation carriers (*LRRK2* and *GBA* mutation carrier focused recruitment).

Participants with required clinical, demographic and genomic (DNA and RNA sequencing) data were identified for inclusion, with excessive missing data (> 15% per feature) as exclusion criteria. 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.

Clinical and demographic data ascertained as part of this project included age at diagnosis for cases and age at baseline visit for controls. Family history (self-reporting if a first or second-degree relative has a Parkinson's disease diagnosis) was also a known clinico-demographic feature of interest. Ashkenazi Jewish status was inferred using principal component analysis comparing those samples to a genetic reference series ¹. Sex was clinically ascertained but also confirmed using X chromosome heterozygosity rates. The University of Pennsylvania Smell Inventory Test (UPSIT) was used in modeling ². For a summary of basic clinical and demographic features, please refer to **Table 1**.

Table 1: Descriptive statistics of studies included from AMP PD.

Study	Status	Age at baseline mean (SD)	UPSIT score mean (SD)	Male (%)	Positive family history of PD (%)	Inferred Ashkenazi ancestry (%)
<i>PPMI</i>	<i>Case</i>	61.75 (9.69)	23.48 (8.35)	65.57	25.53	6.09
	<i>Control</i>	60.61 (10.43)	34.18 (4.71)	63.74	5.85	11.11
<i>PDBP</i>	<i>Case</i>	64.59 (8.99)	19.65 (8.01)	64.18	24.88	3.61
	<i>Control</i>	62.87 (10.96)	32.52 (5.98)	45.25	8.14	4.07

DNA sequencing data were generated using Illumina's standard short-read technology, and the functional equivalence pipeline during alignment was the Broad Institute's implementation³. Jointly genotyped sequencing data using the standard GATK pipeline from AMP-PD was used. This process is described in detail, from sample prep to variant calling, in a separate manuscript detailing the AMP PD whole-genome DNA sequencing effort [under review] ⁴.

Quality control for these samples based on genetic data output by the pipeline included the following inclusion criteria: concordance between genetic and clinically ascertained genders, call rate > 95% at both the sample and variant levels, heterozygosity rate < 15%, freemix estimated contamination rate < 3%, transition:transversion ratio > 2, unrelated to any other sample at a level of the first cousin or closer (identity by descent < 12.5%) and genetically ascertained European ancestry. For inclusion of whole-genome DNA sequencing data, the variants must have passed basic quality control as part of the initial sequencing effort (PASS flag from the joint genotyping pipeline) as well as meeting the following criteria: non-palindromic alleles, missingness by case-control status $P > 1E-4$, missingness by haplotype $P > 1E-4$, Hardy-Weinberg p-value > $1E-4$, minor allele frequency in cases > 5% (in the latest Parkinson's disease meta-GWAS) ⁵. As an *a priori* genetic feature to be included in our modeling efforts, we

also used the basic polygenic risk score from the latest Parkinson's disease meta-GWAS (genome-wide significant loci only) that did not include our testing or training samples as weights ⁵.

RNA sequencing data from whole blood on the same samples was generated by the Translational Genomics Research Institute team using standard protocols for the Illumina NovaSeq technology ⁶. For this study, we focused on blood withdrawn at baseline. Variance stabilized counts were adjusted for experimental covariates using standard limma pipelines⁷. Gene expression counts for protein-coding genes were extracted, then differential expression p-values were calculated between cases and controls using logistic regression adjusted for additional covariates of sex, plate, age, ten principal components, and percentage usable bases.

This project is carried out entirely within an open science framework, and the code and data underlying the summary above can be found in linked notebooks and datasets described in the **Online Repository**.

Data munging

As part of the initial data munging, principal components summarizing genetic variation in DNA and RNA sequencing data modalities are generated separately. For the DNA sequencing, ten principal components were calculated based on a random set of 10,000 variants sampled after linkage disequilibrium pruning that kept only variants with $r^2 < 0.1$ with any other variants in +/- 1MB. As a note, these variants were not p-value filtered based on recent GWAS, but they do exclude regions containing large tracts of linkage disequilibrium⁸. For RNA sequencing data, all protein-coding genes' read counts per sample were used to generate a second set of 10 principal components. All potential features representing genetic variants (in the form of minor allele dosages) from sequencing were then adjusted for the DNA sequence-derived principal components using linear regression, extracting the residual variance. The same was done for RNA sequencing data using RNA sequencing derived principal components. This way, we statistically account for latent population substructure and experimental covariates at the feature level to increase generalizability across heterogeneous datasets. In its simplest terms, all transcriptomic data were corrected for possible confounders, and the same is done for genotype dosages. After adjustment, all continuous features were then Z transformed to have a mean of 0 and a standard deviation of 1 to keep all features on the same numeric scale when possible. Once feature adjustment and normalization were complete, internal feature selection was carried out in the PPMI training dataset using extra decision trees to identify features contributing information content to the model while reducing the potential for overfitting ^{9,10}.

Feature and model selection

After the data munging process (quality control, feature selection, adjustment, and scaling) described above, data from PPMI was randomly split into 70% training and 30% testing. Training of the algorithms was performed on the training set, and validation of the algorithms was performed on the testing set. A total of 12 well-performing ML algorithms were competed to

identify which algorithm could maximize AUC across the two classes (cases and controls). These algorithms were chosen due to their success in other domains, execution in Python's scikit-learn package, and their ability to export probability-based predictions, allowing the training, testing, and interpretation of the model more straightforward. The algorithms included are: logistic regression (LogisticRegression), random forests (RandomForestClassifier), adaptive boosting (AdaBoostClassifier), gradient boosting (GradientBoostingClassifier), stochastic gradient descent (SGDClassifier), support vector machines (SVC), multi-layer perceptron neural networks (MLPClassifier), k-nearest neighbors (KNeighborsClassifier), linear discriminant analysis (LinearDiscriminantAnalysis), quadratic discriminant analysis (QuadraticDiscriminantAnalysis), bagging (BaggingClassifier) and extreme gradient boosting (XGBClassifier). The algorithm with the highest AUC and balanced accuracy in the withheld 30% of PPMI was selected for tuning and cross-validation. Tuning is the process in which multiple algorithm hyperparameters, such as learning rate, are tested to optimize performance. The best hyperparameters were chosen through cross-validation, a technique that estimates model performance on unseen data by training and testing the model on different splits of the dataset. The top competing algorithm was then selected to undergo a computationally intensive hyperparameter tuning phase in the entire PPMI dataset, no longer split into training and testing once, instead, undergoing cross-validation each time parameters were iterated. In this analysis, the top-performing algorithm (AdaBoostClassifier) was tuned for several potential predictors (estimators) between 1 and 1000 for 25 random iterations at 5-fold cross-validation per iteration.

This process detailed in the paragraphs above was carried out 49 times, at varying thresholds of p-values based feature inclusion thresholds. We iterated across all possible combinations of p-value thresholds [1E-2, 1E-3, 1E-4, 1E-5, 1E-6, 1E-7, 1E-8] for genetic data from the most recent published GWAS and for transcriptomic data from our differential expression work also described above⁵. Once the pipeline above was run for all 49 threshold combinations, we picked the p-value thresholds and algorithm that performed best in the training dataset for tuning at cross-validation and external validation in PDBP to evaluate its generalizability and performance (all 49 models are available in the **Online Repository**). Prior to starting the modeling process, we specified that this manuscript would focus on the model trained in PPMI that presented the highest mean AUC, accuracy, and balanced accuracy in withheld samples before moving on to validation in a *de novo* dataset. We acknowledge that incorporating p-values as a pre-filtering step in the feature selection phase may cause data leakage to some degree. The PPMI training set is only a minuscule portion of the most recent GWAS study (less than 0.1% of the sample size); additionally, the algorithmic feature selection described above is generally much more conservative and excluded the majority of features reaching the p-value thresholds of interest. In this report, we only focused on a model with a 1E-5 maximum p-value for genetic data inclusion and a 1E-2 maximum P for transcriptomic data inclusion; however, all potential models were exported and saved for public use in transfer learning for similar datasets for the scientific community (**Online Repository**).

Post-hoc optimization for class imbalance

After training, we refit the model to the withheld samples using an optimized threshold for case probability based on Youden's J calculation to better account for case-control imbalance and

subsequently increase balanced accuracy and related metrics¹¹. This *post-hoc* optimization was done again after fitting the tuned model to the external validation cohort.

Network communities

After building the ML classifiers, we turned our attention to potentially novel PD gene networks that may be hidden within the classifier's selected features. First, we extracted all identified RNA feature counts for this subset of genes. Next, we subsetting to cases only and calculated the correlation between gene-level transcriptomic data for the nominated genes to build a graph space (minimum correlation coefficient (r), the threshold of 0.8 for connections between gene nodes). Then the Leiden algorithm was implemented to cluster the genes within the larger network into related communities; finally, we calculated a modularity score to evaluate the quality of our network clusters¹².

Looking for potential therapeutic connections across the communities within our defined networks, we utilized webGestaltR; we used its over-representation analysis function to explore druggable target enrichments for network genes within the two available drug databases hosted on the website (DrugBank and GLAD4U)^{13,14,15}. These queries were made under default settings. First, we queried the 300 genes comprising our network communities against a background of all 598 genes nominated at the initial feature selection phase in the transcriptomics data. This estimates how genes comprising our network communities, which are highly correlated in cases, might be enriched compared to genes potentially related to case:control differences. We also looked for enrichments similarly comparing all 598 potential genes delineating cases and controls to > 18,000 protein-coding genes. This was then repeated for our 300 network community genes, investigating over-representation of druggable targets against all protein-coding genes. Our primary goal with this analysis was to see if any drug-related annotations were enriched in our network communities based on correlated gene expression between cases compared to other protein-coding genes that were selected as potential case:control classifying features. This database was accessed on February 23rd, 2021.

ML metrics and model interpretation

Our prioritized metric for evaluating model performance was the area under the curve, AUC. The AUC metric is an aggregate metric summarizing the performance of a classifier across all potential probability thresholds for delineating cases:controls (the labels used in this study) and is less affected by class imbalance than other common metrics. In general, an AUC of greater than 80% may be considered qualitatively to be a robust and "very good to excellent" diagnostic¹⁶. For other metrics such as sensitivity and true positive rate (these relate to the proportion of true positive cases identified) or specificity and true negative rate (these relate to the proportion of true controls identified), their values are easily altered by a change to the probability threshold used to split cases and controls. For example, after the model outputs a probability estimate of a sample being a case, a researcher has the option to use the default probability threshold of 50% for binary classification, or use several methods to optimize this

threshold for better performance. As part of our automated ML workflow, we output performance metrics at default, followed by optimized probability thresholds (using Youden's J). We prioritized secondary performance metrics of interest in addition to AUC; these were accuracy and balanced accuracy, the former being the rate of correct predictions, the latter being the mean accuracy weighted across cases, and controls samples used for dealing with imbalanced datasets¹⁷.

The Shapley additive explanations (SHAP) approach was used to evaluate each feature's influence in the ML model. This approach, used in game theory, assigns an importance (Shapley) value to each feature to determine a player's contribution to success¹⁸. In this analysis, the concept of "success" has been replaced by case status. Shapley explanations enhance understanding by creating accurate explanations for each observation in a dataset. The SHAP package was used to calculate and visualize these Shapley values seen in the figures in the manuscript and the interactive website^{19,20}. A surrogate xgboost model was trained in 70% of the data, and later tested in the 30% of withheld data to evaluate the model's contributing features. The interactive website (<https://share.streamlit.io/anant-dadu/shapleypdpredictiongenetics/main>) was developed as an open-access and cloud-based platform for researchers to investigate the top features of the model developed in this study and how these may influence the classification (or in some cases, misclassification) of a particular sample. In its simplest description, the Shapley values are similar to standard regression derived relative importance measures with regard to interpretation, and all the Shapley values sum to a total value of 1.

Details on data from the Accelerating Medicines Partnership - Parkinson's Disease (AMP-PD) and the Global Parkinson's Genetic Project (GP2)

The summaries below are from AMP PD's website on the different studies used in this manuscript.

Cohort Summaries - PPMI

Study Overview for PPMI

The Parkinson's Progression Markers Initiative (PPMI) is a study sponsored by the Michael J. Fox Foundation. It is a longitudinal, observational study where participants can contribute clinical, demographic, and imaging data alongside biological samples used for whole-genome sequencing, whole blood RNA sequencing, and other assays at 33 clinical sites globally. PPMI follows participants for anywhere from five to 13 years. For this manuscript, we have only focused on data collected at baseline. This data is now hosted as part of AMP PD's version 1 release. For more information on the PPMI study, [please visit this link](https://amp-pd.org/unified-cohorts/ppmi#study-overview) (<https://amp-pd.org/unified-cohorts/ppmi#study-overview>).

Study Inclusion Criteria for PPMI - PD Cases

1. Patients must have at least two of the following: resting tremor, bradykinesia, rigidity (must have either resting tremor or bradykinesia); OR either asymmetric resting tremor or asymmetric bradykinesia
2. A diagnosis of Parkinson disease for 2 years or less at Screening
3. Hoehn and Yahr stage I or II at Baseline
4. Confirmation from imaging core that screening dopamine transporter SPECT scan is consistent with dopamine transporter deficit (or for sites where DaTSCAN™ is not available, that VMAT-2 PET scan is consistent with VMAT deficit)
5. Not expected to require PD medication within at least 6 months from Baseline. Male or female age 30 years or older at time of PD diagnosis

Study Exclusion Criteria for PPMI - PD Cases

1. Currently taking levodopa, dopamine agonists, MAO-B inhibitors, amantadine, or other PD medication
2. Has taken levodopa, dopamine agonists, MAO-B inhibitors, or amantadine within 60 days of Baseline
3. Has taken levodopa or dopamine agonists prior to Baseline for more than a total of 60 days
4. Received any of the following drugs that might interfere with dopamine transporter SPECT imaging: Neuroleptics, metoclopramide, alpha methyl dopa, methylphenidate, reserpine, or amphetamine derivative, within 6 months of Screening
5. Current treatment with anticoagulants (e.g., coumadin, heparin) that might preclude safe completion of the lumbar puncture
6. Condition that precludes the safe performance of routine lumbar puncture, such as prohibitive lumbar spinal disease, bleeding diathesis, or clinically significant coagulopathy or thrombocytopenia
7. Use of investigational drugs or devices within 60 days prior to Baseline (dietary supplements taken outside of a clinical trial are not exclusionary, e.g., coenzyme Q10)

Study Inclusion Criteria for PPMI - Healthy Controls

Healthy controls for the PPMI study included males or females 30 years or older at Screening

Study Exclusion Criteria for PPMI - Healthy Controls

1. Current or active clinically significant neurological disorder (in the opinion of the Investigator).
2. First degree relative with idiopathic PD (parent, sibling, child)
3. MoCA score < 26

4. Received any of the following drugs that might interfere with dopamine transporter SPECT imaging: Neuroleptics, metoclopramide, alpha methyldopa, methylphenidate, reserpine, or amphetamine derivative, within 6 months of Screening
5. Current treatment with anticoagulants (e.g. coumadin, heparin) that might preclude safe completion of the lumbar puncture
6. Condition that precludes the safe performance of routine lumbar puncture, such as prohibitive lumbar spinal disease, bleeding diathesis, or clinically significant coagulopathy or thrombocytopenia
7. Use of investigational drugs or devices within 60 days prior to baseline (dietary supplements taken outside of a clinical trial are not exclusionary, e.g., coenzyme Q10)

Cohort Summaries - PDBP

Study Overview for PDBP

The Parkinson's Disease Biomarkers Program (PDBP) is a study sponsored by the National Institute of Neurological Disorders and Stroke (NINDS). It is a longitudinal, observational study where participants can contribute clinical, demographic, and imaging data alongside biological samples used for whole-genome sequencing, whole blood RNA sequencing, and other assays. The goal of this study is to accelerate the discovery of promising new diagnostic and progression biomarkers for Parkinson's Disease. This data is now hosted as part of AMP PD's version 1 release. For more information on the PDBP study, [please visit this link](https://amp-pd.org/unified-cohorts/pdbp#study-overview) (<https://amp-pd.org/unified-cohorts/pdbp#study-overview>).

Study Inclusion Criteria for PDBP - PD Cases

1. Clinically diagnosed with Parkinson's Disease
2. Male or Female aged 21 years or older at screening
3. Able to cooperate with consent procedures (or has appropriate surrogate as defined and approved per local IRB)
4. Able to participate in study activities including all required clinical assessments and biological donations
5. Participation would not lead to hardship or adverse health or mental health conditions

Study Exclusion Criteria for PDBP - PD Cases

1. Clinical Diagnosis uncertain at the time of enrollment
2. Condition that preclude the safe performance of routine lumbar puncture, such as prohibitive lumbar spinal disease, bleeding diathesis, or coagulopathy or thrombocytopenia
3. Current treatment with anti-coagulants (e.g., Coumadin, heparin) that might preclude safe completion of the lumbar puncture
4. Has a history of neuroleptic use or exposure
5. Has a history of schizophrenia
6. Otherwise unable to participate in biological specimen collection due to a medical condition or medication status (other than items listed above)

7. Otherwise unable to participate in clinical assessments due to a medical condition or medication status (other than items listed above)
8. Unable to participate in consent procedures
9. Use of investigational drugs or devices within 60 days prior to baseline visit (dietary supplements such as Coenzyme Q10, for example, are not exclusionary)

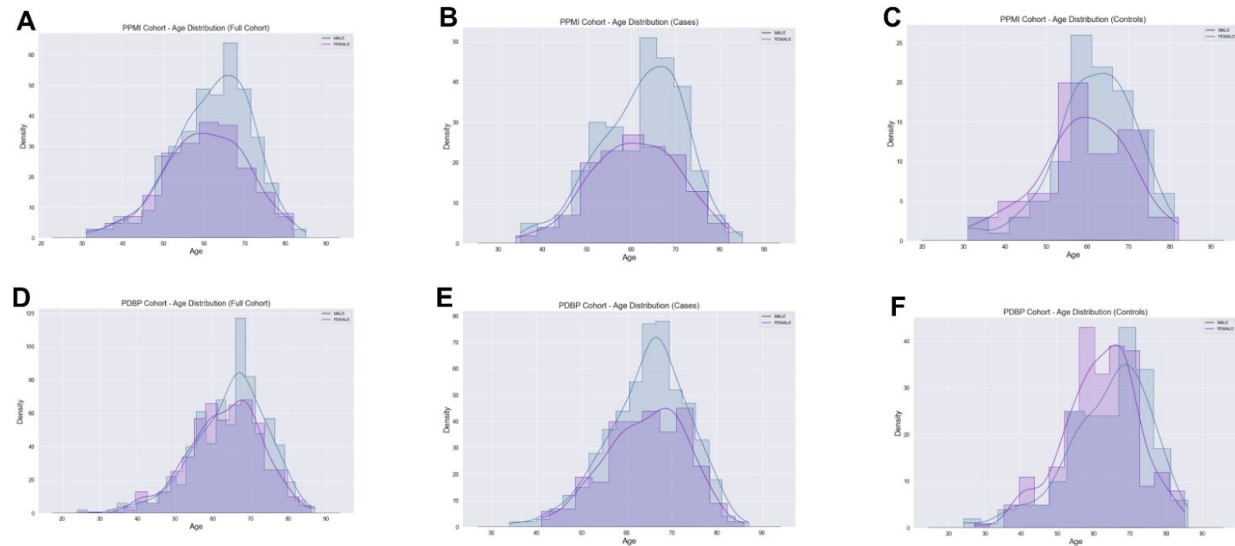
Study Inclusion Criteria for PDBP - Healthy Controls

1. Male or Female aged 21 years or older at screening
2. Able to cooperate with consent procedures (or has appropriate surrogate as defined and approved per local IRB)
3. Able to participate in study activities including all required clinical assessments and biological donations
4. Participation would not lead to hardship or adverse health or mental health conditions

Study Exclusion Criteria for PDBP - Healthy Controls

1. Has a current or clinically significant neurological disorder in the opinion of the investigator
2. Family history of Neurodegenerative disease in a first degree relative or second degree blood relative
3. Condition that preclude the safe performance of routine lumbar puncture, such as prohibitive lumbar spinal disease, bleeding diathesis, or coagulopathy or thrombocytopenia
4. Current treatment with anti-coagulants (e.g., Coumadin, heparin) that might preclude safe completion of the lumbar puncture
5. Has a history of neuroleptic use or exposure
6. Has a history of schizophrenia
7. Otherwise unable to participate in biological specimen collection due to a medical condition or medication status (other than items listed above)
8. Otherwise unable to participate in clinical assessments due to a medical condition or medication status (other than items listed above)
9. Unable to participate in consent procedures
10. Use of investigational drugs or devices within 60 days prior to baseline visit (dietary supplements such as Coenzyme Q10, for example, are not exclusionary)

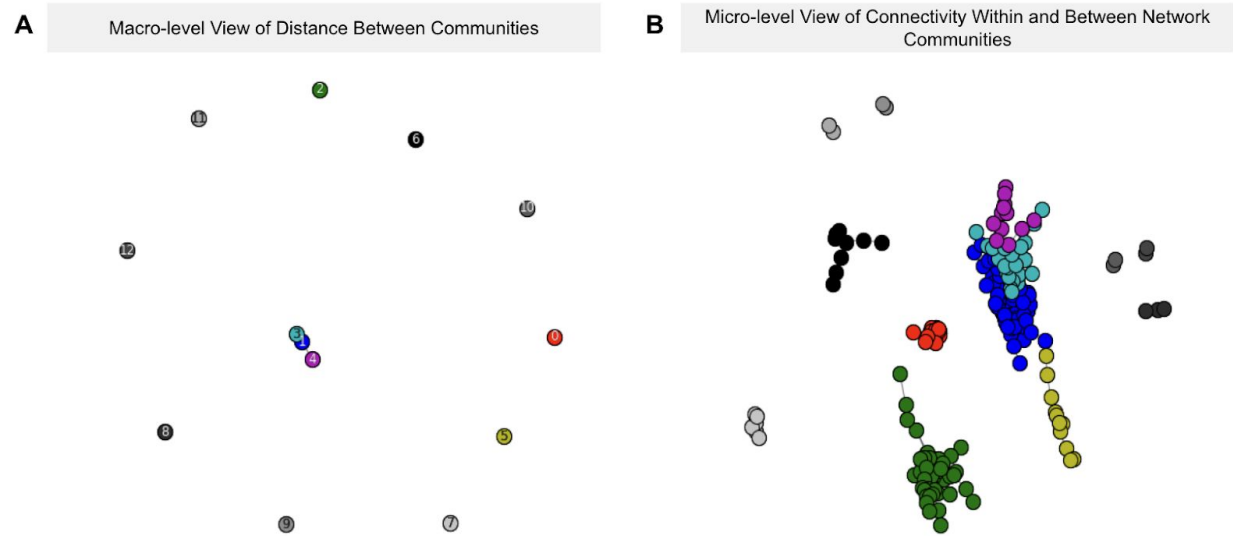
Age distribution per cohort



Supplemental Figure 1: Age distributions of cohorts, broken down by cases and controls
| Blue indicates male, while purple indicates female. Panels A, B, and C show the age distribution of males and females in the full PPMI cohort, cases in PPMI, and controls in PPMI, respectively. Panels D, E, and F show the age distribution of males and females in the full PDBP cohort, cases in PDBP, and controls in PDBP, respectively.

Network Graphical Summary of Nominated Genes

We identified 13 network communities consisting of 300 genes with an Erdos-Renyi modularity score of 0.794 (a modularity score closer to 1 indicates better model fit).

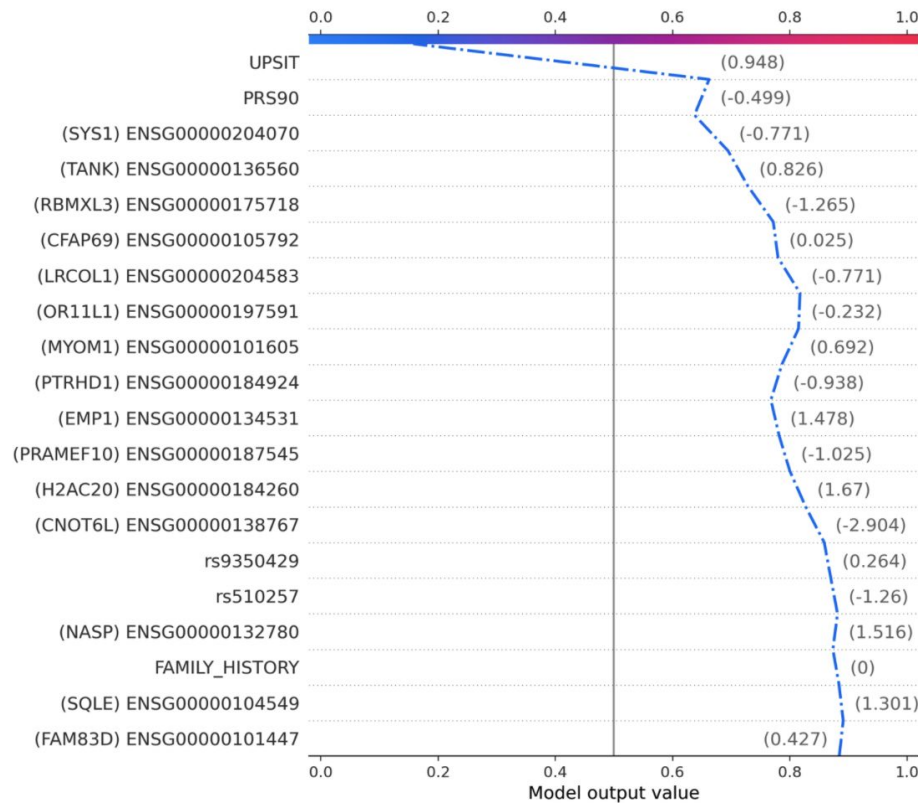


Supplemental Figure 2: Network plot of nominated genes | Panel A provides a macro-level view of the distance between communities (color-coded). Panel B is a micro-level view of connectivity within and between network community modules. The colors of communities in Panel A correspond to those in panel B.

Misclassified cases by the best performing model

Encrypted ID of Misclassified Case: 756ac1345d7068cdc60c8b2583a80092

Decision plots work on visualizing the path a model takes before arriving at a classification. A decision plot shows that a sample that was clinically diagnosed to be a PD case, we see that most of the features seemed to indicate that the individual was about to be classified as a PD case by the model, but ultimately an unexpectedly high UPSIT score misclassified the individual as healthy control.



Supplemental Figure 3: Misclassified case as a healthy control using the best model

Additional References

- 1 Carmi S, Hui KY, Kochav E, *et al.* Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. *Nat Commun* 2014; **5**: 1–9.
- 2 Joseph T, Auger SD, Peress L, *et al.* Screening performance of abbreviated versions of the UPSIT smell test. DOI:10.1101/443127.
- 3 Regier AA, Farjoun Y, Larson D, *et al.* Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. DOI:10.1101/269316.
- 4 Iwaki H, Leonard HL, Makarios MB, *et al.* Accelerating Medicines Partnership: Parkinson's Disease. Genetic Resource. *medRxiv* 2020; : 2020.11.19.20235192.
- 5 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**. DOI:10.1016/S1474-4422(19)30320-5.
- 6 Hutchins E, Craig D, Violich I, *et al.* Quality Control Metrics for Whole Blood Transcriptome Analysis in the Parkinson's Progression Markers Initiative (PPMI). *medRxiv* 2021; : 2021.01.05.21249278.

- 7 Ritchie ME, Phipson B, Wu D, *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015; **43**: e47.
- 8 Abraham G, Qiu Y, Inouye M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* 2017; **33**: 2776–8.
- 9 Extremely randomized trees. *Automation and Control Engineering*. 2010; : 235–8.
- 10 Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 2011; **12**: 2825–30.
- 11 Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biom J* 2008; **50**: 419–30.
- 12 Traag VA, Aldecoa R, Delvenne J-C. Detecting communities using asymptotical surprise. *Phys Rev E Stat Nonlin Soft Matter Phys* 2015; **92**: 022816.
- 13 Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res* 2019; **47**: W199–205.
- 14 Wishart DS, Knox C, Guo AC, *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 2008; **36**: D901–6.
- 15 Jourquin J, Duncan D, Shi Z, Zhang B. GLAD4U: deriving and prioritizing gene lists from PubMed literature. *BMC Genomics* 2012; **13 Suppl 8**. DOI:10.1186/1471-2164-13-S8-S20.
- 16 Šimundić AM. Measures of Diagnostic Accuracy: Basic Definitions. *EJIFCC* 2009; **19**. <https://pubmed.ncbi.nlm.nih.gov/27683318/> (accessed Feb 21, 2021).
- 17 The Balanced Accuracy and Its Posterior Distribution. <https://ieeexplore.ieee.org/document/5597285> (accessed Feb 21, 2021).
- 18 Website. <https://arxiv.org/abs/1705.07874> (accessed Feb 23, 2021).
- 19 Lundberg SM, Nair B, Vavilala MS, *et al.* Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* 2018; **2**: 749–60.
- 20 Lundberg SM, Erion G, Chen H, *et al.* From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell* 2020; **2**: 56–67.

