PARKINSON'S PROGRESSION MARKERS INITIATIVE

Play a Part in Parkinson's Research

Genetic Data Summary: Consensus *APOE* SNPs and *GBA* and *LRRK2* Coding Variant Summary for PPMI Subjects

Kelly N. H. Nudelman, Jessie Browne-Michaels, Trever Jackson, Marco A. Abreu, Dongbing Lai, and Tatiana M. Foroud

Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA

Summary

Genetic data from whole genome sequencing (WGS), whole exome sequencing (WES), GWAS, RNA-sequencing (RNA-seq), Sanger sequencing of select GBA variants, and mutation screening data ('CLIA') was obtained for PPMI participants. This data was supplemented by apolipoprotein E (*APOE)* genotype data generated in-house at the PPMI, LRRK2, ADB, S4, and Bionet Biorepository at Indiana University. Variant data were extracted for the following genes using the same inclusion selection criteria as in the PD GENEration study (https://www.parkinson.org/understanding-parkinsons/causes/genetics/testing-counseling): leucine rich repeat kinase 2 (*LRRK2*); glucosylceramidase beta (*GBA*); VPS35 retromer complex component (*VPS35*); alpha-synuclein (*SNCA*); parkin RBR E3 ubiquitin protein ligase (*PRKN*); Parkinsonism associated deglycase (*PARK7*); and PTEN induced kinase 1 (*PINK1*). Variant data were reviewed to identify those that meet the current American College of Medical Genetics and Genomics (ACMG) criteria for pathogenicity, and data were compared across genetic platforms, to create a consensus variant resource for Parkinson's disease researchers.

Method

Genetic data from PPMI subjects were obtained from chromosome 19 SNPs rs429358 and rs7412 as well as variants from the 7 PD genes also screened in the PD GENEration study (*GBA, LRRK2, VPS35, SNCA, PRKN, PARK7,* and *PINK1*) and compared across genetic platforms to create a consensus reference document for PPMI investigators. Data included: DNA microarray data (GWAS, Project 107); whole exome sequencing (WES, Project 116); whole genome sequencing (WGS, Project 118, hg38 aligned January2021 VCFs); Sanger sequencing of select variants in the *GBA* gene (Sanger, Project 126); RNA-sequencing (Project 133); and genetic screening data ('CLIA', document 'Genetic Testing Results – Screening', included under Biospecimen data), as well as updated WGS (2022 v3 release) from the Accelerating Medicines Partnership Parkinson's disease (AMP-PD; https://amp-pd.org) program, and gVCFs for 281 PPMI participants not included in previous batches or re-runs to replace quality control failed results. RNA-sequencing data was obtained directly from the investigators who generated the data; the other data were downloaded from the Laboratory of Neuro Imaging Image and Data Archive (LONI; https://ida.loni.usc.edu) or AMP-PD (https://amp-pd.org). For *APOE*, data from

LONI was supplemented by data generated in-house at the PPMI, LRRK2, ADB, S4, and Bionet Biorepository at Indiana University with a custom 96-SNP microarray using Fluidigm microfluidic technology (Standard BioTools, South San Francisco, CA). All data was quality controlled for call rate, heterozygosity ratio, genetic vs. reported sex, and specimen genetic concordance to other subject specimen data. For the gVCF files obtained from the Singleton laboratory for new WGS, data were joint-called using GATK best practices (https://gatk.broadinstitute.org) with GenotypeGVCFs. Data generated in-house at the Indiana University biorepository for SNPs rs429358 and rs7412 were tested twice, and only replicated results concordant across all available tested samples per participant were used for the analysis, following an established Standard Operating Procedure (SOP). Finally, structural variant (SV) calling results were obtained from the Singleton laboratory for a subset of individuals with whole genome sequencing. All data used in this analysis are currently available for download on LONI or AMP-PD under specified project IDs and locations or by request from the investigator, as well as methods documents describing in detail how each data set was generated.

For the WGS data, variants within each gene were first extracted using VCFtools[1] from the AMP-PD data set. Some individuals in PPMI did not have WGS included in the AMP-PD data release, but had data available through LONI, or new data generated by the Singleton laboratory. For these individuals, data quality control results were reviewed and notes included where relevant in the consensus document in the 'QC_NOTES' column. There were some individuals who were excluded from the AMP-PD data set due to missing or discrepant clinical data, while others were removed due to other quality control issues such as sample duplication (data includes WGS from identical twins). All relevant issues have been noted in the QC_NOTES column, which we strongly recommend be reviewed prior to analysis. Depending on the goal of the analysis, it might be appropriate to remove some or all of these individuals.

For all variants in the *LRRK2, GBA*, and *SNCA* genes, variant data was compiled and compared for differences between assays. PLINK was used to extract variants from GWAS data sets.[2,3] For variant calls that differed between genetic assays, originating data was reviewed where needed and possible (for example, for whole exome and genome sequencing, and RNA-sequencing, read depth and allele counts were reviewed and data was viewed using Integrative Genomics Viewer (IGV; https://software.broadinstitute.org/software/igv/)[4-6]), and manual curation was completed to create a consensus variant call for the subject. RNA-seq was not used as a primary source of variant data, due to concerns about shallow read depth in some samples and variants as well as concern about potential transcribed allele bias. However, RNA-seq was used as a secondary data source. For the *VPS35, PRKN, PARK7,* and *PINK1* genes, variants were included from the WGS data.

For this version of the consensus document, data was also supplemented with information obtained from the work of Dr. Christos Proukakis. His laboratory, in collaboration with Illumina investigators, produced corrected variant calls for the *GBA* gene[7]. In brief, the investigators developed an improved WGS variant data analysis program, Gauchian, that is able to more accurately call variants in the *GBA* gene region. The investigators validated these findings using

long-read sequencing to confirm accuracy. While the previous version of the consensus document more accurately matched these results compared to variant calls from WGS alone, there were a small number of GBA variant calls that were updated based on these additional data.

For *APOE* genotype analysis, data for rs429358 and rs7412 were obtained from whole genome sequencing data as well as data previously uploaded to LONI (Current Biospecimen Results) and in-house genotype data generated by the Indiana University biobank for samples stored at the biobank. Data from each source was compared and compiled to generate *APOE* genotype calls. For *APOE*, rs429358 and rs7412 were used to calculate alleles (E2/E3/E4); results were compared across platforms, and consensus *APOE* alleles are shown in column 'APOE'. Participants/genes with no data available, or data discrepancies which could not be resolved, are marked 'NA'.

SV calls obtained from Kimberley Billingsley and colleagues for the genes included in this document were assessed for pathogenicity similar to the process used to assess variants (see below). The method used to produce these data is described in more detail in Billingsley et al., 2023[8]. SVs that are likely pathogenic or reported pathogenic are included in the results. There are several variants which are currently undergoing further validation; in this version of the consensus document, results for the relevant gene are listed as 'NA'.

Columns 'CLIA', 'GWAS', 'WES', 'WGS', "SVs", 'Sanger', and 'RNASEQ' document whether data of that type is available for each subject (X=available; '-' = NA; see Table 1). Some variants are only called in WGS and RNA-seq data, so data for these subjects is generally considered incomplete if WGS is missing. For analyses comprising carrier status across all seven genes, it may be best to exclude these subjects, as their carrier status cannot be fully determined from existing data. The summary data includes a 'NOTES' column which includes an indicator 'WGS missing' to allow convenient filtering of these data.

For participants with only CLIA data available, there are many participants who were screened for one or two variants individually, but who do not have any additional data available. For individuals who were negative via CLIA screening, they are listed as '0' for the gene(s) with the screened variant(s); however, if performing an analysis including multiple variants across *LRRK2* or *GBA*, it is important to note that these individuals do not have a complete screen, and thus may not be negative for all pathogenic variants.

All variant data for each gene was annotated using Annovar[9], then manually curated, combined, and reviewed. ClinVar[10], Franklin (https://franklin.genoox.com), and Varsome[11] were also used to evaluate variants of interest for reported or likely pathogenicity. Variants that meet ACMG criteria for 'pathogenic' or 'likely pathogenic' are included. A full list of identified variants and annotation data is available in Table 2.

PARKINSON'S PROGRESSION MARKERS INITIATIVE

Play a Part in Parkinson's Research

Table 1. Variables included in the data set and their descriptions

Variable	Description
PATNO	Patient ID
CLIA	Gene Screening Data ('X' = data, '-' = NA)
GWAS	GWAS data ('X' = data, '-' = NA)
WES	Whole exome sequencing data ('X' = data, '-' = NA)
WGS	Whole genome sequencing data ('X' = data, '-' = NA)
SVs	Structural variant calls from whole genome sequencing ('X' = data, '-' = NA
SANGER	Sanger sequencing data ('X' = data, '-' = NA)
RNASEQ	RNA-sequencing data ('X' = data, '-' = NA)
RNASEQ_VIS	Count of RNA-sequencing visits with data
APOE	*APOE* alleles (E2/E3/E4)
PATHVAR_COUNT	Number of pathogenic variants identified across all seven genes for the participant
VAR_GENE	Gene(s) with pathogenic variant(s)
LRRK2	Pathogenic variants identified in the *LRRK2* gene (0=none); variants identified by amino acid change (N###N). Variants listed twice separated with a '/' indicate homozygous carriers (i..e. G2019S/G2019S).
GBA	Pathogenic variants identified in the *GBA* gene (0=none); variants identified by amino acid change (N###N), transcript location (c.###N>N), or common alias (i.e. IVS2+1G>A). Variants listed twice separated with a '/' indicate homozygous carriers (i..e. N409S/N409S).
VPS35	Pathogenic variants identified in the *VPS35* gene (0=none); variants identified by amino acid change (N###N).
SNCA	Pathogenic variants identified in the *SNCA* gene (0=none); variants identified by amino acid change (N###N).
PRKN	Pathogenic variants identified in the *PRKN* gene (0=none); variants identified by amino acid change (N###N), transcript location (c.###N>N), or SV information (i.e. <DEL>chr6:162316450-162488715). For individuals with more than one pathogenic variant, variants are separated by '/' (i.e. R275W / P113Tfs*51).
PARK7	Pathogenic variants identified in the *PARK7* gene (0=none); variants identified by amino acid change (N###N).
PINK1	Pathogenic variants identified in the *PINK1* gene (0=none); variants identified by amino acid change (N###N).
NOTES	Recommendations for exclusion from analyses based on current data and notes on missing data, and other notes about consensus curation

PARKINSON'S PROGRESSION MARKERS INITIATIVE

Play a Part in Parkinson's Research

Table 2. Included Variant Annotations

Chr	bp_hg38	dbSNP ID	Gene	Ref/Alt Alleles	Variant Annotation	Alias*
12	40309225	rs74163686	LRRK2	A/C	p.N1437H	
12	40310434	rs33939927	LRRK2	C/T	p.R1441C	
12	40310434	rs33939927	LRRK2	C/G	p.R1441G	
12	40310435	rs34995376	LRRK2	G/A	p.R1441H	
12	40340400	rs34637584	LRRK2	G/A	p.G2019S	
12	40340404	rs35870237	LRRK2	T/C	p.I2020T	
1	155235002	rs80356773	GBA	C/T	p.R535H	p.R496H
1	155235196	rs80356771	GBA	G/A	p.R502C	p.R463C
1	155235252	rs421016	GBA	A/G	p.L483P	p.L444P
1	155235727	rs1064651	GBA	C/G	p.D448H	p.D409H
1	155235843	rs76763715	GBA	T/C	p.N409S	p.N370S
1	155238260		GBA	G/C	p.S212*	p.S173*
1	155238630	rs439898	GBA	G/A	p.R159W	p.R120W
1	155240629	rs104886460	GBA	G/A	c.115+1G>A	IVS2+1G>A
1	155240660	rs387906315	GBA	G/GC	p.L29Afs*18	p.84GG
16	46669006		VPS35	C/T	R524Q	
4	89828149	rs104893877	SNCA	C/T	A53T	
6	161350101		PRKN	A/G	X466Q	
6	161350208		PRKN	C/T	G430D	
6	161567720		PRKN	N/<DUP>	<DUP>chr6:161567720-162085660	
6	161785820	rs34424986	PRKN	G/A	R275W	
6	162002000		PRKN	N/<DEL>	<DEL>chr6:162002000-162066790	
6	162002000		PRKN	N/<DUP>	<DUP>chr6:162002000-162066790	
6	162256395		PRKN	N/<DEL>	<DEL>chr6:162256395-162397000	
6	162262537		PRKN	C/CTGG	P133_A134insP	
6	162262560		PRKN	TCAGTGTGCAGAATGACAGCCAGCCCCACAGAGTCTCCTGG/T	P113Tfs*51	
6	162316450		PRKN	N/<DEL>	<DEL>chr6:162316450-162488715	
6	162423548		PRKN	N/<DUP>	<DUP>chr6:162423548-162571629	
6	162443356	rs368134308	PRKN	C/G	R42P	

6	162443371	rs148990138	PRKN	G/A	P37L	
6	162443378	rs55777503	PRKN	CCT/C	Q34Rfs*5	
6	162443383	rs147757966	PRKN	C/T	R33Q	
6	162443408		PRKN	G/A	Q25X	
1	7977708		PARK7	C/T	P127S	
1	20633855		PINK1	GC/G	F104Sfs*3	
1	20649217	rs34208370	PINK1	C/T	R492X	

Chr = chromosome; bp_hg38 = base pair location of variant in human genome build 38; Variant Annotation indicates amino acid change or transcript-based location for splicing variants
*Common aliases for annotation scheme still used frequently by investigators

References

1. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*. 2011 Aug 1;27(15):2156-8. doi: 10.1093/bioinformatics/btr330. Epub 2011 Jun 7. PMID: 21653522; PMCID: PMC3137218.
2. Purcell, S., and Chang, C. PLINK v1.9. www.cog-genomics.org/plink/1.9/
3. 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 Feb 25;4:7. doi: 10.1186/s13742-015-0047-8. PMID: 25722852; PMCID: PMC4342193.
4. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol*. 2011 Jan;29(1):24-6. doi: 10.1038/nbt.1754. PMID: 21221095; PMCID: PMC3346182.
5. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. 2013 Mar;14(2):178-92. doi: 10.1093/bib/bbs017. Epub 2012 Apr 19. PMID: 22517427; PMCID: PMC3603213.
6. Robinson JT, Thorvaldsdóttir H, Wenger AM, Zehir A, Mesirov JP. Variant Review with the Integrative Genomics Viewer. *Cancer Res*. 2017 Nov 1;77(21):e31-e34. doi: 10.1158/0008-5472.CAN-17-0337. PMID: 29092934; PMCID: PMC5678989.
7. Toffoli M, Chen X, Sedlazeck FJ, et al. Comprehensive short and long read sequencing analysis for the Gaucher and Parkinson's disease-associated GBA gene. *Commun Biol*. 2022;5(1):670. Published 2022 Jul 6. doi:10.1038/s42003-022-03610-7
8. Billingsley KJ, Ding J, Jerez PA, et al. Genome-Wide Analysis of Structural Variants in Parkinson Disease. *Ann Neurol*. 2023;93(5):1012-1022. doi:10.1002/ana.26608
9. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data. *Nucleic Acids Research*, 38:e164, 2010.
10. Landrum, M. J., Chitipiralla, S., Brown, G. R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., Lyoshin, V., Maddipatla, Z., Maiti, R., Mitchell, J., O'Leary, N., Riley,

G. R., Shi, W., Zhou, G., Schneider, V., Maglott, D., Holmes, J.B., Kattman, B. L. ClinVar: improvements to accessing data. *Nucleic Acids Res*. 2020;48(D1):D835-D844. doi: 10.1093/nar/gkz972. PMID:31777943.

11. VarSome: the human genomic variant search engine, Christos Kopanos, Vasilis Tsiolkas, Alexandros Kouris, Charles E Chapple, Monica Albarca Aguilera, Richard Meyer, Andreas Massouras. Bioinformatics, Volume 35, Issue 11, 1 June 2019, Pages 1978–1980, https://doi.org/10.1093/bioinformatics/bty897.

About the Author

This document was prepared by **Kelly Nudelman, Indiana University School of Medicine, Department of Medical and Molecular Genetics**. For more information, please contact Kelly Nudelman by email at **kholohan@iu.edu**.