Genomics of Parkinson's Disease: Global and Scalable Approaches Towards Precision Medicine

Mary Botros Makarious

University College London
Queen Square Institute of Neurology
Department of Clinical and Movement Neurosciences

For the Award of Doctor of Philosophy

Declaration

I, Mary Botros Makarious, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Acknowledgements

Reflecting on this journey, I am profoundly moved by the incredible number of individuals I have the opportunity to thank for this experience. This experience has underscored a vital truth: success is not a solitary endeavor but rather the result of numerous contributions, both significant and subtle, from a community that has faith in your potential.

First of all, a heartfelt thanks to the patients and cohorts whose participation has been essential. You are truly the heart of this research.

I extend my deep appreciation to my mentors and supervisors at both the National Institutes of Health (NIH)—Drs. Andrew Singleton, Mike Nalls, and Cornelis Blauwendraat—and University College London (UCL)—Drs. Huw Morris and John Hardy. Your guidance opened doors to growth and learning, fostering a transformative culture of collaboration and fun.

My time at Genentech, particularly in the Human Genetics department under Drs. Bryce van de Geijn and Tushar Bhangale, was an amazing opportunity that significantly deepened my understanding of research.

Big thanks to my colleagues and friends from UCL, especially Dr. Mina Ryten, Dr. Raquel Real, Alejandro Martinez Carrasco, and Lesley Wu. Your support and teamwork made all the difference. A special shoutout to my colleagues and friends from the Global Parkinson's Genetics Program—notably Dr. Bradford Casey, Dr. Zih-Hua Fang, Dr. Thiago Peixoto Leal, Sumit Dey, Dr. Teresa Periñán Tocino, Dr. Kajsa Brolin, Dr. Paula Saffie Awad, Dr. Artur Schumacher Shuh, Dr. Lara Lange, Dr. Niccolo Mencacci, Dr. Nacho Mata, Dr. Njideka Okubadejo, Dr. Henry Houlden, and Dr. Alastair Noyce—for their collaborative spirit and friendship.

The NIH's Center for Alzheimer's and Related Dementias and the Laboratory of Neurogenetics have been vibrant hubs of support. Special thanks to Peter Wild Crea, Mat Koretsky, Nicole Kuznetsov, Samantha Hong, Dr. Kim Billingsley, Pilar Alvarez Jerez, Dr. Emma Price, Dr. Sara Bandrés Ciga, and Jeff Kim. A massive shoutout to the team at Data Tecnica—especially Hampton Leonard, Kristin Levine, Chelsea Alvarado, Dr. Cory Weller, and Dan Vitale—for your guidance and insights.

I would also like to thank my mentors—Drs. Mark V. Albert, Jessica Brann, Jennifer Mierisch, Heather Wheeler, and Catherine Putonti—at my undergraduate alma mater, Loyola University in Chicago, for encouraging me to pursue graduate school and seriously consider a career as a scientist.

I owe a tremendous amount of gratitude to more of my friends—Somin, Estefany, Nick, Margaret, John, Theresa, Alex, Christina, and Shilpa. Your support, encouragement, and laughter have been the light during challenging times, and your faith in me has propelled me forward.

To my family, particularly my brothers George and Mina, and my parents, Mona and Botros, as well as our friends and extended families in Cairo and Chicago—your sacrifices and belief in me have been nothing short of transformative. The choices you made, rooted in love and hope, have provided me with strength and courage to pursue my PhD. To David's family, your warm embrace and steadfast support have offered me a second home filled with love and encouragement.

And finally, to my fiancé, Dr. David Saffo, without your endless support, love, and belief in me, this journey would not have been possible. You have been my loudest cheerleader and my fiercest advocate.

To everyone I mentioned here—this success is ours to share, as it was built on the foundation of your unwavering support and belief in me. For that, I am forever grateful.

Impact Statement

Throughout my research in Parkinson's disease (PD) genomics, I have assessed genetic variations and their association with the disease. With a global approach, I have used both common and rare genetic variants to gain valuable insights into the genetic underpinnings of PD. Importantly, my work has highlighted the significance of studying diverse populations, revealing novel genetic risk factors specific to African and African admixed populations.

In a pivotal discovery, our research identified a novel common risk factor at the *GBA1* locus, uniquely prevalent in African ancestry populations, highlighting a distinct genetic architecture and pointing towards a founder effect. This work underscores the importance of including ancestrally diverse groups in genetic studies, revealing the *GBA1* rs3115534-G allele's rarity outside of these populations and its significant contribution to the population-attributable risk for PD, despite its lower frequency compared to known coding variants in other populations. These findings not only bridge the diversity gap in PD research but also pave the way for RNA-based therapeutic strategies and more inclusive clinical trial designs, emphasizing the need for personalized, efficient treatments. As my work continues, I will delve deeper into the realm of PD-associated biological networks, focusing on community clustering and network analyses. I am optimistic that by understanding networks associated with genes such as *GBA1*, we will unlock the potential for more targeted therapies and move a step closer to personalized treatment strategies for PD patients.

One of the primary challenges in PD research has been early detection due to its heterogeneous nature. My advancements in integrating multi-omics and machine learning with the development of GenoML enable a more precise assessment of PD risk, integrating multiple data types for enhanced accuracy and a biobank-scale model that can identify 10 individuals for every real case to follow-up with. GenoML has 16,000 downloads at time of writing, and is rooted in an open source framework. My hope is that its accessibility has ensured researchers across the globe can replicate and benefit from its capabilities on their own data.

All the findings from my research have been peer-reviewed and published. I firmly believe in an open science approach, and to that end, my code and analysis pipelines are publicly available on GitHub for the wider scientific community to not only use, but hopefully learn, expand, and improve upon as well.

Outside of the lab, I have been fortunate enough to be able to actively learn from and collaborate with researchers and physicians also deeply dedicated to better understanding Parkinson’s disease through the International Parkinson’s Disease Genomics Consortium (IPDGC), Accelerating Medicines Partnership in Parkinson’s disease (AMP-PD), and the Global Parkinson’s Genetics Program (GP2). I have also been able to present my research at the NIH Research Festival, the International Congress of Parkinson’s Disease and Movement Disorders (MDS), the American Society for Human Genetics (ASHG), and the

International Conference on Alzheimer’s and Parkinson’s Diseases and related neurological disorders (AD/PD).

My intention has always been to translate the complexities of my research into actionable insights and interventions for those affected by PD. Through the spectrum of my research, from genomics to machine learning, I am committed to unraveling the mysteries of Parkinson's disease and other neurodegenerative diseases.

As I look ahead, my goal remains unchanged: to increase diversity in our studies, develop tools and pipelines to facilitate early diagnosis, pave the way for tailored therapeutic interventions, and contribute substantially to the global PD research community.

Abbreviations

AAC	African Admixed
AAO	Age at Onset
ACAT	Aggregated Cauchy association test
ACB	African Caribbean in Barbados
AD	Allelic depth
AF	Allele frequency
AFR	African
AI	Artificial intelligence
AJ	Ashkenazi Jewish
ALS	Amyotrophic lateral sclerosis
AMD	Age-related macular degeneration
AMP-PD	Accelerating Medicines Partnership in Parkinson’s disease
AMR	Amerindian
ANN	Artificial neural network
ARJP	Autosomal juvenile recessive parkinsonism
ASHG	American Society for Human Genetics
ASW	African ancestry in Southwest United States of America
AUC	Area Under the Curve
BAF	B-Allele frequency
BLAAC	Black and African American Connections to Parkinson's Disease
BWA	Burrows-Wheeler Aligner
CMC	Combined Multivariate and Collapsing
CNV	Copy number variant
CSF	Cerebrospinal fluid
DLB	Dementia with Lewy bodies
DNA	Deoxyribonucleic acid
EAS	East Asian
EHRs	Electronic health records
eQTL	Expression quantitative trait locus
ESN	Esan in Nigeria

EUR	European
FDR	false discovery rate
FOR	false omission rate
FTD	frontotemporal dementia
GABA	gamma-aminobutyric acid
GAS	Genetic Association Study
GATK	Genome Analysis Toolkit
GCase	Glucocerebrosidase
GD	Gaucher disease
gDNA	Genomic DNA
GERP	Genomic Evolutionary Rate Profiling score
GNE	Genentech
GO	Gene ontology
GP2	Global Parkinson's Genetics Program
GQ	Genotype quality
GRS	Genetic risk scores
GSA	Global screening Array (from Illumina)
GWAS	Genome-wide association study
GWD	Gambian in Western Division
HBS	Harvard Biomarker Study
HGMD	Human Gene Mutation Database
HLA	Human leukocyte antigen
HMM	Hidden Markov model
HRC	Human Reference Consortium
HWE	Hardy-Weinberg Equilibrium
IBD	Identity-by-descent
IBS	Irritable bowel syndrome
ICD-10	ncluded data derived from the International Classification of Diseases, 10th Revisior
IGV	Integrative Genomics Viewer
IPDGC	International Parkinson's Disease Genomics Consortium
IPDGCAF-NG	International Parkinson's Disease Genomics Consortium - Nigeria
IPDGCAN	International Parkinson's Disease Genomics Consortium - Africa

IPF	Idiopathic pulmonary fibrosis
IRB	Institutional review board
KNN	K-Nearest Neighbors
L2R	Log2 Ratio
LBD	Lewy Body Dementia
LCC	LRRK2 Cohort Consortium
LCLs	Lymphoblastoid cell lines
LD	Linkage disequilibrium
LDA	Linear Discriminant Analysis
LDSC	LD Score Regression
LNG	Laboratory of Neurogenetics
LoF	Loss-of-function
LOFTEE	Loss-Of-Function Transcript Effect Estimator
LOWESS	Locally Weighted Scatterplot Smoothing
LWK	Luhya in Webuye, Kenya
MAC	Minimum allele count
MAF	Minor allele frequeny
MDS	Movement Disorders
ML	Machine learning
MLP	Multi-layer Perceptron
MLPA	Multiplex ligation-dependent probe amplification
MSA	Multiple system atrophy
MSL	Mende in Sierra Leone
NCBI	National Center for Biotechnology Information
NDDs	Neurodegenerative disorders/diseases
NIH	National Institutes of Health
NINDS	National Institute of Neurological Disorders and Stroke
NPV	Negative Predictive Value
ONT	Oxford Nanopore Technologies
OR	Odds ratio
PAF	Pure autonomic failure
PAR	Population attributable risk

PCA	Principal component analysis
PCs	Principal components
PD	Parkinson's disease
PDBP	Parkinson's disease biomarkers program study
PP	Posterior probabilities
PPMI	Parkinson’s Progression Markers Initiative
PPV	Positive Predictive Value
pQTL	Protein quantitative trait locus
PRS	Polygenic risk score
PSP	Progressive supranuclear palsy
QC	Quality control
QDA	Quadratic Discriminant Analysis
QTL	Quantitative trait locus
RBD	REM sleep behavior disorder
REM	rapid eye movement
RNA	Ribonucleic acid
ROC	Receiver Operating Characteristic
ROHs	Runs of homozygosity
SD	Standard deviation
SE	Standard error
SGD	Stochastic Gradient Descent
SKAT	Sequence Kernel Association Test
SKAT-O	Sequence Kernel Association Test - Optimized
SNP	Single nucleotide polymorphism
SNVs	single nucleotide variations
SURE-PD3	Study of URate Elevation in Parkinson’s Disease, phase 3” trial
SVC	Support Vector Machine Classification
SVs	Structural variants
SWEDD	Scans without evidence for dopaminergic deficit
UKB	United Kingdom Biobank
UKBEC	United Kingdom Brain Expression Consortium
UMAP	Uniform Manifold Approximation and Projection

UPSIT	University of Pennsylvania Smell Inventory Test
USHUS	Uniformed Services University
VCF	Variant call format
VEP	Variant Effect Predictor
VIF	Variance inflation factor
VQSR	Variant Quality Score Recalibration
WES	Whole exome sequencing
WGS	Whole genome sequencing
XGBoost	eXtreme Gradient Boosting
YRI	Yoruba in Ibadan

List of Tables

Supplementary Tables

These are not printed in-line, but can be downloaded from the accompanying files with this thesis here: https://github.com/m-makarious/ucl_phd_thesis

- **Supplementary Table 2:** Overview of *SNCA* copy number variant carriers
- **Supplementary Table 3:** Frequency of variants for each gene in cases and controls, stratified by variant class and cohort (MAF <0.1%; excluding Genentech)
- **Supplementary Table 4:** Frequency of variants for each gene in cases and controls, stratified by variant class and cohort (MAF <1%; excluding Genentech)
- **Supplementary Table 5:** N variants and positions for each gene in cases and controls, stratified by variant class and cohort (MAF <0.1%; excluding Genentech)
- **Supplementary Table 6:** N variants and positions for each gene in cases and controls, stratified by variant class and cohort (MAF <1%; excluding Genentech)
- **Supplementary Table 7:** Variant IDs for genes reaching exome-wide significance (p<1E-6; MAF <0.1%; excluding Genentech)
- **Supplementary Table 8:** Variant IDs for genes reaching exome-wide significance (p<1E-6; MAF <1%; excluding Genentech)
- **Supplementary Table 9:** Lambda and lambda 1000 values per dataset
- **Supplementary Table 10:** Lambda and lambda 1000 values per meta-analysis
- **Supplementary Table 11:** Exome-wide significant genes (p<1E-6) and variant count information in meta-analyses and separate cohorts
- **Supplementary Table 12:** *LRRK2* conditional analysis (excluding Genentech)
- **Supplementary Table 13:** Genes previously reported in the literature, across both meta-analyses, if exome-wide or nominally significant
- **Supplementary Table 14:** Possible compound heterozygous/recessive counts per gene for high confidence LoF or CADD>20 variant carriers (excluding Genentech)
- **Supplementary Table 15:** Genes identified at each Parkinson's disease GWAS locus within -/+ 1Mb
- **Supplementary Table 16:** The 23andMe Reference Panel Information
- **Supplementary Table 17:** Meta-GWAS Fine-mapping Analyses
- **Supplementary Table 18:** Genome-wide Replication Assessment of Known PD Risk Loci
- **Supplementary Table 19:** Genome-wide significant SNPs identified in the African and African Admixed GWAS meta-analysis with frequency metrics
- **Supplementary Table 20:** Complete performance metrics for best combined method comparing training in withheld samples in PPMI
- **Supplementary Table 21:** SHAP values for final combined multi-omic model
- **Supplementary Table 22:** Complete QTL analysis between all nominated SNPs and nominated transcripts in top performing model
- **Supplementary Table 23:** Complete summary statistics for QTL Mendelian randomization

List of Figures

- **Figure 26:** Population attributable risk comparison for *GBA1* known coding variants in the EUR population versus the novel *GBA1* intronic variant in the AFR population
- **Figure 27:** GenoML Workflow
- **Figure 28:** PPMI and PDBP Age Distribution
- **Figure 29:** Workflow and data summary for multi-omics of PD risk prediction using ML study
- **Figure 30:** Correlation matrix of top 5% contributing features in ML model
- **Figure 31:** 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
- **Figure 32:** 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
- **Figure 33:** Feature importance plots for top 5% of features in XGBoost surrogate of best combined *omics model
- **Figure 34:** Network plot of nominated genes following best combined multi-model ML prediction model
- **Figure 35:** Misclassified case as a healthy control using the best multi-modal model

Supplementary Figures

These are not printed in-line, but can be downloaded from the accompanying files with this thesis here: https://github.com/m-makarious/ucl_phd_thesis

- **Supplementary Figure 1:** Complex *SNCA* genomic events of interest
- **Supplementary Figure 2:** Ten random "negative control" subjects for the UK biobank exome sequencing replication of *SNCA* alteration carriers
- **Supplementary Figure 3:** Genentech IGV Plot (*B3GNT3* LoF; chr19:17807816:T:G)
- **Supplementary Figure 4:** Genentech IGV Plot (*B3GNT3* LoF; chr19:17808270:G:T)
- **Supplementary Figure 5:** Genentech IGV Plot (*B3GNT3* LoF; chr19:17808270:G:T)
- **Supplementary Figure 6:** Genentech IGV Plot (*B3GNT3* LoF; chr19:17807919:G:T)
- **Supplementary Figure 7:** UK Biobank IGV Plot for Control (*B3GNT3* LoF; chr19:17807982:GC:G)
- **Supplementary Figure 8:** UK Biobank IGV Plot for PD Parent Proxy #1 (*B3GNT3* LoF; chr19:17808033:C:T)
- **Supplementary Figure 9:** UK Biobank IGV Plot for PD Parent Proxy #2 (*B3GNT3* LoF; chr19:17812105:C:CA)
- **Supplementary Figure 10:** *GBAP1* Duplication and MTX1, MTX1P1 Fusion vs Long Read Sequencing

Table of Contents

Chapter 1: Historical Overview and Contemporary Challenges in Parkinson's Disease

Overview

Parkinson's disease (PD) is a complex progressive neurodegenerative disease with multifactorial etiology. Both rare and common genetic variants contribute to the disease's risk, onset, severity, and progression. There are no treatments that cure PD, but there are multiple treatments designed to help patients manage their symptoms. For progressive neurodegenerative diseases such as PD, early and accurate diagnosis is likely to be key in effectively developing and using new interventions, while the disease process is most amenable to therapeutic intervention. This work aims to leverage large multi-omic datasets to identify patients early in their disease course and design personalized risk and progression predictions. In addition to investigating multiple *omics and common genetic variations associated with PD, I will investigate the contribution of rare variants to PD risk. The outcome of this work aims to improve early diagnosis and identify potential therapeutic targets for new disease-modifying treatments.

Background: Parkinson's Disease

PD was initially described by British physician Dr. James Parkinson in 1817, in *An Essay on the Shaking Palsy*. He described multiple men between the ages of 50 and 65 with the following characteristic motor symptoms of PD: tremor of limbs even at rest, impaired and stooped posture, and slow, erratic movement (bradykinesia) coupled with the weakness of muscles (Parkinson 1817; Lees 2017). Now, the diagnostic criteria for PD are characterized by a composite of motor symptoms termed "parkinsonism" or Parkinson's syndrome, together with supportive features consistent with an underlying pathological diagnosis of Lewy body PD. The Queen Square Brain Bank criteria focused on motor symptoms, progression and the response to treatment and the more recent Movement Disorder Society (MDS) diagnostic criteria for PD include supportive non-motor features. Parkinsonism is an umbrella term to describe the motor symptoms typically found in PD, such as bradykinesia, rigidity, tremor, and postural instability. Typically the prodromal phase of PD has characteristics such as anosmia, sleep disturbances, and constipation, sometimes up to 20 years prior to the development of overt symptoms (Savica et al. 2009; Gelber, Launer, and White 2012).

Synucleinopathies

Synucleinopathies are a group of neurodegenerative diseases characterized by abnormal intracellular aggregations of the alpha-synuclein protein encoded by *SNCA*. They include PD, dementia with Lewy bodies (DLB), pure autonomic failure (PAF), and multiple system atrophy (MSA). The role of alpha-synuclein was established following the identification of *SNCA* as a gene that contains mutations linked to the autosomal dominant form of PD, and of alpha-synuclein as a core component of Lewy bodies (Fayyad et al. 2019; Goedert et al. 2013). The pathological characteristics of PD are Lewy bodies in

multiple brain regions including the midbrain and a loss of dopaminergic neurons in the substantia nigra pars compacta. Other synucleinopathies include MSA and DLB, suggesting there might be both clinical and genetic overlap between PD and these diseases (Blauwendraat, Nalls, and Singleton 2020).

Role of Genetics in Parkinson's Disease

Traditionally perceived as having limited genetic involvement, PD now stands recognized as a condition underpinned by a spectrum of genetic contributions. The genetic underpinnings of PD have been routinely viewed through a dichotomous lens: monogenic determinants or common small-effect variants. However, the more we delve into the genetics of PD, the more we recognize the fluidity between these categories. This spectrum of genetic influence in PD is becoming increasingly nuanced, challenging the conventional boundaries of classification (**Figure 1**).

On one end, we have monogenic forms of PD, characterized by a mutation or several mutations in a single gene with dominant or recessive inheritance. These forms, while rare, are pivotal as they have high, although age-dependent penetrance. Familial PD, cases with a clear family history, comprises 10% of all PD cases, of which known monogenic PD accounts for about 30% (S. Bandres-Ciga et al. 2020; Klein and Westenberger 2012). Such mutations are predominantly caused by rare variants with significant impacts, and certain monogenic forms intriguingly deviate from the prototypical Lewy body pathology seen in many PD patients (Johansen et al. 2018).

Traversing the spectrum, there are rare high-risk variants that exhibit incomplete penetrance but demonstrate some level of familial clustering. These rare variants, while not assuring the manifestation of PD, contribute substantially to its risk, often in conjunction with environmental factors.

Finally, at the opposite end lies common small-effect variants. Contrary to the earlier belief that PD had negligible genetic contribution, research over the past two decades underscores the importance of these variants in apparently sporadic forms of PD. Within the PD research community, a strong example of this was the extensive genome-wide association study (GWAS) by Nalls and colleagues, analyzing over 37,000 PD cases, 18,000 proxy cases, and 1.4 million controls in individuals of European ancestry, identifying 90 distinct signals associated with PD risk across 78 loci (Nalls et al. 2019a). While common in the population and enriched in PD cases, each of these variants only marginally escalate the risk. This adheres to the common disease-common variant hypothesis, which posits that multiple genetic markers nominated by GWASs are prevalent but confer minimal risk individually (Schork et al. 2009).

In the broader context, defining "variant" as any change in the DNA sequence that may influence gene function, potentially contributing to conditions like monogenic PD, and "risk" as the likelihood conferred by genetic variants—whether they are high or low frequency—to develop the disease, delineates the expansive scope of genetics in PD research (Lake et al. 2021). The uncovering of these genes and genetic variants not only offers hope to carriers but also highlights pathways for potential therapeutic interventions.

Genes Previously Associated with Parkinson's Disease

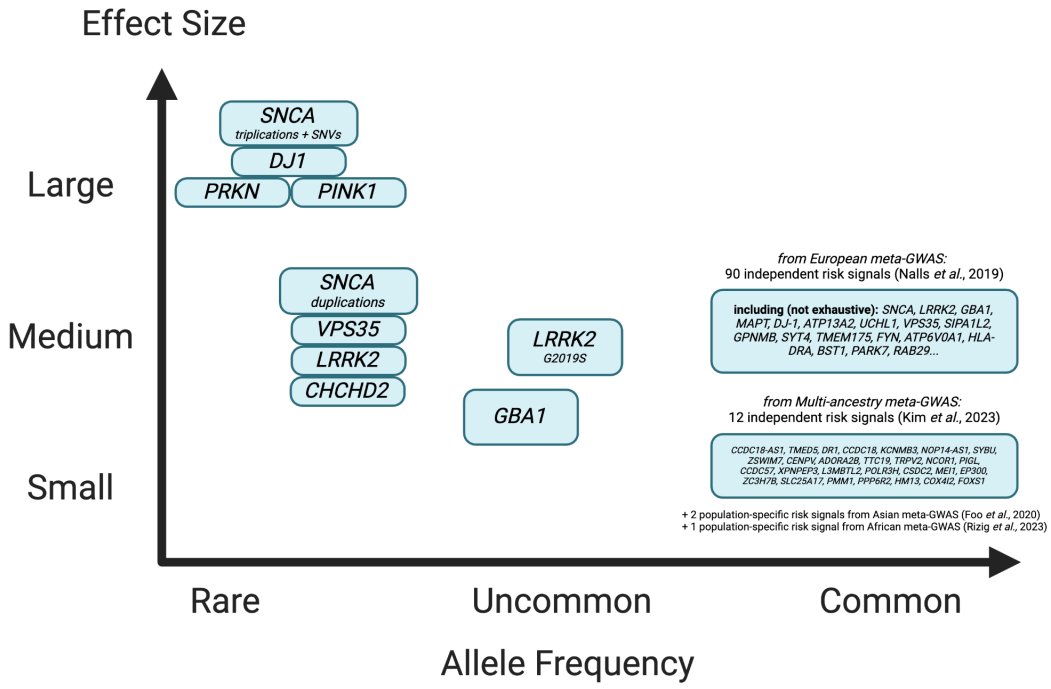


Figure 1: Manolio Plot for Parkinson's Disease

SNCA

In the intricate landscape of disease etiology, monogenic forms of the disease typically arise from rare mutations. These mutations are typically identified through examination of families, particularly in extensive families with a high incidence of affected members. A key discovery was made by Polymeropoulos and colleagues, when they identified mutations in *SNCA* within an Italian lineage (M. H. Polymeropoulos et al. 1997). This marked the first instance where mutations in a protein-coding gene were associated with autosomal dominant PD (Mihael H. Polymeropoulos et al. 1997, 1996; Klein and Westenberger 2012). The protein in question, alpha-synuclein, encoded by *SNCA*, undergoes misfolding and aggregation, culminating in the formation of Lewy bodies and neurites—both hallmark features of PD. Further advancing our understanding, in 2003, Singleton and colleagues discovered that a triplication in the *SNCA* locus is causative of PD (A. B. Singleton et al. 2003). This revelation underscored the notion that an overproduction of the alpha-synuclein protein, instigated by multiple copies of *SNCA*, is a sufficient trigger for the neurodegeneration characteristic of PD. Currently, the scientific community is directing considerable efforts towards targeting the alpha-synuclein protein. The objective is to mitigate the deleterious accumulation it induces within Lewy bodies and neurites, as this therapeutic avenue holds promise in both modifying the disease trajectory and stalling its onset and progression (Fields, Bengoa-Vergniory, and Wade-Martins 2019; Oliveira et al. 2021).

The genetic intricacies of PD further unravel when considering the rare missense variants and copy number augmentations of *SNCA*, both of which have been implicated in the autosomal dominant form of PD (A. B. Singleton et al. 2003; M. H. Polymeropoulos et al. 1997; Chartier-Harlin et al. 2004). Additionally, a range of non-coding common variants have been spotlighted for their role in elevating PD risk (Nalls et al. 2019a), and are associated with an earlier disease onset (Blauwendraat et al. 2019). Intriguingly, the underlying mechanism for this seems to range on the increased expression of *SNCA* (Pihlstrøm et al. 2018; Soldner et al. 2016). Collectively, this body of evidence positions *SNCA* at a central position regarding PD pathogenesis, prompting the current pharmacological endeavors aimed at modulating *SNCA* levels.

Pathogenic missense *SNCA* variants are rare in the general population, with five disease-causing mutations identified (p.A30P, p.E46K, p.G51D, p.A53T and p.A53G) (Blauwendraat, Nalls, and Singleton 2020) and these mutations can be associated with Parkinson's disease dementia, Lewy body dementia (LBD), and MSA (Scholz and Bras 2015). In a relative comparison, *SNCA* copy number variants emerge with a higher frequency, albeit remaining rare, with documentation spanning approximately 60 families (Kasten and Klein 2013). It is noteworthy that carriers of these *SNCA* mutations often exhibit a more aggressive disease trajectory, accompanied by a heightened prevalence of symptoms such as dementia, rapid eye movement (REM) sleep behavior disorder, and autonomic dysfunction. The age of onset for these carriers tends to skew earlier than idiopathic PD, with a predominant clustering around the late forties to early fifties. Beyond the conventional autosomal dominant inheritance patterns, there are intriguing reports of *SNCA* copy number gains manifesting through mosaic patterns (Perandones et al. 2014; Perez-Rodriguez et al. 2019; Mokretar et al. 2018). However, the broader implications of this in the overarching PD pathogenesis narrative warrant more expansive studies.

SNCA mutations and their role in PD underscores the pleomorphic risk locus hypothesis, illustrating how variations within a single gene can manifest in diverse outcomes (A. Singleton and Hardy 2011). Such findings illustrate the pleomorphic nature of the *SNCA* locus, with mutations leading to a spectrum of clinical presentations, from typical PD symptoms to more aggressive forms involving dementia and autonomic dysfunction.

GBA1

The *GBA1* gene encodes the lysosomal enzyme glucocerebrosidase (GCase), a pivotal player in PD etiology. GCase is crucial for the breakdown and recycling of glucosylceramide, a glycolipid; its dysfunction leads to the accumulation of glucosylceramide and related glycolipids in cells, affecting macrophage-monocyte cell lineage (Welsh et al. 2020).

GBA1 mutations are linked to both Gaucher's disease, an autosomal recessive lysosomal storage disorder, and significantly elevated PD risk, with a five to ten-fold increase depending on the population (Sidransky et al. 2009). *GBA1* represents a classic pleomorphic locus, showcasing a spectrum of coding, structural, and non-coding variants, each imparting varying degrees of PD risk (A. Singleton and Hardy 2011). *GBA1*-positive PD also presents with more severe non-motor symptoms clinically (Toffoli et al. 2023).

It is believed that reduced GCase activity due to *GBA1* mutations leads to the accumulation of glucocerebrosides and alpha-synuclein, contributing to neurodegeneration (Schapira 2015). This relationship highlights the therapeutic potential of enhancing GCase activity as a novel intervention in PD (Migdalska-Richards and Schapira 2016). However, the hypothesis that *GBA1* mutations lead to reduced GCase activity, causing accumulation of glucocerebrosides and alpha-synuclein and contributing to neurodegeneration (Schapira 2015), has been complicated by the clinical trial results of Venglustat. This drug, aimed at reducing glucocerebroside levels based on the theory that such reduction could mitigate alpha-synuclein aggregation and provide neuroprotective effects, did not significantly impact disease progression in trials (Giladi et al. 2023). This outcome suggests that the pathophysiological relationship is more complex than initially believed, pointing to the possibility of additional or alternative mechanisms of disease beyond the simple accumulation of glucocerebrosides and alpha-synuclein.

Among the specific mutations in *GBA1*, p.N370S (rs76763715) and p.L444P (rs421016) are notable. The p.N370S mutation is most prevalent in Ashkenazi Jewish individuals and contributes significantly to PD risk, presenting as a milder form of Gaucher's disease (Balwani et al. 2010). The prevalence and impact of these *GBA1* mutations demonstrate notable variation across populations, especially among Ashkenazi Jews (Toffoli, Smith, and Schapira 2020; Gan-Or et al. 2008).

LRRK2

Leucine-Rich Repeat Kinase 2, or *LRRK2*, also known as *PARK8*, is integral to PD research, particularly following the discovery of mutations in families with late-onset PD (Funayama et al. 2005; Paisán-Ruíz et al. 2004; Zimprich et al. 2004). These mutations exhibit an autosomal dominant inheritance pattern, appearing in about 5-6% of familial PD cases and in less than 1% of sporadic PD cases, thus highlighting a significant genetic factor in familial PD (Di Fonzo et al. 2005; Brice 2005).

The prevalence of *LRRK2* mutations exhibits remarkable variation among different populations: they are identified in up to 28% of PD cases in the Ashkenazi Jewish community (Ozelius et al. 2006) and in as many as 41% of cases in the North African population (Inzelberg, Hassin-Baer, and Jankovic 2014), indicating a significant ethnic and geographical influence on PD risk associated with *LRRK2*.

A key feature of *LRRK2* in PD pathology is its influence on kinase activity; mutations in the *LRRK2* gene have been shown to increase this activity (Gloeckner et al. 2006). This finding is crucial for understanding the molecular mechanisms driving PD and for developing targeted therapies. Further research into *LRRK2* has also suggested potential links between these mutations and specific clinical features of PD, such as the rate of disease progression and the severity of motor and non-motor symptoms (Wojewska and Kortholt 2021). Additionally, ongoing studies are exploring how *LRRK2* interacts with other proteins and pathways within cells. These insights offer promising avenues for personalized medicine approaches in PD treatment, tailoring interventions to individual genetic profiles (Sen and West 2009; Eschbach and Danzer 2014).

Recessive Genes: *PINK1* and *PRKN*

Two genes usually associated with early-onset PD are *PRKN* (also known as parkin) and *PINK1*. *PRKN* mutations were first identified in individuals with autosomal recessive juvenile parkinsonism (ARJP) in Japanese families (Ishikawa and Tsuji 1996). *PRKN* encodes the E3 ubiquitin ligase parkin (Cookson 2012), and patients with these mutations do not typically have Lewy bodies when examined postmortem (Takahashi et al. 1994). *PINK1* mutations are rare, and were first identified in families with autosomal recessive inheritance and have earlier parkinsonism onset with slower progression (Valente et al. 2004). The proteins encoded by *PRKN* and *PINK1* are important for mitochondrial regulation and mitophagy, and mutations disrupt these functions, indicating that dysregulation of mitochondrial pathways are key mechanisms in PD pathogenesis (Pickrell and Youle 2015; Cookson 2012).

Other Rare Genetic Mutations Associated with Parkinson's Disease

In addition to the well-characterized genes such as *GBA1*, *SNCA*, *LRRK2*, *PINK1*, and *PRKN*, PD is also associated with mutations in less commonly implicated genes that contribute to its complex genetic landscape. Notable among these are genes like *DJ1*, *ATP13A2*, *PLA2G6*, *FBXO7*, *VPS35*, *CHCHD2*, and *VPS13C*, which, although less common, play significant roles in PD pathogenesis. These rare genetic mutations showcase the broader spectrum of genetic contributors that affect various cellular pathways, including mitochondrial function, autophagy, and inflammatory responses in PD (Jia, Fellner, and Kumar 2022; W. Li et al. 2021; Reed et al. 2019).

Overlap between Monogenic and Idiopathic Forms of Parkinson's Disease

There is, however, some overlap between the monogenic and idiopathic forms of PD, with some monogenic PD genes representing pleomorphic risk loci for idiopathic PD. Pleomorphic risk loci are defined as a risk at individual loci that fall on a spectrum of allele frequency and effect size, with loci capable of enveloping low-risk and high-risk variants. Both SNVs and CNVs can result in increased expression of certain genes, indicating a common pathway through which genetic variations exert their effects on disease risk (A. Singleton and Hardy 2011).

Polygenic Risk Scores

Known rare and/or common variants can be combined to calculate a polygenic risk score (PRS). PRS are subtypes of genetic risk scores (GRS). They are used to evaluate one's overall risk or predisposition to a disease or phenotype, derived from case-control GWAS considering comprehensive genetic architecture, including the effects of both protective and risk variants. PRS per individual is calculated by taking the sum of risk alleles weighted by the effect size (in an external dataset), calculated from a GWAS, and indicates an individual's susceptibility to the disease based on alleles present in their genome. An area of active research, PRS for PD affection status, age-at-onset, and even specific symptoms using different methodologies, have been developed (Koch et al. 2021). For example, PRS can be used to quantify the effect of common variant risk for PD, for example in modifying the penetrance of the disease, as in the case of *LRRK2* p.G2019S carriers on PD (Iwaki et al. 2020). The heritability of PD driven by common

genetic variation is estimated to be around 22% and approximately a third of this variation is explainable by the largest common genetic variation study in individuals of European ancestry (Elsayed et al. 2021; Nalls et al. 2019b). PRSs highlight the importance of understanding the genetic predisposition for disease and evaluation for risk prediction. However, given that GWASs performed in PD are primarily on individuals of European origin, the statistics generated and subsequently used for PRS are best suited for Europeans and underperform in non-European populations is a significant limitation to consider (Duncan et al. 2019; A. R. Martin et al. 2020).

Role of Genomics in Parkinson's Disease

Machine Learning

In the realm of computational biology, machine learning (ML) emerges as a promising direction allowing computational systems to learn and make inferences from large amounts of data. ML algorithms can analyze large volumes of data, recognizing complex patterns and relationships that might be unobservable to humans.

ML is a subset of artificial intelligence (AI) focused on building systems that learn from data, identifying patterns, and making decisions with minimal human intervention. ML specifically involves algorithms that improve automatically through experience. In the computational biology space, ML algorithms are particularly adept at weighting genetic variants and developing predictive models, thanks to their ability to handle the vast complexity and volume of genomic data. These algorithms can assign importance to various genetic factors based on their influence on a particular trait of interest, enabling the creation of models that can predict disease risk or therapeutic outcomes. A successful application of ML in genomics is in cancer genomics, where ML models have identified novel cancer-driving mutations and have helped in predicting patient outcomes and treatment responses (Moon et al. 2023). Another example is in predictive models for cardiovascular diseases, where ML algorithms analyze genetic and clinical data to predict an individual's risk, demonstrating the potential of ML to develop personalized medicine and disease prevention strategies (Pal et al. 2022).

This ability is particularly useful in genomics, where the datasets are not only large but also complex (Libbrecht and Noble 2015). Genomic datasets are characterized by their high dimensionality and volume. ML provides tools to handle this complexity efficiently, enabling the extraction of meaningful insights from vast amounts of data (Stephens et al. 2015; Min, Lee, and Yoon 2017). It allows for more precise and comprehensive analyses, which are essential for understanding genetic diseases and developing personalized medicine strategies (Angermueller et al. 2016).

In recent years, the demand for machine learning (ML) expertise has outpaced the supply, despite the surge of people entering the field. To address this gap, there have been significant strides in the development of user-friendly machine learning software that can be used by non-experts (**Table 1**). The first steps toward simplifying machine learning involved developing simple, unified interfaces to a variety

of machine learning algorithms (e.g., scikit-learn (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, and Others 2011), XGBoost (Chen and Guestrin 2016), LightGBM (Ke et al. 2017), TensorFlow (Abadi et al. 2016), PyTorch (Paszke et al. 2019). Although these packages have made it easy to experiment with machine learning, there is still a fair bit of knowledge and background in data science required to produce high-performing and usable machine learning models.

However, different data types require different ML pipelines and different custom functionality defined by domain expertise. The development of ML models for genomics (genetics and multi-omics) data, in particular, is notoriously difficult for a non-expert. These data modalities require significant domain expertise to clean, pre-process, harmonize and perform quality control (QC) (Eraslan et al. 2019). Furthermore, tuning, validation, and interpretation involve taking into account the biology and the limitations of the underlying data collection, protocols, and technology.

The most recent strategic vision published by the National Human Genome Research Institute stated that the features of the epigenetics and transcriptomics will be incorporated into predictive models of the effect of genotype on phenotype routinely by the year 2030 (Green et al. 2020). Biomedical researchers are currently at the convergence of two scientific advances that will allow 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 ML pipelines and artificial intelligence, to maximize the value of this massive amount of readily available data (Sudlow et al. 2015). Previous work in this space to use ML to predict PD risk and onset use data types such as analyzing gait (Palmerini et al. 2017), fall detection methods (Silva de Lima et al. 2017), and other motor data (Noyce et al. 2014), or sleep behaviors (Campabadal et al. 2021).

To encapsulate the evolving landscape of machine learning in the field of PD research, it is evident that the future holds great promise. The integration of machine learning with extensive genomic, clinical, and *omics data offers a path for understanding PD's complex etiology (**Table 1**). As computational methodologies advance, they will enable researchers to unearth deeper insights into the disease's progression, risk factors, and individual variability in response to treatment, offering hope for improved patient outcomes and a deeper understanding of this multifaceted, complex disease.

Supervised Learning	**Unsupervised Learning**
Requires "teaching" the model by training the data with a *labeled* dataset	Requires machine learning algorithms draw conclusions that may not be visible to the human eye on *unlabeled* data
Analogous to learning in the presence of a	Analogous to self-guided learning where one

teacher or supervisor	works on their own to discover information
Relevant for data where we know what kind of data we are dealing with and know what to expect	Relevant for data where we know little to nothing about the data
Includes algorithms for *classifying* labeled data and *regression* (predicting trends using previous labeled data)	Includes algorithms for *clustering* (finding patterns and groupings from unlabeled data) and *association* (establishing similarities amongst data objects inside large databases)
Pros include less computational complexity and higher level of accuracy	Pros include not requiring output data and being able to classify big data

Table 1: Overview of differences between supervised and unsupervised machine learning.

Overarching Themes and Aims in this Dissertation

The overarching challenges and aims addressed in this dissertation are as follows:

1. **Reviewing the historical landscape and current limitations:** To synthesize previous research findings, emphasize the existing lack of diversity in genetic studies, highlight the potential benefits of Machine Learning (ML) in this space, and position the current research within the broader context of PD studies (Chapter 1).
2. **Assessing the structural variants using UKBiobank:** To investigate structural variants in *SNCA* by leveraging data from the UKBiobank, providing insights into potential genetic markers or factors and highlighting the importance of a scalable approach (Chapter 2).
3. **Investigate rare variants in all genes across multiple large-scale datasets:** To analyze large-scale cohorts, examining rare variants in both recognized and potential candidate genes, with an emphasis on their influence on broader genetic pathways and PD manifestations and the benefits of a scalable approach (Chapter 3).
4. **Broadening the GWAS horizon beyond European populations:** To conduct the first African and African admixed genome-wide association study, a population outside the European ancestry, aiming to unearth population-specific genetic risk factors for PD and the benefits of a global approach (Chapter 4).
5. **Leveraging machine learning and multi-omics data:** To introduce the GenoML platform and use the AMP-PD dataset—encompassing genetic information, clinicodemographic details, and transcriptomics data—to develop a robust ML model that evaluates PD diagnoses, followed by an in-depth analysis of its findings and the benefit of global and scalable approaches (Chapter 5).

6. **Conclusions and future directions:** To collate the research findings, derive conclusions, engage in thorough discussions, and evaluate the overall impact of the presented work, while also charting potential future avenues in PD genetic research (Chapter 6).

Chapter 2: Investigating Structural Variants in *SNCA* Leveraging the UK Biobank

Overview and Broader Relevance

SNCA, encoding alpha-synuclein, plays a pivotal role in the pathogenesis of PD. Variations in *SNCA*, particularly structural variants, are of significant interest in understanding the molecular mechanisms underlying PD. Structural variants include a range of genomic alterations such as duplications, deletions, and more complex rearrangements, which can disrupt the gene's function or alter its expression levels.

In the context of PD, duplications and triplications of *SNCA* have been linked to familial forms of the disease, suggesting a dosage effect where increased alpha-synuclein protein leads to neuronal toxicity. Additionally, SNPs in *SNCA* have been associated with sporadic PD , further supporting the gene's importance in the context of PD. For instance, therapies aiming to reduce alpha-synuclein expression or mitigate its toxic aggregation could be particularly effective in individuals with *SNCA* duplications or triplications, or in patients with SNPs that increase the level of *SNCA* expression (eQTLs).

In this chapter, I conduct a study focusing on UK Biobank data, investigating the frequency and nature of these structural variants in *SNCA*. Exploring these variants in a large cohort like the UK Biobank provides valuable insights into their prevalence in the general population and their contribution to PD risk.

This work has been published here:

> Blauwendraat C*, **Makarious MB***, Leonard HL, Bandrés-Ciga S, Iwaki H, Nalls MA, Noyce AJ, and Singleton AB: "**A Population-Scale Analysis of Rare *SNCA* Variation in the UK Biobank**" *Neurobiology of Disease (2021); https://doi.org/10.1016/j.nbd.2020.105182*

Introduction

PD is a complex neurodegenerative disorder which is likely to be caused by both genetic and environmental risk factors. Its pathogenesis is closely linked to genetic components, with *SNCA* playing a crucial role. This gene encodes the alpha-synuclein protein, which is found in aggregates in Lewy bodies, a pathological hallmark of PD.

The genetic landscape of PD is intricate. There are rare damaging variants and more common risk variants, contributing to the disease in different ways. Notably, rare missense variants and copy number gains in *SNCA* cause autosomal dominant PD, as identified in several studies (A. B. Singleton et al. 2003; M. H. Polymeropoulos et al. 1997; Chartier-Harlin et al. 2004). These pathogenic missense *SNCA* variants are rare in the general population, but there are well reported families who carry *SNCA* mutations, such as p.A30P, p.E46K, p.G51D, p.A53T, and p.A53G. They are associated with more rapid progression and

extreme PD presentations, often including dementia, rapid eye movement sleep behavior disorder, and autonomic dysfunction. *SNCA* CNVs are more frequent than missense variants but still rare, with about 60 families reported. These variants are linked to an earlier onset of PD, typically in the late forties or early fifties (Book et al. 2018; Konno et al. 2016). Large scale studies of familial PD indicate that *SNCA* multiplications account for ~0.5% of familial PD cases (Schmaderer et al. 2023).

Moreover, non-coding common variants in *SNCA* have been found to moderately increase PD risk (Nalls et al. 2019a) and lead to earlier disease onset (Blauwendraat et al. 2019). These variants likely increase *SNCA* expression, contributing to disease pathogenesis (Pihlstrøm et al. 2018; Soldner et al. 2016). The presence of these variants in the broader population was examined in large-scale studies like the UK Biobank cohort (Bycroft et al. 2018).

In addition to the more typical autosomal dominant inheritance patterns, *SNCA* copy number gains have been reported through mosaic patterns (Perandones et al. 2014; Perez-Rodriguez et al. 2019; Mokretar et al. 2018). However, further research is needed to establish whether this represents a general PD pathogenesis mechanism. Current research efforts are focusing on developing drugs that target *SNCA* expression. The central role of *SNCA* in PD pathogenesis makes it a prime target for therapeutic intervention. By lowering *SNCA* expression, for example using anti-sense olgionuclelotide therapy it is hoped that the progression of PD can be slowed or halted.

Understanding the interplay of genetic factors in PD is crucial for developing effective treatments and for the early diagnosis and management of the disease. The background population frequency of copy number variants in *SNCA* is unknown. Here, I analyze the frequency and types of these *SNCA* structural variants in the UK Biobank to understand prevalence in the general population and their contribution to PD risk using scalable approaches.

Methods

The data used in this study can be accessed publicly by submitting an application to the UK Biobank (available at [https://www.ukbiobank.ac.uk/]) and gnomAD (accessible at [https://gnomad.broadinstitute.org/]). The code developed for preprocessing, visualization, and analysis is available on our GitHub repository, which can be found at [https://github.com/neurogenetics/UKbiobank_SNCA]. This study encompassed data from 488,377 individuals from the UK Biobank, accessed in 2020, focusing on the *SNCA* gene region.

UK Biobank Phenotype Data

Phenotype Data Collection and Prodromal Phenotype Assessment

Phenotypic data from the UK Biobank was comprehensively sourced using a variety of field codes to gather detailed health and demographic information. This included data derived from the International Classification of Diseases, 10th Revision (ICD-10) codes (field code: 41270), which provided insights into a wide range of diagnosed conditions. Specific attention was paid to Parkinson's Disease (PD; field code:

131023), and further enriched by gathering information about the illnesses of both the father and mother of participants (field codes: 20107 and 20110). Additional data were collected regarding cases of parkinsonism (field code: 42031) and dementia (field code: 42018), both of which are critical for understanding neurodegenerative diseases. The genetic ethnic grouping of participants was also considered (field code: 22006), along with their year of birth (field code: 34) and age at the time of recruitment into the study (field code: 21022).

In order to assess potential prodromal phenotypes of PD, the MDS Prodromal Criteria was employed, as detailed elsewhere (Heinzel et al. 2019). This criteria is a set of guidelines developed to identify early signs of PD before the onset of its classic motor symptoms. The use of these criteria in conjunction with the detailed phenotypic data collected from the UK Biobank allowed for a more nuanced and early detection of potential PD cases and if they are associated with the structural variation in *SNCA*.

UK Biobank Genotype Data

Data Acquisition and Preparation

The genotype data, including B-allele frequency and Log2 ratio values, were obtained from the UK Biobank's version 2 June 2020 release. This dataset included information on the larger *SNCA* gene region located on chromosome 4 (hg19; chr4:84992449-96412248).

B-allele Frequencies

B-allele frequency (BAF) refers to the frequency of the “B” allele at a given SNP in a sample. SNPs are often denoted as A/B, where “A” and “B” represent different alleles. BAF is the proportion of the “B" allele in the sample, with values ranging from 0 (no “B” alleles) to 1 (all “B” alleles). To calculate BAF, the number of reads for the “B” allele is divided by the total number of reads for both “A” and “B” alleles at a specific SNP. For example, if there are 20 reads for the “A” allele and 80 reads for the “B” allele, the BAF would be $\frac{80}{20+80} = 0.8$. BAF is particularly useful in detecting CNVs in a genome.

In humans, a normal diploid organism, the BAF for heterozygous SNPs should be around 0.5. Deviations from this expected value can indicate the presence of duplications or deletions. For example, a BAF closer to 0.33 or 0.66 might suggest a trisomy at that locus, as one of the alleles is present in an extra copy. Note that deletions typically do not have any 0.5 values because most “A” genotypes will be 0 and most B genotypes will be 1, however this should not be confused with homozygous stretches of DNA, where all AA genotypes will be 0 and all “B” genotypes will be 1. At a deletion or homozygous stetch, there will be some single “A” and “B” (where not all will be homozygous “A”, some will be homozygous “B”). See **Table 2** for a summary of how to interpret BAF. These value ranges do not capture somatic mosaicism, where partial deletions and duplications can confound the values, but we consider somatic mosaicism in the manual revision of the plots.

BAF Approximate Value	Genotype	Type
0	AA	Normal; 2 genotypes
0.75	ABBB	Triplication; 4 genotypes
0.66	ABB	Duplication; 3 genotypes
0.5	AB	Normal; 2 genotypes
0.33	AAB	Duplication; 3 genotypes
0.25	AAAB	Triplication; 4 genotypes
1	BB	Normal; 2 genotypes

Table 2: Summary of B-allele frequency values and corresponding genotypes and type of structural variation. BAF: B-allele frequency.

Log2 Ratios

The Log2 Ratio (L2R) is a measure used to detect CNVs based on the intensity of the signals obtained during genotyping. The expected intensity is typically derived from a reference set of normal samples. For example, if the observed intensity is twice the expected intensity, the L2R would be log2(2/1) = 1. L2R is valuable in identifying genomic gains and losses. A higher than normal L2R value indicates a gain (duplication), while a lower than normal value suggests a loss (deletion). See **Table 3** for a summary of how to interpret L2R.

L2R Value	Genotypes	Type
Less than 0	A or B	Deletion; 1 genotype
0	AA, AB, or BB	Normal; 2 genotypes
Greater than 0	ABB or AAB	Duplication; 3 genotypes
Greater than or equal to 0.75	ABBB or AAAB	Triplication; 4 genotypes

Table 3: Summary of Log2 ratios and corresponding genotypes and the type of structural variation. L2R: Log2 ratio.

Specific Region Analysis

SNCA was analyzed in three specific regions:

- The larger *SNCA* region gene ±5 megabases (Mb)(hg19; chr4:85645250-95759466)
- The *SNCA* gene ±0.5 Mb (hg19; chr4:90145250-91259466)
- The *SNCA* gene body (hg19; chr4:90645250-90759466)

For a broader chromosomal analysis, the ±20 Mb region surrounding *SNCA* (hg19; chr4:70645250-110759466) and the entirety of chromosome 4 (hg19; chr4:1-180915260) were examined.

BAF values approximately at 0.66 and 0.33 suggested the likelihood of duplications. For each region, I computationally quantified the number of BAF values for variants within specific intervals. Specifically, I counted the number of BAF values for variants within two ranges: greater than 0.65 but less than 0.85, and greater than 0.15 but less than 0.35. This counting was conducted for each of the three regions under study. In cases where the count of variants exceeded six, those samples were selected for more detailed examination. Additionally, I calculated the average L2R values for each region. High average L2R values were indicative of potential duplications, whereas low average values pointed towards the possibility of deletions. This approach allowed for a more nuanced understanding of the genomic variations in the regions of interest leveraging information from the BAF and L2R values to prioritize samples for visual inspection. In our analysis, we inferred mosaicism events through careful examination of both L2R and BAF plots, looking for patterns indicative of mixed cell populations with different copy number states. These patterns included intermediate shifts in L2R values and BAF patterns deviating from the expected 0, 0.5, and 1 positions. We adopted a conservative approach by using "mosaicism" as an umbrella term for complex structural events that did not present exactly expected representations in both L2R and BAF plots. By choosing to do this, it allowed us to capture and report genomic complexity without overcommitting to specific interpretations, thereby reducing the risk of false positives and acknowledging the limitations of our semi-automated approach. Such an approach is particularly valuable in the context of mosaic events, which often manifest as subtle or ambiguous signals in genomic data.

Among the almost 500,000 individuals in this release, a thorough manual inspection was conducted on 363 cases prioritized using BAF and L2R values. This examination resulted in the discovery of 26 samples exhibiting signs of possible genomic events. These events encompassed a range of genetic alterations such as duplications, deletions, and other more intricate variations. For samples exhibiting significant genomic events, all other autosomal chromosomes were also thoroughly examined. For effective visualization and analysis, the LOWESS (Locally Weighted Scatterplot Smoothing) method was applied with a variable smoother span, adjusted according to the total graph size (ranging from 0.01 to 0.0001). This allowed for the calculation of more refined smoothed averages. Additionally, bespoke functions were developed in R to visually represent and interpret the genomic variations.

Calculation of Relatedness

The UK Biobank genotype data (v2) was also used to calculate relatedness among individuals, to identify potential relatives between individuals studied, using PLINK (v1.9; (C. C. Chang et al. 2015).

UK Biobank Exome Sequencing Data

Data Acquisition and Processing

The exome sequencing data from the UK Biobank (FE dataset encompassing field codes 23160 and 23161) were acquired in June 2020. This dataset included exome sequencing from a total of 49,960 individuals. ANNOVAR was used for variant annotation (K. Wang, Li, and Hakonarson 2010). This process involved a screening for known pathogenic *SNCA* variants, which included p.A30P, p.E46K, p.G51D, p.A53E, and p.A53T. These variants were cross-referenced with reported pathogenic data from reputable sources such as ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) and the Human Gene Mutation Database (HGMD; Professional version 2020.2 from Qiagen), along with analysis of loss-of-function (LoF) variants encompassing stop codons, frameshift mutations, and splicing errors.

Additionally, exome sequencing data were used to validate and replicate genomic events, leveraging the allelic depth (AD) field from each individual's exome sequencing VCF (variant call format) files. A ratio of AD was calculated by dividing the highest depth value by the lowest for each variant, providing a measure of how many times each allele of a variant is observed in the sequencing data, a quantitative measure of allelic representation. By comparing the AD ratio across different samples, we can confirm the presence of these events and assess their frequency. This AD ratio was then visualized using the ggplot2 package in R (Wickham 2016), offering a visual representation of the genomic variations.

For quality control and comparative analysis, ten random samples were selected as negative controls, ensuring the robustness of the data interpretation. Furthermore, to provide a broader context and enhance the reliability of the findings, additional population allele frequencies were sourced from the Genome Aggregation Database (gnomAD, https://gnomad.broadinstitute.org/, version 2.1.1, June 2020; (Karczewski et al. 2020).

Results

Assessment of SNCA Mutations in UK Biobank Cohort

In this comprehensive study of the UK Biobank, my focus was on assessing the presence and frequency of known, potentially pathogenic mutations in the *SNCA* gene, which is widely recognized for its association with PD. The analysis was centered around five well-established *SNCA* variants, specifically p.A30P, p.E46K, p.G51D, p.A53E, and p.A53T. None of these variants were detected in the exome sequencing data from the UK Biobank. Their absence suggests either a low prevalence of these specific mutations in the studied population or potential gaps in the detection capabilities of the sequencing methods used.

Despite this, the study did uncover two *SNCA* variants with uncertain or conflicting pathogenic status, detailed in **Supplementary Table 1**. The frequencies of these variants in the UK Biobank cohort were found to be in line with those reported in the gnomAD database, a global reference for human genetic

variation. This consistency reinforces the reliability of the findings and provides a comparative context for understanding the prevalence of these variants.

None of the individuals carrying these *SNCA* variants had a reported clinical diagnosis of PD in the UK biobank. However, a familial link to PD was noted in a small subset of these individuals. Specifically, thirteen subjects were identified as carriers of the *SNCA* p.P117T variant. Among them, one individual had a parent with a PD diagnosis. Additionally, 28 subjects carried the *SNCA* p.H50Q variant, and within this group, two had a parent who had been diagnosed with PD. This familial pattern hints at a possible hereditary component or increased susceptibility to PD in carriers of these variants.

It is, however, crucial to note that these observations did not attain statistical significance, as indicated by p-values greater than 0.1 in a 2x2 Fisher exact test. This lack of statistical significance suggests that while there may be a biological or genetic link, the evidence is not strong enough to establish a definitive connection between these *SNCA* variants and PD.

Analysis of SNCA Copy Number Variants

Identification of Subjects with *SNCA* Variants

In my examination of *SNCA* CNVs in the UK Biobank genotyping data, I identified 30 individuals with distinct genetic variations. These individuals were classified into three groups based on their specific types of *SNCA* variations. The first group included six individuals with duplications of *SNCA*. This type of genetic change is noteworthy for its potential impact on the gene's function and expression. Another six individuals formed the second group, characterized by deletions in *SNCA*. Deletions represent a significant alteration in the gene and may have implications for gene function and disease risk. The third group, comprising 14 individuals, was distinguished by large, complex genetic events involving *SNCA*. These events were notable for showing significant mosaicism, indicating a varied genetic composition within the individuals' cells. Further details about these large complex events, including the specific nature of the genetic alterations, are outlined in **Supplementary Table 2**.

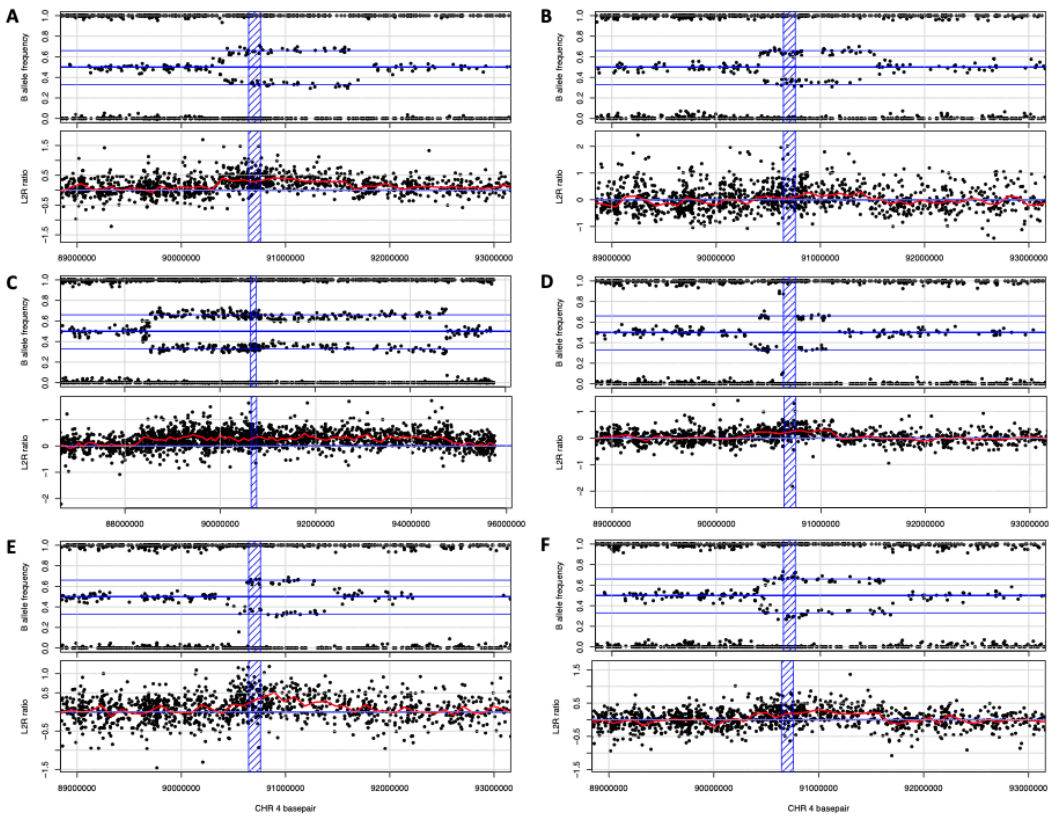


Figure 2: Six whole gene *SNCA* duplications were identified in the UK Biobank cohort

Identification of Subjects with Duplications and Deletions in *SNCA*

Regarding *SNCA* duplications, we identified six potential cases. Each of these cases displayed clear changes in both BAF and L2R, consistent with a *SNCA* duplication (**Figure 2**). The duplications varied in size but averaged around 2Mb. The smallest observed duplication was approximately 0.8Mb, while the largest extended up to about 6Mb. This variation in size suggests a range of effects that these duplications could have on gene function and expression. Regarding deletions, we found six potential cases. Five of these were complete deletions of *SNCA*, while one appeared to be a partial deletion. The partial deletion is of particular interest as it could, like a complete deletion, also lead to haploinsufficiency; where a partial loss of gene function occurs due to having only a single functional copy of a gene that is not sufficient to maintain normal function (**Figure 3**). The average size of these deletions was estimated to be around 0.76Mb, with the smallest being about 0.2Mb and the largest close to 1Mb. The range in sizes of both duplications and deletions highlights the genetic diversity in *SNCA* variations.

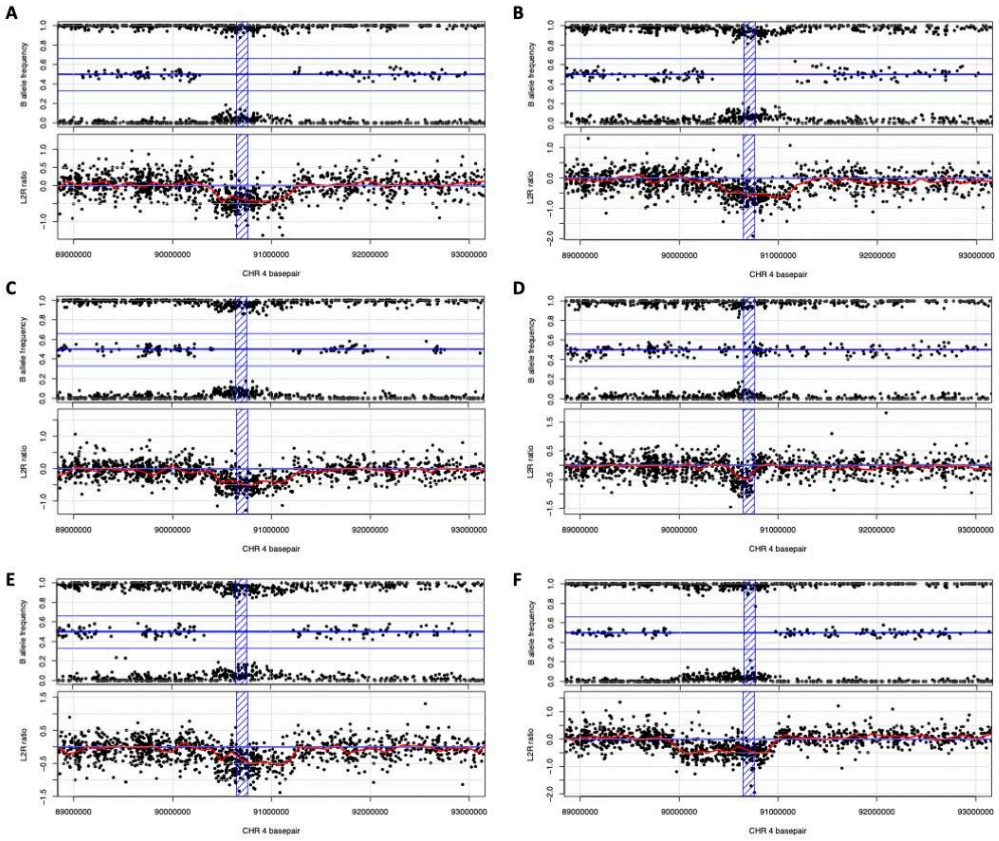


Figure 3: Five full *SNCA* deletions and one (likely) partial *SNCA* deletion (D) were identified in the UK Biobank cohort

Identification of Subjects with Complex Events in *SNCA*

The discovery of 14 complex events in the *SNCA* region were discovered through a thorough examination of the *SNCA* region, extending ±20Mb from the gene, as well as a comprehensive review of the entirety of chromosome 4 and all other autosomes. I identified 14 of these complex events (**Supplementary Figure 1**).

My findings indicated that the majority of these complex events could be instances of mosaicism, where different cells within the same individual have different genetic compositions. Notably, these mosaicism events predominantly spanned a substantial portion of the long arm of chromosome 4, specifically from regions 4q11 to 4q35. This suggests a widespread genetic variability within individual cells, particularly in the region of chromosome 4 where *SNCA* is found. Interestingly, no significant events akin to these were identified on other chromosomes within the individuals studied, underscoring the unique nature of these occurrences in the context of *SNCA* and, on a broader scale, chromosome 4.

Upon a more detailed analysis, 11 of these 14 complex events were classified as large complex events, each exhibiting varying levels of mosaicism. This variation highlights the genetic diversity and complexity

inherent in these cases. Two of these eleven events were particularly noteworthy, as they were likely mosaic deletions resulting from uniparental disomy events, where an individual receives two copies of a chromosome, or part of a chromosome, from one parent and no copy from the other parent.

Validation Using Exome Sequencing Allelic Depth Data

In order to validate the results I obtained from the genotyping array data regarding the *SNCA* CNVs, we also examined the exome sequencing data. This process focused on analyzing the AD data for all heterozygous variants present on chromosome 4. Out of the 30 subjects identified with potential *SNCA* variations, I was able to apply this method of validation to two specific cases. These were subject #dup3, who had an identified *SNCA* duplication, and subject #comp8, who exhibited a complex *SNCA* event. In both instances, these subjects showed clear differences in AD at the genomic locations corresponding to their respective *SNCA* events (**Figure 4**).

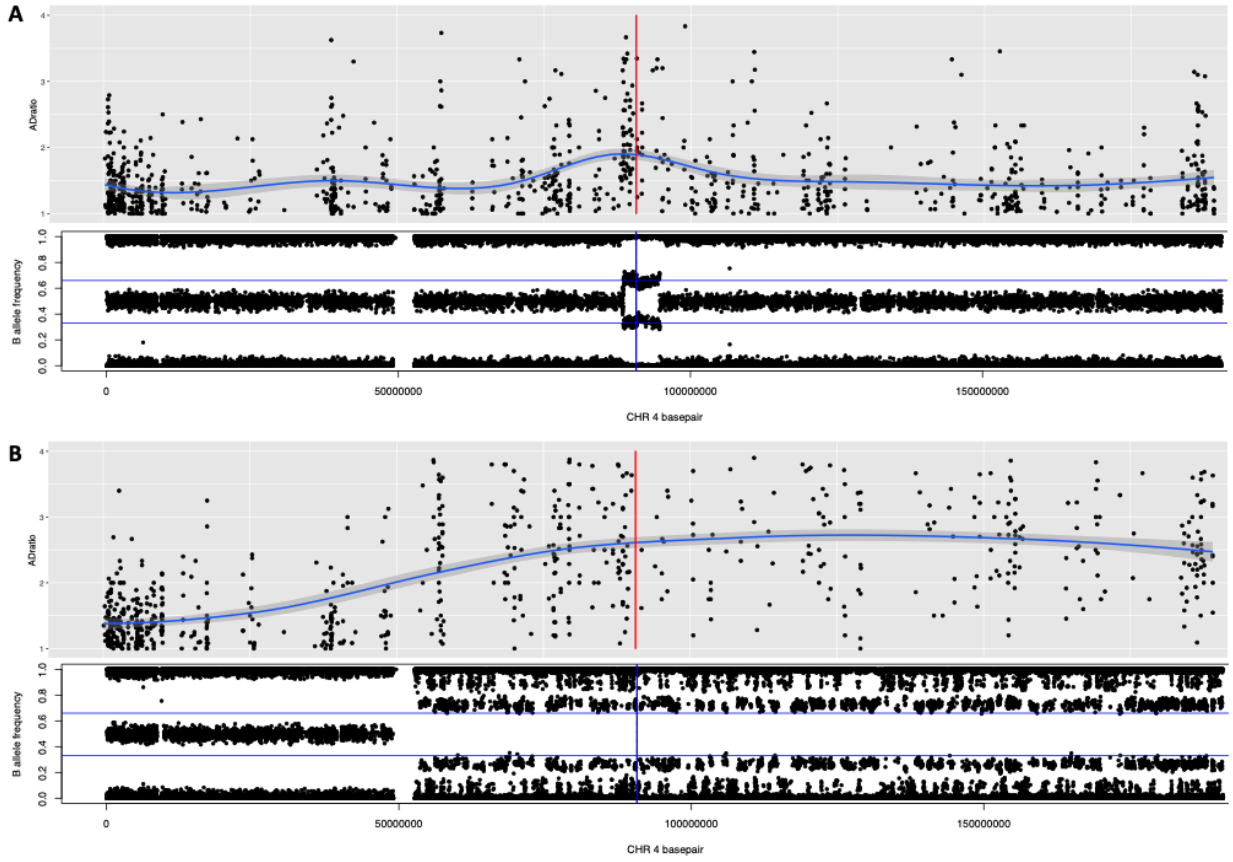


Figure 4: Partial validation of genomic events of the genotyping array data using exome sequencing data

A) Duplication in the *SNCA* gene region (subject #dup3), B) Partial duplication of the long arm of chromosome 4 (subject #comp8). Red and blue vertical line represents the *SNCA* gene body.

To ensure the robustness of our findings, we included a comparative analysis using ten randomly selected subjects from the exome sequencing data as negative controls. These control subjects were crucial for establishing a baseline of AD in the absence of *SNCA* copy number variations. None of these control subjects exhibited any evidence of allelic imbalance (**Supplementary Figure 2**). This lack of imbalance in the control group helped to reinforce the validity of the allelic depth differences observed in subjects #dup3 and #comp8.

Phenotypic Data Analysis of Subjects with SNCA Alterations

Interestingly, none of the individuals with potential *SNCA* alterations were reported to have PD, parkinsonism, or dementia. This was further confirmed by reviewing the ICD-10 codes, where no diagnosis of PD was found among these individuals. This absence of explicit PD or related diagnoses in the records of these individuals was an unexpected yet important observation, given the established links between variations in *SNCA* and PD.

However, considering the complexity and variability of PD's clinical presentation, especially in its early stages, I acknowledged that some of these individuals might be in the prodromal phase of PD, where non-specific symptoms precede the more definitive motor symptoms of PD. To explore this further, I investigated the three participants who had prodromal PD symptoms noted. These three were then evaluated using the revised MDS Prodromal Criteria, an evaluation designed to identify individuals who are likely in the prodromal phase of PD. Despite this assessment, none of the three subjects met the threshold for probable prodromal PD (**Table 4**).

LNG ID	Mutation	Age	Sex	Prodromal features	MDS pooled LR	MDS Probablity
del2	Deletion	55	Male	Depression, anxiety, IBS (constipation)	4.8	3.50%
comp7	Complex/mosaic	52	Male	Depression, head injury	1.9	0.10%
comp13	Complex/mosaic	67	Male	Autonomic dysfunction	22.2	31%

Table 4: Overview of prodromal ICD-10 codes in *SNCA* copy number variant carriers.

LNG: Laboratory of Neurogenetics; IBS: Irritable bowel syndrome; MDS: Movement Disorders Society; LR: Likelihood ratio.

Additionally, I observed familial links to other neurodegenerative disorders in some indviduals. Two individuals (#comp3 and #comp6) reported having a parent with dementia. Notably, one of these subjects (#comp6) also had a parent with PD. While these familial links are intriguing, particularly in the context of genetic predisposition to neurodegenerative diseases, they were not statistically significant. Using a Chi-square test yielded a p-value greater than 0.1, indicating that these observations could not be linked to the variations in *SNCA*.

Of the 14 individuals with mosaic patterns, four of them (approximately 29%) had a documented history of blood-based cancers. These cancers included lymphoma, multiple myeloma, and myelodysplastic syndrome (**Supplementary Table 2**). This suggests a potential link or coincidence between mosaicism in *SNCA* and the occurrence of certain blood-based cancers. Originally believed to be exclusive to neurons, alpha-synuclein is now known to be significantly present in erythroid cells according to mRNA findings, with its protein components identifiable across all blood elements, including packed red blood cells, PBMCs, platelets, and plasma. Individuals carrying *SNCA* triplications exhibit alpha-synuclein levels in their whole blood that are roughly twice as high as those in family members without the mutation and in unrelated individuals without the mutation. However, it remains unclear whether there is a link between common *SNCA* polymorphisms that increase risk and levels of alpha-synuclein in the periphery (Mata et al. 2010). The exact nature of this association requires further investigation.

The demographic profile of the individuals with *SNCA* variations also provided interesting insights. The average age of recruitment for those carrying *SNCA* deletions was 52 years, with an age range of 41 to 56 years. For individuals with *SNCA* duplications, the average recruitment age was higher, at 61 years, and ranged from 49 to 69 years. Similarly, the average age for individuals with complex *SNCA* events was 61 years, spanning an age range of 42 to 69 years. Notably, two subjects in this group were not classified as European from the demographic information provided by the UK biobank, indicating some diversity in the genetic background of the study cohort.

Two pairs of individuals, based on their relatedness value (Proportion of Identity by Descent; PIHAT approximately 0.5), were identified as siblings. The reported birth years of each pair differed by about two years. The first pair, individuals #dup1 and #dup6, both had a 1.5Mb *SNCA* duplication. The second pair, individuals #del2 and #del5, both had a 0.8Mb *SNCA* deletion. The occurrence of the same type of *SNCA* variation in siblings is particularly noteworthy, as it suggests a possible genetic inheritance pattern and highlights the importance of family history in studies focused on genetics.

Conclusions and Discussion

Missense Mutations and Copy Number Gains in SNCA

In an effort to perform a comprehensive assessment of the UK Biobank cohort for pathogenic variants in *SNCA*, I focused on detecting known missense mutations that are often linked to PD. Surprisingly, none of the five missense mutations typically associated with PD and traditionally recognized as pathogenic, were identified among the cohort participants. Beyond these known mutations, our analysis revealed the presence of other *SNCA* missense mutations. These mutations, however, come with a degree of uncertainty regarding their role in PD. Some of these identified mutations have conflicting reports in the literature, with varying interpretations of their pathogenicity. Interestingly, despite the identification of these missense mutations in *SNCA*, none of the carriers were reported to have PD.

Among the individuals carrying these missense *SNCA* mutations, three reported a positive family history for PD. This familial link could suggest a genetic predisposition, but due to the low allele frequency of these variants in the population, it is challenging to robustly determine if their occurrence is higher than what might be expected by chance. This difficulty in establishing a clear statistical link between these variants and PD underscores the need for further research and larger sample sizes. In previous work, we suggested that the *SNCA* p.H50Q mutation is likely not pathogenic, despite existing functional evidence that might suggest otherwise (Blauwendraat et al. 2018).

I also identified six instances of *SNCA* copy number gains and another six of copy number losses. Additionally, we discovered 14 complex *SNCA* genetic events. Again, unexpectedly, none of these individuals with *SNCA* variations had been diagnosed with PD. Three of these individuals had been assigned ICD-10 codes for symptoms that are commonly recognized as prodromal. Despite this, they did not meet the established criteria for a prodromal PD diagnosis as outlined elsewhere (Heinzel et al. 2019). This finding suggests that while these individuals exhibit some early symptoms, they do not conclusively indicate the onset of PD or are classified as prodromal PD cases according to current diagnostic standards. Moreover, one subject with an *SNCA* deletion was diagnosed with schizophrenia, based on their ICD-10 code. Previous studies have suggested schizophrenia as a potential prodromal symptom in individuals with *SNCA* duplication (Takamura et al. 2016). Regarding genetic predispositions, this finding might raise the question of whether certain psychiatric conditions may be early indicators of neurodegenerative diseases like PD.

It is well-established that *SNCA* copy number gains are a clear cause of autosomal dominant PD. However, the role of SNCA copy number losses in PD is less clear, as they are not typically reported to cause the disease and are likely protective. The absence of a PD diagnosis in these subjects with *SNCA* variations could indicate a possibility of later development of PD. This is particularly relevant considering that the average age of onset for PD in duplication carriers is around 46.9 years, with an age range from 30 to 73 years, as reported in a recent meta-analysis (Book et al. 2018). Therefore, it is possible that some of these individuals might develop PD later in life, beyond the age at which they were assessed in our study.

Mosaic Events and Their Implications

In our investigation, we uncovered 14 potential mosaic events related to *SNCA*. Out of these, 11 events were classified as large complex events. These complex events were notable for their varying levels of mosaicism, indicating a diversity in the extent of genetic variation present within the individuals studied, aligning with previous studies which discussed the implications of such large-scale mosaic events (Conlin et al. 2010). Two of the 14 identified mosaic events were characterized as likely being large mosaic deletions, each spanning between 20 to 25Mb. These sizable deletions suggest a significant alteration in the genetic structure within the affected individuals, potentially impacting either the function or expression of *SNCA*. One of the mosaic events appeared to be the result of a uniparental disomy event and seemed to affect a major portion of the long arm of chromosome 4, where *SNCA* is located.

Previous studies have demonstrated a link between *SNCA* mosaicism and PD as well as other synucleinopathies, providing evidence supporting the role of *SNCA* mosaicism in these neurodegenerative conditions in brains (Perez-Rodriguez et al. 2019; Mokretar et al. 2018; Perandones et al. 2014). This evidence highlights the importance of understanding mosaicism and its implications in the pathogenesis and progression of PD and related disorders.

Four individuals among the 14 identified with potential mosaic *SNCA* events, constituting approximately 29% of the group, had a history of blood-based cancers, such as lymphoma and multiple myeloma. This significant proportion of individuals with both mosaic *SNCA* events and a history of hematologic malignancies raises questions about a possible link between these genetic alterations and cancer predisposition. However, a key uncertainty in this observation is whether the mosaic events observed in blood-derived DNA are also mirrored in tissues more directly related to PD, such as brain tissue and neurons. Since *SNCA* variations and the resultant expression of alpha-synuclein protein are critical factors in the development of PD, understanding the presence and impact of these mosaic events in PD-relevant tissues is crucial. Moreover, the exact effect of these mosaic events on the expression of *SNCA* and the alpha-synuclein protein is not fully understood. While these genetic alterations in blood cells are evident, their influence on gene and protein expression within the central nervous system, and consequently their role in PD pathogenesis, remains a subject of ongoing research.

Limitations and Future Directions

Multiplex ligation-dependent probe amplification (MLPA) is recognized as the gold standard technique for detecting CNVs. This method offers high sensitivity and specificity in identifying variations in the number of copies of a particular gene segment. However, a primary challenge in this study was the lack of immediate access to the DNA samples of the subjects with potential *SNCA* variation. This limitation hindered our ability to use MLPA for direct validation of the identified CNVs.

To circumvent this limitation, we used available exome sequencing data to partially validate two of the genomic events identified. Exome sequencing, while not as targeted for CNV detection as MLPA, still provides valuable insights into the genetic composition of an individual and AD information collected as a part of exome sequencing can be used to infer CNVs to some extent. Through this approach, we were able to confirm the presence of these two specific genomic events, lending support to our initial findings based on genotyping array data.

Nevertheless, the partial nature of this validation highlights the need for more comprehensive future efforts to fully understand these genetic variations. This is particularly important for the complex mosaic events we identified. Such events, involving variations in the genetic makeup across different cells of the same individual, present a more complex scenario that requires follow-up investigation.

Despite the UK Biobank being an extensive resource, it is not immune to certain limitations, particularly regarding the completeness of phenotype data. The primary source of this limitation stems from the reliance on electronic health records (EHRs) and the use of ICD-10 coding. EHRs, while comprehensive,

may not capture every aspect of a patient's health history, especially more subtle symptoms or conditions that are not routinely coded. In the context of PD and its prodromal stages, this limitation is significant. Prodromal PD often presents with non-specific symptoms that might not be adequately captured or coded in EHRs, leading to potential underreporting or misclassification of PD cases in our study.

Additionally, while our prioritzation methods are scalable and rapid, our methodological approach incorporated manual interpretation of visualized genetic data. While this allows for a nuanced understanding of the data, it also poses a risk of missing certain types of genetic variations. Smaller deletions or complex genomic rearrangements, which might be less apparent in the data, could be overlooked. These smaller or more intricate genetic alterations can be crucial in understanding the full genetic landscape of diseases like PD. The potential for missing smaller deletions or complex rearrangements is a notable concern, as these genetic changes could play a significant role in disease pathogenesis or progression. This gap highlights the need for a more refined analytical approach, possibly involving more advanced genetic analysis techniques, to ensure a comprehensive understanding of the genetic variations present in the UK Biobank cohort.

The results of this study highlight the occurrence of *SNCA* CNVs and mosaicism among individuals in the general population who do not exhibit symptoms of PD. Here, I point to the existence of genetic alterations typically associated with PD in individuals without any explicit reported clinical manifestation of the disease. These individuals, who carry *SNCA* CNVs or display mosaicism but remain free of PD symptoms, present a unique opportunity for further scientific exploration, aimed at understanding the mechanisms by which *SNCA* variation contribute to PD. Investigating why some individuals with these genetic alterations do not develop PD could provide insights into potential protective factors or mechanisms.

Moreover, these individuals are potential candidates for inclusion in clinical trials. Research could focus on monitoring these individuals over time to observe if and when they might develop PD symptoms, or exploring therapeutic interventions aimed at preventing the onset of PD in those with a particular genetic predisposition.

Additionally, the role of mosaicism in the *SNCA* region, especially as observed in blood-derived genotype data, warrants further investigation to understand how moscaism relates to disease development. Broader case-control studies are necessary to determine the association of *SNCA* mosaicism with PD or other neurodegenerative diseases. Such studies should aim to compare individuals with and without PD symptoms, focusing on the prevalence and patterns of mosaicism and their potential role in disease pathogenesis. This approach will help clarify whether these genetic events are mere bystanders or active contributors to the disease process.

Chapter 3: Large-scale Rare Variant Burden Testing in Parkinson's Disease

Overview and Broader Relevance

While genome-wide association studies have already identified over 90 loci with common variants associated with PD, offering valuable insights into the disease's biology, there has been a notable lack of large-scale analyses focusing specifically on rare genetic variants.

To address this, this study leverages extensive whole genome and exome sequencing data, examining a broad cohort that includes individuals with PD, proxy-cases, and healthy controls. The objective of this study is to conduct thorough analyses on these participants' genetics and how these less common genetic variations might influence the risk of developing PD. I revisit and explore both established and potentially new genetic contributors to PD. This includes genes such as *GBA1* and *LRRK2*, where mutations within these genes have been previously linked to PD risk, and extends to probing variants in other genes potentially implicated in the disease. I conduct burden tests to study the potential impact of specific genetic changes, like small insertions/deletions (indels) and single nucleotide variations (SNVs) that alter protein function, chosen for their anticipated functional significance. In brief, burden tests collectively assess the individual variants that may be rare or impart a small effect by aggregating them and assessing their association with PD (in this study, by variant class restricted to a gene).

Here, I conduct the first large-scale, extensive investigation at a genome-wide scale to explore rare genetic variants associated with PD risk, focusing specifically on those with minor allele frequencies (MAF) below 1%. This analysis delves deeply into the rare genetic elements of PD, aiming to uncover their potential contributions to the susceptibility of disease risk.

This work has been published here:

> **Makarious MB**, Lake J, Pitz V, Fu AY, …, Beach TG, Serrano GA, Real R, Morris HR, Ding J, Gibbs RG, Singleton AB, Nalls MA, Bhangale T, Blauwendraat C: "**Large-scale Rare Variant Burden Testing in Parkinson's Disease**" *Brain (2023);* https://doi.org/10.1093/brain/awad214

Introduction

PD emerges from a complex interplay of aging, environmental influences, and genetic factors. The study of common genetic variants in PD has been extensively pursued through large genome-wide association studies (GWAS), revealing significant insights. These studies have identified 92 independent risk signals linked to PD, encompassing common variants near genes such as *SNCA*, *TMEM175*, and *MAPT* (Nalls et al. 2019a; Foo et al. 2020a). Typically found in non-coding regions, these array-based GWAS risk alleles usually have a frequency above 5% in the target population and tend to exert relatively small effects.

In contrast, the exploration of rare, more damaging variants, often implicated in familial and sporadic forms of PD, has primarily employed family-based methods. Notable examples include coding variants in genes like *SNCA* (M. H. Polymeropoulos et al. 1997) and *PRKN (Kitada et al. 1998)*. Another intriguing aspect in PD genetics is the diversity observed in pleomorphic genes, where multiple variants across a spectrum of allele frequencies exhibit a wide array of effect sizes (A. Singleton and Hardy 2011). This phenomenon is exemplified in PD, where GWAS have pinpointed common variants near genes such as *GBA1*, *GCH1*, *LRRK2*, *SNCA*, and *VPS13C* with moderate effects (Nalls et al. 2019a), while familial studies have identified rare variants in these same genes with more substantial impacts, such as *GBA1* p.N370S , *LRRK2* p.G2019S, and *SNCA* p.A53T (Jansen et al. 2017; Gaare et al. 2020; Rudakou et al. 2021; Mencacci et al. 2014).

However, the landscape of rare variants in PD has not been thoroughly investigated on a genome-wide scale. While candidate gene studies have reported rare variant associations in PD-related genes like *ARSA* and *ATP10B* (J. S. Lee et al. 2019; S. Martin et al. 2020), these findings are subject to debate due to inconsistent replication across independent PD datasets (Makarious et al. 2019; Fan et al. 2020; Tesson et al. 2020; Real et al. 2020). A significant challenge in rare variant analysis is the diminishing quality and reliability of imputation methods as allele frequencies decrease. As most large-scale GWAS datasets still predominantly rely on more cost-effective genome-wide genotyping methods over sequencing, they often use imputed genotype data.

This study addresses these challenges by analyzing whole genome (WGS) and whole exome sequencing (WES), which are more adept at analyzing rare variants. We conduct the most extensive genome-wide analysis of rare variants in PD to date, encompassing 7,184 PD cases, 6,701 proxy cases (individuals with a parent or sibling diagnosed with PD), and 51,650 neurologically healthy controls of European ancestry from various large-scale sequencing projects. Through gene-level burden testing, we aim to decipher the contributions of moderate- to large-effect rare variants to the genetic landscape of PD, thus enhancing our understanding of its genetic etiology.

Methods

Access to the Accelerating Medicines Partnership in Parkinson's Disease (AMP-PD) data and quality control notebooks requires individual registration at [https://amp-pd.org/]. UK Biobank data also necessitate application for access via [https://www.ukbiobank.ac.uk/]. Other cohort data were sourced through collaborations with the National Institutes of Health (NIH) and Genentech, with all studies adhering to their respective institutional ethics guidelines and obtaining informed consent from participants. All generated data and full results available at [https://github.com/neurogenetics/PD-BURDEN]. NABEC data can be accessed through NCBI dbGaP, study accession phs001300.v2.p1.

Cohorts

AMP-PD and NIH Clinic

AMP-PD is a collaborative initiative designed to consolidate data from various cohorts into a unified research platform dedicated to PD research. This platform unifies data storage and analysis from several cohort studies, encompassing clinical metrics, whole genomes, whole blood RNA sequencing transcriptomics, and proteomics data for participants.

Version 2.5 of the AMP-PD release includes whole genome sequencing data from approximately 10,000 participants across 8 cohorts. These samples were subjected to joint genotyping using the TOPMed Freeze 9 Variant Calling Pipeline. The cohorts contributing to this release include BioFIND, Harvard Biomarker Study (HBS), NINDS Lewy Body Dementia (LBD), LRRK2 Cohort Consortium (LCC), NINDS Parkinson's Disease Biomarkers Program (PDBP), Parkinson's Progression Markers Initiative (PPMI), the randomized Phase 3, 2-arm, double-blind, parallel-group trial STEADY-PD3, and the randomized, double-blind, placebo-controlled trial of urate-elevating inosine treatment named the "Study of URate Elevation in Parkinson's Disease, phase 3" trial (SURE-PD3). Detailed information about each cohort's recruitment methodology is available on the AMP-PD website (https://amp-pd.org/).

In addition to the AMP-PD datasets, parallel sequencing efforts were conducted at the Laboratory of Neurogenetics (LNG) and the U.S. Uniformed Services University (USHUS). These included samples from the NIH PD clinic, the United Kingdom Brain Expression Consortium (UKBEC; Trabzuni, United Kingdom Brain Expression Consortium (UKBEC), and Thomson 2014), the North American Brain Expression Consortium (NABEC; (Gibbs et al. 2010), and the Wellderly study (Erikson et al. 2016). The AMP-PD cohorts (PPMI, PDBP, HBS, BioFIND, SURE-PD3, and STEADY-PD3) underwent processing in accordance with the Genome Analysis Toolkit (GATK) Best Practices guidelines by the Broad Institute's joint discovery pipeline (Poplin et al. 2018).

Other cohorts were processed separately but similarly, adhering to the GATK Best Practices and Broad Institute's workflow for joint discovery and Variant Quality Score Recalibration (VQSR). The data processing and QC protocols are detailed in previous studies (S. Bandres-Ciga et al. 2020; Iwaki et al. 2021). These sequencing efforts achieved median/mean coverages ranging between 33.3x and 35.0x. Additional QC included removing closely related individuals (with PIHAT >0.125, removing second-degree relatives or closer) using PLINK software (v1.9; Purcell et al. 2007). All participants were of European ancestry, as established by principal component analysis against HapMap3 European ancestry populations. Biased or genetically targeted datasets, such as *LRRK2* and *GBA1* variant carriers in specific PPMI efforts, were excluded from this analysis. A minimum allele count (MAC) threshold of 1 was set for all variants within gene boundaries. Exonic regions were isolated from the whole genome sequencing data, using exome calling regions from gnomAD mapped to the hg38 reference genome (Karczewski et al. 2020). The AMP-PD and NIH datasets were merged prior to gene burden analysis, with 3,848 duplicated samples removed prior to analysis.

UK Biobank

The UK Biobank is a population-based prospective study tracking around 500,000 individuals based in the United Kingdom, where the data is available upon application. Made available to researchers are multiple types of data for the recruited individuals including phenotypes, genotypes, imaging, and physical activity monitoring data.

To investigate PD, exome sequencing data from a total of 200,643 individuals (OQFE dataset, field codes: 23151 and 23155) were downloaded from the UK Biobank (Bycroft et al. 2018) in December of 2020. Standard quality control was performed to exclude non-European outliers. Closely related individuals (PIHAT >0.125) were excluded by selecting one sample at random using PLINK (v1.9; Purcell et al. 2007). Standard exome sequencing data filtering, performed by the UK Biobank Exome Sequencing Consortium, sequenced these exomes with 95.8% of targeted bases covered at a depth of 20x or higher, described in previous UK biobank exome sequencing studies (Backman et al. 2021).

UK Biobank phenotype data were obtained from ICD-10 codes (field code: 41270), PD (field code: 131023), illnesses of father and mother (field codes: 20107 and 20110), parkinsonism (field code: 42031) or dementia (field code: 42018), genetic ethnic grouping (field code: 22006), year of birth (field code: 34) and age of recruitment (field code: 21022). Cases were defined as any individual identified as having PD using the above field code. Proxy-cases were defined as having a parent or sibling with PD as previously reported (Nalls et al. 2019a). Controls were filtered to exclude any individuals with an age of recruitment <59 years, any reported nervous system disorders (Category 2406), a parent with PD or dementia (field codes: 20107 and 20110), and any of the following reported neurological disorders: (disorder/field code): Dementia/42018, Vascular dementia/42022, frontotemporal dementia (FTD)/42024, amyotrophic lateral sclerosis (ALS)/42028, Parkinsonism/42030, PD/42032, progressive supranuclear palsy (PSP)/42034, or MSA/42036.

Genentech

In a comprehensive whole-genome sequencing effort, Genentech contributed data comprising 2,710 PD cases and 8,994 control individuals. PD cases included 2,318 individuals from 23andMe, a subset of those included in the analysis by Chang and colleagues (D. Chang et al. 2017) who were contacted and provided consent for this analysis. Additionally, 392 PD cases were incorporated from the Roche clinical trial of tolcapone (TASMAR), a now-approved oral medication designed to be used alongside carbidopa and levodopa, aimed at alleviating the symptoms of PD. For the control group, individuals were selected from various Genentech clinical trials and studies. These control subjects were diagnosed with diseases that do not exhibit significant genetic correlation with PD. The diseases include age-related macular degeneration (AMD; n=1,735 individuals), asthma (n=3,398 individuals), idiopathic pulmonary fibrosis (IPF; n=1,532 individuals), and rheumatoid arthritis (RA; n=2,329 individuals).

The whole-genome sequencing was conducted using the Illumina HiSeq platform. This platform facilitated 30x coverage of the genome with 150bp paired-end reads. The sequencing data was of high fidelity, with genotypes having a genotype quality (GQ) score lower than 20 marked as missing. These reads were aligned to the GRCh38 human reference genome using the Burrows-Wheeler Aligner (BWA)

(H. Li and Durbin 2009). Post-alignment, the GATK Best Practices pipeline was employed for essential processes like base quality score recalibration, indel realignment, and duplicate removal (H. Li and Durbin 2009; McKenna et al. 2010). Subsequently, SNP and indel discovery, along with genotyping, were executed across all samples concurrently, adhering to the Variant Quality Score Recalibration (VQSR) as recommended in the GATK Best Practices (DePristo et al. 2011; Van der Auwera et al. 2013; Van der Auwera and O'Connor 2020).

The dataset, comprising 11,704 samples, met rigorous QC criteria. These criteria included a genotype missing rate below 0.1, ensuring no sample pair had a kinship coefficient (k0, representing the probability of zero alleles shared IBD or the value Z0 as reported by the PLINK `--genome` module) less than 0.4. Additionally, all samples were validated through five iterations of outlier removal using principal component analysis (PCA; (Price et al. 2006).

Variant Annotation

Variants in this study were annotated using the SnpEff and SnpSift annotation tools (version 4.3t; (Cingolani et al. 2012) along with the Ensembl Variant Effect Predictor (VEP; version 104; McLaren et al. 2016). We also used the Combined Annotation Dependent Depletion (CADD; version 1.4; Rentzsch et al. 2019) and the Loss-of-Function Transcript Effect Estimator (LOFTEE; version 1.02; Karczewski et al. 2020) as VEP plugins. For gene burden analysis in this study, I classified the variants into the following variant classes: 1) missense variants as identified by SnpEff, 2) variants with moderate or high impact as determined by SnpEff/SnpSift, 3) high confidence LoF variants as classified by LOFTEE, and 4) variants possessing a CADD PHRED score greater than 20 or classified as high confidence LoF by LOFTEE.

SnpEff

SnpEff is a tool is designed for annotating genetic variants found in sequencing data and predicting their functional impact. Using a comprehensive database that encompasses approximately 38,000 genomes, SnpEff can process variants across a broad range of organisms, providing valuable insights into their potential biological significance. The annotation process involves categorizing variants based on their locations (such as exonic, intronic, intergenic, etc.) and types (like missense, nonsense, synonymous, etc.). SnpEff also predicts the functional impacts of these variants, such as whether they might alter protein function or gene expression. This prediction is crucial for understanding how variations in DNA sequence contribute to phenotypic differences, disease susceptibility, and drug response.

SnpSift

Working in conjunction with SnpEff, SnpSift offers advanced filtering and prioritization capabilities for genetic variants. It integrates data from multiple databases, enhancing the depth and breadth of variant analysis. SnpSift can filter variants based on various criteria, such as their frequency in the population, predicted impact, or association with specific diseases or traits. One of the key features of SnpSift is its ability to predict amino acid changes resulting from non-synonymous variants and to assess the potential impacts of these changes. These impacts are often classified into categories such as "moderate" or "high," depending on the predicted degree of alteration in protein function or structure.

CADD

CADD PHRED was developed to assess the potential pathogenicity of SNVs and small indels within the human genome. Its significance stems from the growing need to interpret the enormous volume of variant data generated by modern high-throughput sequencing technologies.

CADD has a scoring system, conceptualized in a PHRED-like format. This format, familiar to genomic researchers due to its application in quantifying the quality of base calls in sequencing, is repurposed in CADD to quantify the deleteriousness of genetic variants. These scores are expressed in PHRED-scaled units, which essentially log-transform raw scores to a more interpretable scale. Higher CADD scores indicate a greater likelihood that a variant is deleterious and could potentially impact biological function, indicating that a score of 20 places the variant in the top 1% of harmful variants in the human genome.

CADD integrates a wide array of genomic features, including evolutionary conservation, functional annotations, and other sequence-based characteristics, to assess each variant. This integration takes into account both the direct impact of a variant on protein structure and function, as well as more subtle effects that might influence gene regulation or transcript stability. This assists in prioritizing variants for further study, especially in large-scale genomic projects. Additionally, there is ongoing refinement and updates to CADD, leveraging deep learning networks and splicing information. CADD continues to incorporate new data and improved algorithms, ensuring its relevance and accuracy.

LOFTEE

LOFTEE's main function is to assess and categorize genetic variants that potentially lead to the loss of function in genes. LOFTEE specifically targets mutations such as stop-gain (nonsense), frameshift, and disruptions in splice sites. These types of mutations are known for their capacity to drastically alter gene function, making the accurate identification and categorization of these variants important in research. The types of variants analyzed by LOFTEE include stop-gain mutations, point mutations causing premature termination of protein translation, often resulting in truncated and usually non-functional proteins. Frameshift mutations, caused by insertions or deletions, change the reading frame of a gene, typically producing different and non-functional protein products. Additionally, variants that disrupt normal splice sites can significantly affect pre-mRNA splicing, potentially leading to aberrant or incomplete proteins (Vihinen 2023).

A key feature of LOFTEE is its ability to classify these LoF variants into two categories: low confidence and high confidence. High confidence LoF variants are those with strong evidence suggesting a true loss of gene function, supported by factors like the evolutionary conservation of the affected region, available functional data, or alignment with known disease phenotypes. In contrast, low confidence LoF variants are those where the evidence is less definitive or where conflicting data exist regarding their impact. These variants require further experimental validation to fully understand their effects (Karczewski et al. 2020).

Gene Burden Tests

The Combined Multivariate and Collapsing (CMC) Wald and Optimal Sequence Kernel Association Test (SKAT-O) tests are statistical methods to analyze the impact of rare variants. These methods are designed to improve the detection of associations between genetic variants and complex diseases, especially when individual variants might have subtle effects that are challenging to detect with conventional approaches (Seunggeung Lee et al. 2014).

CMC Wald

The CMC Wald tests are used for gene-based association studies, and are particularly effective for analyzing rare variants. This method combines multiple rare variants within a gene into a single analytical framework, increasing the power to detect associations that might be missed when examining variants individually. The “collapsing” component of this test involves grouping rare variants based on specific criteria like their location within a gene or predicted functional impact. The “multivariate” aspect refers to the statistical technique used to assess the collective effect of these variants on the trait of interest (Seunggeun Lee, Wu, and Lin 2012; Seunggeun Lee et al. 2016). This approach is necessary for studying rare genetic variants, which might individually occur at a low frequency in the population but collectively exert a significant influence on disease risk.

The benefit of the CMC Wald test is its enhanced power in detecting associations with rare variants by considering their cumulative effect within a gene. However, it can suffer from reduced specificity if the collapsing includes variants not associated with the trait, and it often assumes that grouped variants exert a similar effect on the trait. CMC Wald tests also assume that the variants within the collapsed group are independent of each other, which does not take into consideration regions of linkage disequilibrium (LD) (Seunggeung Lee et al. 2014).

SKAT-O

SKAT-O is a versatile and robust statistical method used to test the association of a set of variants with a phenotype, especially useful in the context of rare variants and complex traits. It integrates elements of the SKAT (Sequence Kernel Association Test) and burden tests, offering a flexible approach that adapts to different genetic architectures. SKAT-O can assess both the aggregate effect of multiple rare variants and identify individual variants with more substantial effects. The key advantage of SKAT-O is its adaptability, maintaining high statistical power across various genetic effect scenarios, from many variants with small effects to fewer variants with larger effects. SKAT-O has demonstrated reliability in gene detection compared to either the burden or SKAT tests used independently. This enhanced effectiveness is achieved through its adaptive selection of the optimal linear combination of both SKAT and burden tests, which is designed to maximize the test's power (Seunggeun Lee, Emond, Bamshad, Barnes, Rieder, Nickerson, Christiani, et al. 2012). However, it requires careful selection of the variants for analysis and can be computationally intensive, as because it is a mixed effect model, assumes a normal distribution and that the genetic variants have random, not fixed, effects.

SKAT-O, compared to CMC Wald, is better suited for complex traits as it can account for rare and common variants, and especially when there is heterogeneity in genetic effects (Seunggeun Lee, Emond, Bamshad, Barnes, Rieder, Nickerson, NHLBI GO Exome Sequencing Project—ESP Lung Project Team, et al. 2012). It provides flexibility in handling various genetic effect sizes and directions, but it might be less effective when the effect sizes are very uniform. SKAT-O is also limited when handling sparse alternatives, referring to the scenarios where only a small proportion of variants in a set are actually associated with a disease or trait. It is important to note that sparse alternatives are quite common in sequencing studies since many variants in a set are typically not influential for the risk or related traits of a disease (Liu et al. 2019). In this study, I attempt to mitigate this by analyzing particular variant classes that have a higher probability of being associated with disease risk.

Gene Burden Analysis and Meta-analyses

Analyses

All analyses were stratified by the four variant classes described above and by maximum control MAF levels of 1% and 0.1%. These algorithms were run using the RVtests package (v2.1.0; Zeggini and Morris 2015) for the UK Biobank, and combined AMP-PD and NIH dataset. For Genentech data, SKAT-O and CMC-Wald tests were performed using the R package SKAT (Wu et al. 2011). Unless otherwise stated, all results reported in this manuscript correspond to the SKAT-O rare variant test, and all meta-analyses were performed using the combined p-values reported following Fisher's test.

The AMP-PD and NIH dataset were adjusted for the following factors: sex, age, and the initial five principal components. In contrast, the UK Biobank datasets were calibrated for sex, Townsend scores, and the first five principal components. For the analysis involving the UK Biobank, only neurologically healthy individuals aged 60 and above were considered, thus excluding age as a variable in the analysis.

Analyses of rare variants were conducted separately for each dataset, all using the hg38 genome build. Two comprehensive meta-analyses were carried out as follows: firstly, a case-control meta-analysis involving the integrated AMP-PD and NIH dataset, the Genentech dataset, and the UK Biobank case-control dataset; secondly, a meta-analysis combining case-control and proxy-control results from the amalgamated AMP-PD and NIH dataset, the Genentech dataset, the UK Biobank PD case-control dataset, and both the sibling and parent proxy-cases datasets from the UK Biobank. A detailed description of this analytical process is presented in **Figure 5**. A visualized distribution of the PCs per dataset and age distribution within datasets can be found in **Figure 6 and Figure 7.**

The population substructure evident in **Figure 6** likely reflects varying ancestry among cohort participants, particularly in the American cohorts (AMP-PD, NIH, and Genentech), which have a higher proportion of individuals with AJ ancestry. This clustering aligns with the known genetic differences between AJ and broader European populations. In my analysis, I did not further divide European and AJ ancestry; however, we confirmed European ancestry for all participants through PCA against reference populations from HapMap3. HapMap3 provides a global reference panel that includes individuals from well-characterized populations, such as Europeans (CEU), Africans (YRI), and Asians (CHB and JPT).

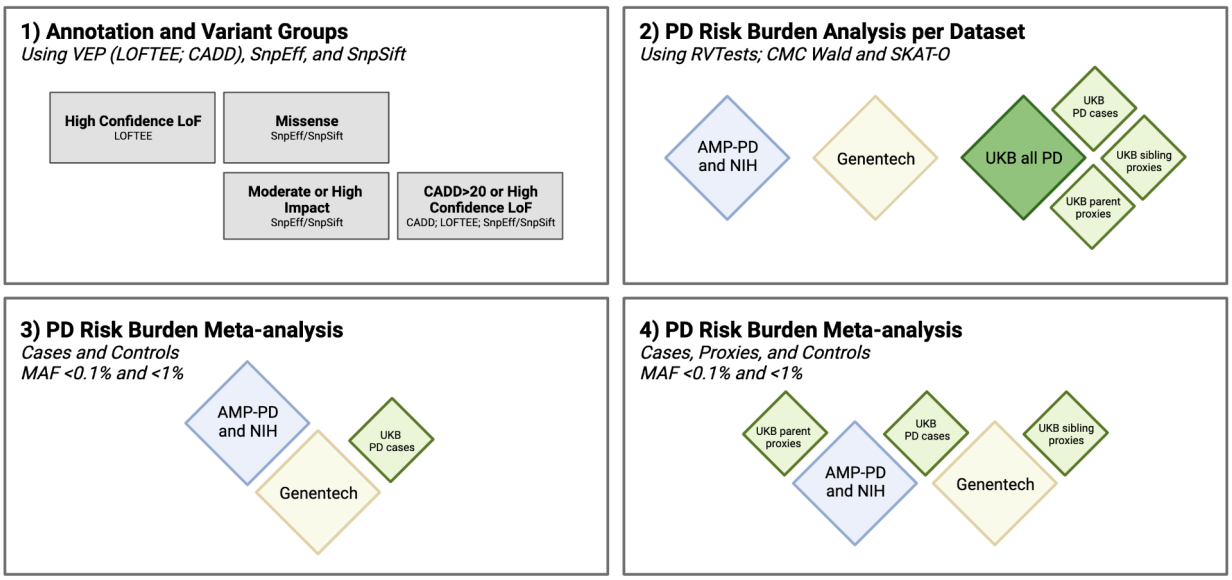


Figure 5: Graphical representation of the analytical process for conducting large-scale rare variant burden testing in Parkinson's Disease

(1) Variant annotation was executed through the Variant Effect Predictor (VEP), leading to the categorization of variants into four distinct groups: (i) missense variants, as identified by SnpEff; (ii) variants with moderate or high impact according to SnpEff/SnpSift definitions; (iii) high confidence loss-of-function (LoF) variants, as determined by LOFTEE; and (iv) variants that either possess a CADD PHRED score greater than 20 or are high confidence LoF variants, as classified by LOFTEE. (2) For each dataset, burden analysis was carried out independently, focusing on variants classified as rare (minor allele frequency, MAF, less than 1%) and ultra-rare (MAF below 0.1%). (3) The first meta-analysis approach, referred to as the 'case-control' meta-analysis, involved solely Parkinson's disease (PD) cases and control subjects. (4) The second meta-analysis strategy, known as the 'case-control-proxies' meta-analysis, incorporated PD cases, PD proxy cases (including siblings and parents), and control subjects.
AMP-PD: Accelerating Medicines Partnership in Parkinson's Disease; NIH: National Institutes of Health; PD: Parkinson's disease; LoF: Loss-of-Function; LOFTEE: Loss-of-Function Transcript Effect Estimator; VEP: Variant Effect Predictor; CADD: Combined Annotation Dependent Depletion; MAF: Minor allele frequency

Meta-analysis

The combined P-Value approach using Fisher's test aggregates p-values from different studies to ascertain an overall effect. Fisher's test, a classic technique for combining p-values, operates under the null hypothesis that all individual studies are independent and exhibit no effect. The method involves summing the natural logarithms of the p-values from each study, and then this sum is multiplied by -2. The resultant statistic adheres to a chi-square distribution with degrees of freedom equal to twice the number of studies. Notably, Fisher's method has been shown to be effective in detecting over 75% of causal effects (whether deleterious or protective), particularly when these effects align in direction across studies (Derkach, Lawless, and Sun 2013).

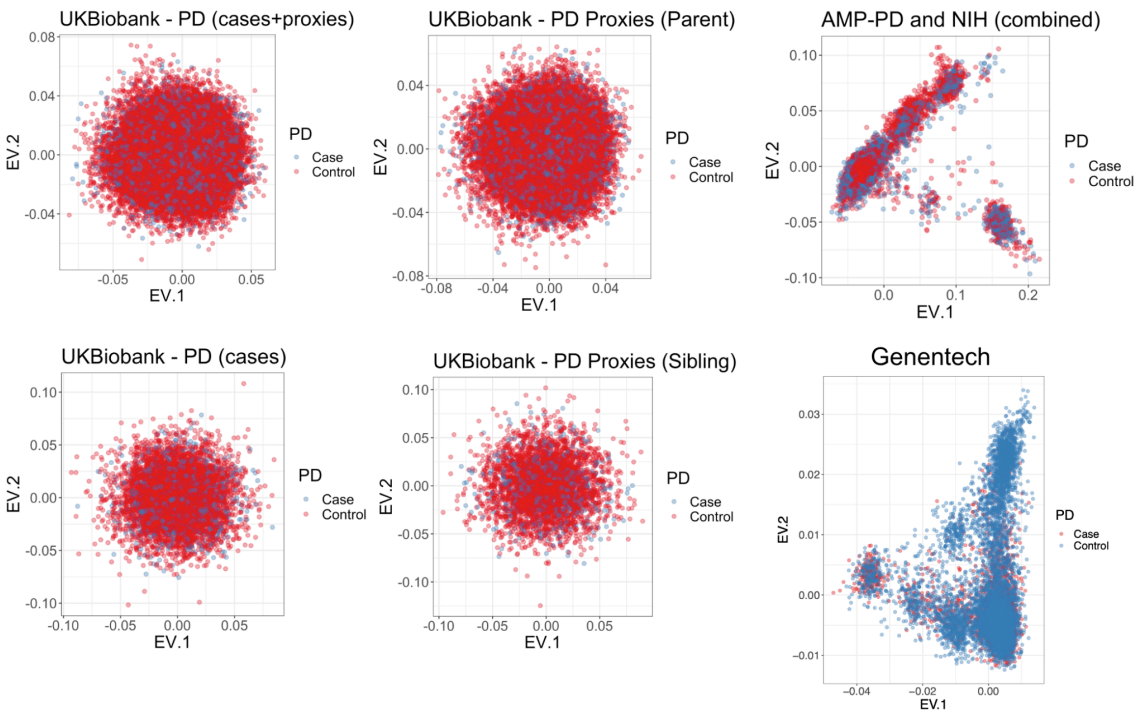


Figure 6: Principal component plots per dataset included in large-scale rare variant burden meta-analyses

AMP-PD: Accelerating Medicines Partnership in Parkinson's Disease; NIH: National Institutes of Health; PD: Parkinson's disease; EV: Eigenvalue.

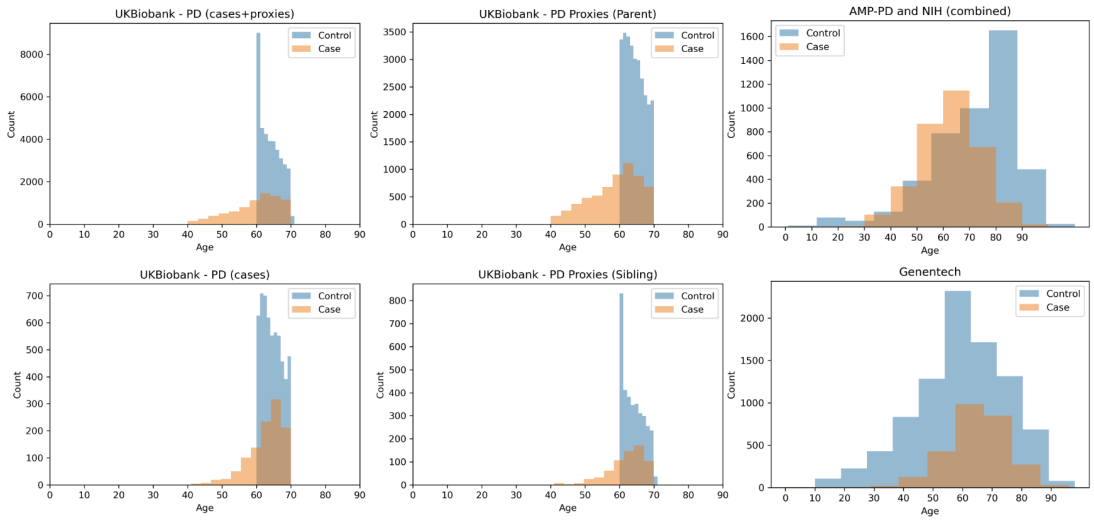


Figure 7: Age distribution per dataset included in large-scale rare variant burden meta-analyses

AMP-PD: Accelerating Medicines Partnership in Parkinson's Disease; NIH: National Institutes of Health; PD: Parkinson's disease

Power Calculations

In this study, I conducted 100 gene simulations using the power calculation feature of the SKAT R package (version 2.0.1; Seunggeun Lee, Emond, Bamshad, Barnes, Rieder, Nickerson, Christiani, et al. 2012), with European haplotypes as the default setting. The simulations projected a total sample size of 65,535, which included 7,184 PD cases, 6,701 proxy cases with a weighting equivalent to one-fourth of a PD case (effectively 1,675 cases), and 51,650 control participants. This configuration resulted in a case ratio of 13.5%. PD prevalence was assumed to be 1%, as previously reported (Tysnes and Storstein 2017).

For exome-wide significance, I applied a Bonferroni-corrected threshold of 2.50E-6, considering an estimated 20,000 protein-coding genes. With the application of two distinct burden testing algorithms, we established a final significance threshold at 1E-6. The power analyses, which were varied based on causality percentages (at 10%, 5%, 3%, 1%, and 0.5%) and MAF of causal variants (at 0.05%, 0.1%, 0.5%, 1%, 3%, and 5%), are detailed in **Table 5**. The analyses indicate that with a minimum assumption of 3% causal rare alleles, there is at least 80% power to detect associations at the specified MAF thresholds (**Table 5**).

		Causal Minor Allele Frequency (MAF; %)					
		0.05	**0.1**	**0.5**	**1**	**3**	**5**
Causal Percentage (%)	**10**	100	100	100	100	100	100
	5	75.483	99.999	100	100	100	100
	3	27.853	89.383	99.989	100	100	100
	1	3.867	14.843	49.704	71.429	96.362	96.594
	0.5	0.001	4.462	21.325	35.344	67.970	74.767

Table 5: Power calculations at various causal percentages and MAFs (given 1% disease prevalence and alpha=1E-6)

MAF: Minor allele frequency

	Sample Size		Age^ (Mean ± SD)		Sex (Male; %)	
Dataset	**Cases**	**Controls**	**Cases**	**Controls**	**Cases sex (Male; %)**	**Controls sex (Male; %)**
AMP-PD and NIH Genomes (Includes: PPMI, PDBP, HBS, BioFIND, NIH PD clinic, UKBEC, NABEC, Wellderly)	3369	4605	62.1 (11.8)	71.9 (16.2)	63.6	47.6
UKB case-control (WES)	1105	5643	62.9 (5.24)	64.1 (2.84)	62.4	47.6
UKB sibling proxy-control (WES)	668*	3463	62.2 (5.59)	64.1 (2.83)	45.5	49.5
UKB parent proxy-control (WES)	6033*	28945	58.1 (7.23)	64.1 (2.82)	42.5	48.7
Genentech case-control (WGS)	2710	8994	64.7 (10.4)	59.2 (15.6)	59.2	40.7
Total	**7184 cases; 6701 proxies**	**51,650 controls**				

Table 6: Datasets included in rare variant burden analysis after quality control

AMP-PD: Accelerating Medicines Partnership Parkinson's disease; NIH: National Institutes of Health; PPMI: Parkinson's Progression Markers Initiative; PDBP: Parkinson's disease Biomarkers Project; HBS: Harvard Biomarker Study; UKBEC: UK Brain Expression Consortium; NABEC: North American Brain Expression Consortium
WES: Whole exome sequences; WGS: whole genome sequences
* indicates proxy cases
^age for AMP-PD and NIH datasets reported at recruitment or baseline, ages reported for UKBiobank datasets at recruitment, ages reported for Genentech at recruitment

Results

Study Overview

In this study, I curated substantial datasets comprising 7,184 PD cases, along with 6,701 sibling or parent proxies, and a control group of 51,650 individuals. The data were sourced from whole genome sequencing (from AMP-PD, NIH, and Genentech) and exome sequencing (from the UK Biobank; **Table 6**). To assess the genetic burden of rare variants, gene-level burden tests were conducted across all genes. This analysis was segmented into four distinct classes of variants and applied two different MAF thresholds for determining causality (overview in **Figure 5**).

One key observation was the inverse relationship between the deleteriousness of a variant class and the number of variants identified per gene; more harmful variant classes tended to yield a smaller number of variants per gene. For a comprehensive understanding of the distribution and quantity of these variants within each gene among both cases and controls, and how these vary according to the class of variant and the specific cohort (with the exception of the Genentech cohort), see **Supplementary Tables 3, 4, 5, and 6.** These tables offer a detailed breakdown, stratified by variant class and cohort for every gene.

Exome-Wide Significance of Genetic Variants in Parkinson's Disease Case-Control Studies

In this study, I conducted gene burden analyses across various PD case-control datasets, including AMP-PD and NIH, Genentech, and UK Biobank cases, as well as sibling and parent proxies from the UK Biobank. This multi-dataset approach yielded significant findings known PD genes, notably *GBA1* and *LRRK2*, which demonstrated exome-wide significance under both MAF thresholds (P < 1E-6; **Table 7 and Table 8; Supplementary Tables 7 and 8**). Lambda values calculated for each dataset (λ_{1000}; **Supplementary Table 9**) indicated minimal genomic inflation, adjusted for the varying numbers of cases, proxy-cases, and controls. Notably, smaller datasets, such as the UK Biobank sibling proxy-control dataset, exhibited more pronounced genomic deflation (λ_{1000} < 0.9) when analyzed independently.

VARIANT CLASS (MAF <1%)	GENE	CASE ONLY META PVAL	CASE PROXIES META PVAL	AMP NIH PVAL	GNE PVAL	UKB CASE PVAL	UKB SIBLING PVAL	UKB PARENT PVAL
Missense	*GBA1***	**3.27E-14**	**1.46E-21**	1.05E-05	**1.32E-08**	3.14E-04	0.247	**2.15E-10**
	*LRRK2**	**7.15E-07**	9.46E-06	**1.96E-07**	0.047	0.372	0.615	0.482
Moderate or High Impact	*GBA1***	**9.10E-15**	**1.32E-22**	1.05E-05	**5.70E-08**	1.89E-05	0.073	**2.15E-10**
	*LRRK2**	**7.23E-07**	9.85E-06	**2.09E-07**	0.040	0.413	0.584	0.527
	TUBA1B	0.69	9.02E-05	NA	0.647	0.501	0.352	**9.48E-07**
LOF	*B3GNT3***	**4.40E-09**	**3.36E-09**	NA	**4.40E-09**	NA	NA	0.032
	*CAPN10***	**3.60E-07**	**7.84E-07**	NA	0.005	3.75E-06	0.053	0.394
CADD>20 or LOF	*GBA1***	**3.72E-14**	**9.12E-22**	1.24E-05	**6.99E-08**	5.77E-05	0.130	**2.15E-10**
	*LRRK2**	**2.49E-07**	4.22E-06	**2.23E-07**	0.012	0.409	0.735	0.485
	ADH5	4.62E-06	6.15E-05	0.512	0.170	**3.13E-07**	0.491	0.768
	OR1G1	0.215	6.56E-06	0.848	0.029	0.620	**6.58E-07**	0.063

Table 7: Genes reaching exome-wide significance (p<1E-6) in MAF <1% in meta-analyses and individual datasets following SKAT-O

AMP NIH: Accelerating Medicines Partnership Parkinson's disease; NIH: National Institutes of Health; GNE: Genentech; UKB: UK Biobank

MAF: Minor allele frequency
*denotes genes that pass exome-wide significance (P<1E-6) in meta-analyses

VARIANT CLASS (MAF <0.1%)	GENE	CASE ONLY META PVAL	CASE PROXIES META PVAL	AMP NIH PVAL	GNE PVAL	UKB CASE PVAL	UKB SIBLING PVAL	UKB PARENT PVAL
Missense	*GBA1**	1.86E-05	**4.48E-12**	0.022	2.30E-02	2.55E-04	5.41E-03	**6.86E-08**
Moderate or High Impact	*GBA1**	1.71E-06	**4.87E-16**	NA	0.088	1.13E-06	0.001	**5.13E-10**
	*AUNIP***	**1.54E-08**	**2.70E-07**	NA	0.023	**3.04E-08**	0.170	1
	TUBA1B	0.690	9.02E-05	NA	0.647	0.501	0.352	**9.48E-07**
	*TREML1**	0.048	**3.58E-07**	NA	0.010	0.858	0.001	1.41E-05
LOF	*B3GNT3***	**4.40E-09**	**3.36E-09**	NA	**4.40E-09**	NA	NA	0.032
	*AUNIP***	**1.64E-08**	**2.04E-07**	NA	0.024	**3.13E-08**	0.116	1
	*CAPN10***	**3.60E-07**	**7.84E-07**	NA	0.005	3.75E-06	0.053	0.394
CADD>20 or LOF	*GBA1***	**2.33E-07**	**1.20E-14**	0.017	0.127	**4.56E-07**	8.93E-04	**7.89E-08**
	LRRK2	3.46E-06	2.65E-06	0.727	**6.15E-07**	0.044	0.771	0.014
	*AUNIP**	**2.12E-07**	1.53E-06	0.886	0.032	**3.15E-08**	0.125	1
	OR1G1	0.215	6.56E-06	0.848	0.029	0.620	**6.58E-07**	0.063

Table 8: Genes reaching exome-wide significance (p<1E-6) in MAF <0.1% in meta-analyses and individual datasets following SKAT-O

AMP NIH: Accelerating Medicines Partnership Parkinson's disease; NIH: National Institutes of Health; GNE: Genentech; UKB: UK Biobank
MAF: Minor allele frequency
*denotes genes that pass exome-wide significance (P<1E-6) in meta-analyses

Differential Impact of Variant Classes in PD Gene Burden Analysis

The initial rare variant burden analysis highlighted the significance of *GBA1* and *LRRK2* across multiple variant class categories, including missense, moderate/high impact, and LoF or highly deleterious variants (CADD PHRED > 20). Specifically, *GBA1* variants showed significant results in both the Genentech (P-values ranging from 1.32E-08 to 6.99E-08) and UK Biobank parent proxies (P = 2.15E-10 for all variant class categories) datasets. Similarly, *LRRK2* variants were significant in the combined AMP-PD and NIH dataset (P-values ranging from 1.96E-07 to 2.23E-07).

Additionally, LoF variants in *B3GNT3* were found to be exome-wide significant in the Genentech dataset (P=4.40E-09), with nominal significance in the UK Biobank parent proxies dataset (P=0.032; **Supplementary Figures 3 through 9**; Genentech [hg38: chr19:17807816:T:G, chr19:17807816:T:G,

chr19:17807816:T:G] and UK Biobank [hg38: chr19:17807982:GC:G; chr19:17808033:C:T; chr19:17812105:C:CA]). *TUBA1B* showed significance in moderate and high impact variants in the UK Biobank parent proxies dataset (P=9.48E-07). In the UK Biobank cases-control dataset, significant results were observed for LoF or highly deleterious variants in *ADH5* (P=3.13E-07) and in the sibling proxies dataset for *OR1G1* (P=6.58E-07; **Table 7; Supplementary Table 8**).

Ultra-rare variant (MAF < 0.1%) burden analysis further demonstrated the significance of variants in *GBA1* within the UK Biobank parent proxies dataset (P-values ranging from 6.88E-08 to 5.13E-10). *GBA1* also showed significance in the case-control dataset (P=4.56E-07). *LRRK2* variants were significant in the Genentech dataset (P=6.15E-07). Noteworthy findings in the UK Biobank case-control dataset included significant moderate/high impact variants in *AUNIP* (P=3.04E-08), and *TUBA1B* in the parent proxies dataset (P=9.48E-07). Additionally, LoF variants in *B3GNT3* (P=4.40E-09) and *AUNIP* (P=3.13E-08) were significant in the Genentech and UK Biobank case-control datasets, respectively. Finally, the ultra-rare variant burden analysis found no significant genes exome-wide within the AMP-PD and NIH genomes across the four variant classes (P < 1E-6; Table 8; **Supplementary Table 7**).

Meta-analyses

We designed and I executed two distinct meta-analyses. The first, referred to as the case-control meta-analysis, did not incorporate any UK Biobank proxy-cases, while the second, named the case-control-proxies meta-analysis, did include these proxy-cases alongside the standard PD cases and neurologically-healthy controls. In both meta-analyses, lambda values remained within the expected range (λ_{1000} = 0.97-1.00), indicating no significant deviation from the norm (**Supplementary Table 10**).

In terms of rare variant burden analysis, both meta-analyses demonstrated significant exome-wide associations for *GBA1* across different variant classes – missense, moderate/high impact, and LoF or predicted highly deleterious variants. The results were notably robust, with the case-control analysis yielding P-values of 3.27E-14, 9.10E-15, and 3.722E-14 for these variant classes, respectively. The case-control-proxies analysis presented even more striking findings, with P-values of 1.46E-21, 1.32E-22, and 9.12E-22, respectively. Furthermore, high confidence LoF variants in *CAPN10* and *B3GNT3* also showed significant exome-wide associations in both analyses (**Table 7**), with P-values of 3.60E-07 and 4.40E-09 in the case-control and 7.84E-07 and 3.36E-09 in the case-control-proxies meta-analyses, respectively.

The ultra-rare variant burden analysis further reinforced these findings. It demonstrated significant exome-wide associations for moderate/high impact and high confidence LoF variants in *AUNIP* in both meta-analyses. The case-control analysis showed P-values of 1.54E-08 and 1.64E-08, respectively, while the case-control-proxies analysis reported 2.70E-07 and 2.04E-07. Additionally, *TREML1* showed significance with the inclusion of proxy-cases. In line with the rare variant analysis, ultra-rare LoF variants in *CAPN10* and *B3GNT3* also remained significant. Importantly, both rare (MAF < 1%) and ultra-rare (MAF < 0.1%) variants in *GBA1* were associated with an increased risk of PD (**Table 7 and Table 8**).

B3GNT3 emerged as a gene of interest in the high confidence LoF variant class with significant p-values, especially in the Genentech dataset (P=4.40E-09) and to a lesser extent in the UK Biobank parent proxies dataset (P=0.032). Nevertheless, no variants meeting the criteria for high confidence LoF were found in the AMP-PD and NIH genomes, leaving the association of rare LoF variants in *B3GNT3* within these datasets unconfirmed. A majority of the novel candidate genes identified in this study, including *B3GNT3, AUNIP, ADH5, TUBA1B, OR1G1, CAPN10, and TREML1*, achieved exome-wide significance exclusively through the SKAT-O test (**Supplementary Table 3**). Comprehensive results from both the SKAT-O and CMC Wald burden tests for each variant class, MAF cutoff, and meta-analysis group are available on our dedicated GitHub repository.

Conditional *LRRK2* Analysis

I also conducted a conditional analysis on *LRRK2*, a gene known to be associated with PD, particularly focusing on the p.G2019S (rs34637584) variant, a relatively prevalent as a risk factor. To understand the extent to which the rare variant association at *LRRK2* is influenced by p.G2019S, we incorporated the allelic status of *LRRK2* p.G2019S into our analysis. Each individual's *LRRK2* p.G2019S status was quantified as 0, 1, or 2, reflecting their allelic dosage. This method allowed me to condition this analysis on the *LRRK2* p.G2019S status while retaining carriers in the dataset.

We then included the allelic status as a covariate in our burden analyses. Interestingly, the initial association observed at *LRRK2* dissipated (P > 0.05) after this conditional adjustment across all tested variant categories and MAF thresholds in the discovery datasets, with the exception of the Genentech dataset (**Supplementary Table 12**). This suggests that the rare variant association at *LRRK2* might be predominantly driven by the p.G2019S variant.

Furthermore, apart from *LRRK2* p.G2019S, we did not identify any other significant coding risk factors within *LRRK2*. It is notable that other rare coding variants previously identified as risk factors for PD, such as *LRRK2* p.R1441H (J. Liao et al. 2014), were not observed in our study. This absence underscores the unique impact of the p.G2019S variant in the context of *LRRK2*-associated PD risk and highlights the necessity of considering specific variant influences in genetic analyses of complex diseases like PD.

Assessing Previously Reported PD Causal or High Risk Genes and GWAS Regions

In this comprehensive analysis, we evaluated numerous genes previously associated with PD based on rare variant associations. A detailed list of these genes is available in **Supplementary Table 13**, and for specific information on the frequency and number of variants for each gene, variant class, and dataset, refer to **Supplementary Tables 3 to 6**. Apart from *GBA1* and *LRRK2*, which were already discussed, none of the genes in our study reached exome-wide significance (P > 1E-6). Nevertheless, we observed notable sub-significant signals for LoF or highly deleterious variants in *ARSA* (P=8.73E-05) and *DNAJC6* (P=8.08E-04; **Supplementary Tables 3, 4, and 13**).

Interestingly, the gene *PRKN*, known for its strong association with early onset PD, did not show a significant P-value in our analysis (P=0.30). To further investigate, I examined the enrichment of

homozygous and potentially compound heterozygous *PRKN* mutations in PD. In the most stringent variant class (LoF or highly deleterious variants), I identified a frequency of 0.41% in cases and 0.07% in controls within the combined AMP-PD and NIH dataset (**Supplementary Table 14**). Additionally, I did not find significant P-values in *VPS35* (present only in the Genentech dataset), known for its autosomal dominant association with PD, or *SNCA*, associated with earlier onset and more severe PD. *VPS35* showed a P-value of 0.235 in the case-control meta-analysis, and *SNCA* had a P-value of 0.274 in the case-proxy-control meta-analysis, suggesting limited impact of these genes in our dataset (**Supplementary Tables 3 and 5**).

We also sought to determine if known PD loci identified through GWAS presented rare variant associations. This approach was inspired by previous findings near genes like *SNCA, GBA1, GCH1, VPS13C,* and *LRRK2* (Jansen et al. 2017; Gaare et al. 2020; Rudakou et al. 2021; Mencacci et al. 2014). Our assessment covered 82 PD GWAS regions: 78 from the largest GWAS of Europeans (Nalls et al. 2019a), two from the largest PD GWAS of East Asians (Foo et al. 2020b), and two from the largest PD GWAS on progression (Tan et al. 2021); **Supplementary Table 15**). In our meta-analyses, only *GBA1* and *LRRK2* reached significance after applying the Bonferroni correction for 2,361 unique genes within 1 Mb of known PD loci. This suggests that the GWAS signals in these regions are likely driven by non-coding variants rather than coding variants, underscoring the complex genetic architecture of PD and the potential predominance of non-coding genetic factors in its pathogenesis.

Conclusions and Discussion

In this study, I present the results from comprehensive rare variant gene burden tests for PD, using 7,184 PD cases, 6,701 proxy-cases, and 51,650 healthy controls. My meta-analysis on gene burden tests reinforces the association of rare variants in *GBA1* and *LRRK2* with PD risk among individuals of European ancestry.

Rare Variant Gene Burden in PD and Novel Gene Associations

In addition to *GBA1* and *LRRK2*, this study evealed several novel PD-associated genes, namely *B3GNT3, AUNIP, ADH5, TUBA1B, OR1G1, CAPN10,* and *TREML1*, that achieved exome-wide significance (P < 1E-6) in my analysis in sample sets that are not enriched for early onset PD or familial PD (**Supplementary Table 11**). However, it is important to note that these genes did not demonstrate consistent significance across all datasets. The lack of replication at exome-wide significance in independent datasets may be attributed to varying dataset powers, influenced by differences in sample size and geographical population diversity, which in turn affect the presence of these rare variants.

A notable finding from my study was the strong evidence of a novel rare variant association in *B3GNT3*. LoF variants in this gene showed a significant meta-analysis P-value (P=4.40E-09), primarily driven by the Genentech (P=4.40E-09) and UK Biobank parent proxies (P=0.032) datasets. However, such variants were not found in the combined AMP-PD and NIH genomes, suggesting the necessity of additional data to

confirm their association with PD risk. Further investigation into *B3GNT3* revealed three LoF variants potentially linked to an increased risk of PD. Among the four individuals in the Genentech dataset carrying these *B3GNT3* variants, three were diagnosed with PD, and one was a control. Notably, all three PD cases reported a family history of the disease, which is relatively common in this cohort but does not necessarily indicate familial PD. These patients experienced an earlier onset of PD symptoms (in their 30s or younger), although they did not exhibit a notable prevalence of tremors, gait disturbances, REM sleep disturbances, or anosmia. Additionally, no significant identity-by-descent (IBD) linkage was found among these three PD cases (the proportion of the genome where zero alleles are shared IBD, average k_0=0.91). Given the rarity of these variants in *B3GNT3*, driving the association primarily in the Genentech and UK Biobank parent proxies datasets, it's plausible that they are absent in other datasets examined.

This not only corroborates the role of *GBA1* and *LRRK2* in PD, but also introduces several other genes potentially associated with the disease. The observed variability in gene significance across different datasets underscores the challenges in studying rare genetic variants in PD and highlights the need for further research to validate these findings.

Mitigating Bias in Genetic Burden Analysis

Rare genetic variants often exhibit population specificity, potentially introducing bias in burden analyses if population substructure is not adequately addressed. These variants may emerge in specific populations due to genetic drift, founder effects, or selective pressures, leading to strong associations within one group while being rare or absent in others. If not properly accounted for, this can result in spurious associations that do not accurately reflect genetic risk factors for the disease under study. To mitigate this issue, I implemented two strategies: inclusion of top eigenvectors from PCA as covariates in all models to control for population stratification and minimize ancestry-related confounding, and application of a meta-analysis approach to reduce potential bias arising from subpopulation differences between cohorts, thereby limiting the influence of population-specific variants.

An example of a population-specific variant is *LRRK2* p.G2019S, a known risk factor for PD that is enriched in Ashkenazi Jewish populations. Given its prevalence in the study cohorts, I conducted a conditional analysis to determine whether the observed rare variant associations within the *LRRK2* gene were primarily driven by p.G2019S. This conditional analysis incorporated the allelic dosage of *LRRK2* p.G2019S (0, 1, or 2) as a covariate in the burden analysis. Post-conditioning, the initial associations at *LRRK2* were no longer significant ($P > 0.05$) in all cohorts except Genentech.

Assessment of Previously Suggested PD GWAS Loci and Genetic Testing Limitations

In my investigation, I examined rare variants in previously suggested PD GWAS loci, including genes like *SYT11*, *FGF20*, and *GCH1* (Pu et al. 2022). Despite thorough analysis, I did not identify significant p-values in these genes, a finding that aligns with a parallel, albeit smaller, study conducted in the East Asian population (Pu et al. 2022). This raises questions about the underlying mechanisms at these PD GWAS loci. Although it seems plausible that some risk variants may influence gene expression differences, it remains unclear whether all such variants contribute to PD risk through this mechanism.

A notable observation from my research was the absence of several genes typically associated with PD, such as *PINK1* and *PRKN*, known as common genetic causes of early onset PD (Kasten et al. 2018). This was somewhat anticipated, as the burden testing algorithms employed are more effectively tuned to identify dominant and high-risk variants, like those found in *GBA1* and *LRRK2*, but less sensitive to recessive and ultra-rare mutations. It is also important to consider the distinctive clinical presentations in PD patients carrying *PRKN*, *PINK1*, and *SNCA* mutations, which often include earlier onset, variable progression rates, and rapid dementia onset, differing from the general PD population (Klein and Westenberger 2012). Given that the majority of PD cases in my study manifested symptoms in their sixties, it is less probable that these cases would harbor pathogenic *PRKN* mutations, which are more characteristic of early onset PD (**Table 6**).

Additionally, it is crucial to highlight the extreme rarity of certain disease-causing variants. For instance, pathogenic missense variants in *SNCA* have only been documented in approximately 25 reports, suggesting their scarcity in the broader population. Consequently, such mutation carriers are likely underrepresented in the datasets used in my study. This underrepresentation underscores a significant challenge in genetic research of PD, particularly when examining rare variants and their contributions to disease etiology.

Exploring the Role of Immune Response and Microtubule Defects

The involvement of the immune system, particularly the adaptive T lymphocyte response in PD, has been well-documented (Mosley et al. 2012). A gene of interest in my study is *B3GNT3*, which is responsible for encoding an enzyme crucial in the synthesis of L-selectin. This enzyme plays a pivotal role in lymphocyte homing, especially in the rolling of leukocytes on endothelial cells, a process vital for their migration to inflammatory sites. This suggests a possible link between *B3GNT3* and the immune responses observed in PD.

Another gene, *TUBA1B*, encodes the 1B chain of alpha-tubulin, a primary constituent of the cytoskeleton. Recent studies suggest that defects in microtubules, which are integral to the cytoskeleton, might contribute to the progressive neuronal loss characteristic of PD (Calogero et al. 2019; Pellegrini et al. 2017). The aggregation of alpha-tubulin, resulting from mutations in genes like parkin and alpha-synuclein, which are known to be implicated in PD, further underscores the potential role of *TUBA1B* in the disease's pathology (Ren, Zhao, and Feng 2003; Cartelli et al. 2016).

TREML1, a TREM receptor, has been increasingly implicated in various neurodegenerative disorders, including Alzheimer’s disease, PD, and multiple sclerosis (Dardiotis et al. 2017; Feng et al. 2019; Piccio et al. 2008). This highlights the potential importance of *TREML1* in the context of PD. Additionally, *ADH5*, which encodes one of the alcohol dehydrogenases, has been studied for its association with PD risk, albeit with conflicting results (J. J. Kim et al. 2020; Buervenich et al. 2005; García-Martín et al. 2019). The examination of *ADH5* in PD may provide further insights into its role in disease progression or risk.

Other genes nominated in this analysis were *AUNIP, OR1G1,* and *CAPN10*. *AUNIP* is involved in cell cycle regulation and DNA damage response, playing a critical role in maintaining genomic stability (Yang et al. 2019). *OR1G1* encodes a member of the olfactory receptor family, which are G-protein-coupled receptors involved in the initiation of neuronal responses for odor perception (Ben-Arie et al. 1994). *CAPN10* encodes calpain-10, a cysteine protease implicated in proinsulin processing and insulin secretion, with genetic variations in this gene associated with type 2 diabetes susceptibility and altered glucose-stimulated insulin secretion (Stumvoll et al. 2001). Despite these insights, there remains no clear connection between the known biology of PD and the functions of *AUNIP*, *OR1G1*, and *CAPN10*. As such, it is imperative to conduct further studies that offer both genetic support and functional data for these genes. This will be crucial in elucidating their potential roles in PD.

Future Directions and Limitations in Rare Variant Analysis of PD

Despite this being the largest effort to identify rare genetic variants in PD, there are several key limitations to this work. A significant limitation of my study is its restriction to individuals of European ancestry. It is essential to extend rare variant analyses to non-European populations and consider different age-at-onset ranges, especially as more whole genome and exome sequencing data become available. This expansion is crucial for a more comprehensive understanding of PD's genetic diversity and its implications across various demographics.

My analysis focused on four variant classes, which inherently limits the scope to specific types of genetic mechanisms. For instance, in analyzing LoF variants, I concentrated on mutations impairing protein function and their disease risk impact. This approach may overlook other disease mechanisms, such as gain-of-function mutations. Additionally, while the sample size is substantial, there is a lack of power to detect associations in genes where a small percentage of variants are functional or causal (**Table 5**). This limitation is particularly relevant as some rare variant tests differentially weigh variants based on penetrance and effect size.

The comprehensive literature search for rare variant associations was not limited to late-onset PD, potentially influencing the failure to replicate these associations in our analysis focusing on late-onset PD. Additionally, not all datasets in the meta-analysis were jointly called from raw read alignments, raising the possibility of batch effects influencing the results. However, the meta-analysis model used was designed to mitigate these biases by leveraging the power of the datasets without combining them. Further investigation, including segregation analysis in multiplex families or resequencing in large case-control datasets, is warranted, especially for early onset and familial cases.

The inclusion of parent and sibling proxy-cases from the UK Biobank aimed to enhance statistical power. While proxy-cases are valuable in large-scale common variant studies (Nalls et al. 2019a), caution is necessary when investigating recessive diseases. Most patients in this study represent the general PD population, where a smaller fraction reports a positive family history. Future studies should focus on recruiting patients suspected of having monogenic forms of PD, as they are more likely to carry highly

pathogenic or causal mutations not typically associated with PD (Global Parkinson's Genetics Program 2021). The clinical heterogeneity within PD cases necessitates further validation of the pathogenicity of rare or ultra-rare variants (Campbell et al. 2020; Mu et al. 2017; Sauerbier et al. 2016). My analysis was restricted to SNVs and small indels, not encompassing copy number variants due to data access limitations. Future studies should include copy number variants, as they have been shown to be significant in PD, using long-read sequencing for a more robust identification of these variants (Billingsley et al. 2023; (Daida et al. 2023).

In conclusion, while my study has reaffirmed *GBA1* and *LRRK2* as genes harboring rare variants associated with PD and has nominated several other previously unidentified genes, the need for further research is evident, prioritizing familial PD cases and individuals of non-European ancestry. This study underscores the complex interplay between genetics and PD, highlighting the potential contributions of specific genes to the disease's pathophysiology. While certain genes have established links to PD, others like *AUNIP*, *OR1G1*, and *CAPN10* warrant further investigation to fully understand their involvement in this neurodegenerative disorder. Deeper exploration into the biological mechanisms of these genes, particularly *B3GNT3* and *TREML1*, and replication in more diverse and familial datasets, will enhance our understanding of PD genetics.

Chapter 4: Expanding GWAS: Assessing Genome-wide Parkinson's Disease Risk in the African and Admixed Populations

Overview and Broader Relevance

The current understanding of PD heavily relies on studies conducted in populations of European ancestry, leading to a critical knowledge gap in PD genetics, particularly in African and African admixed populations. In contrast to the previous chapter, here we focus on genome-wide common variation and how they are associated with disease risk. Despite the advancements in GWAS across various ethnicities, these populations have been notably underrepresented. This oversight has limited our comprehension of PD's genetic diversity and risk factors. In studies involving European, Asian, and Latin American populations, numerous loci linked to disease risk have been identified. Specifically, research within European populations has discovered 78 loci and 90 distinct signals related to the risk of PD. In Asian populations, nine loci have been replicated, along with the identification of two new signals unique to this ancestry. Additionally, recent genome-wide association studies encompassing multiple ancestries have brought to light 12 new loci. However, the genetic landscape of PD in African populations remained largely unexplored.

Addressing this gap, this study represents the first extensive genome-wide assessment of PD genetics in African and African admixed populations, with the analysis including a total of 197,918 individuals. Using data from various cohorts, mainly from the Global Parkinson's Genetics Program (GP2) and 23andMe, the study focuses on identifying ancestry-specific genetic risk factors, analyzing haplotype structures, and exploring coding and structural variations.

My research uncovered a novel common risk factor at the *GBA1* locus in African populations, and identified a novel disease mechanism via expression changes consistent with decreased *GBA1* activity levels. This study bridges the diversity gap in PD research, unveiling a critical genetic risk factor in *GBA1* unique to African and African admixed populations. It underscores the importance of ancestry-specific genetic research and highlights the distinct genetic risk factors in these populations. The findings offer novel avenues for RNA-based therapeutic strategies and inform the design of future clinical trials, emphasizing the need for inclusive representation, ultimately paving the way for personalized, efficient treatments and advancing the field towards equitable healthcare solutions.

This work has been published here:

> Rizig M*, Bandrés-Ciga S*, **Makarious MB***, Oluwadamilola O, Wild Crea P, Abiodun O, Levine KS, …, Nigeria Parkinson Disease Research Network, International Parkinson's Disease Genomics Consortium - Africa, Black and African American Connections to Parkinson's Disease (BLAAC PD) Study Group, the 23andMe Research Team, Blauwendraat C, Houlden H, Singleton AB,

Okubadejo N, Global Parkinson's Genetics Program: **"Genome-wide Association Identifies Novel Etiological Insights Associated with Parkinson's Disease in African and African Admixed Populations"** *Lancet Neurology (2023);* https://doi.org/10.1016/S1474-4422(23)00283-1

Introduction

Globally, PD affected approximately 6.1 million people in 2016, with projections estimating a rise to 17.5 million by 2040, owing to an aging population and increased longevity (GBD 2016 Parkinson's Disease Collaborators 2018; Dorsey et al. 2018). Notably, ethnic variations in monogenic causes and genetic risk factors have been observed across different populations. For example, the *LRRK2* p.G2019S mutation, while prevalent in familial and sporadic PD cases in Zambia and Northern Africa, has not been reported in some sub-Saharan African populations (N. Okubadejo et al. 2008; Cilia et al. 2012; N. U. Okubadejo et al. 2018; Yonova-Doing et al. 2012).

Africa, a continent of rich ethnic diversity, is home to numerous ethno-linguistic groups spread across its varied geographical landscapes. In West Africa, with Nigeria being the largest country, the population primarily belongs to the Niger-Congo phylum. Conversely, North African and Northeast African populations predominantly fall under the Afroasiatic and Nilo-Saharan phyla (Schlebusch and Jakobsson 2018). However, despite this diversity, comprehensive and high-quality data on the prevalence of PD in African populations is notably lacking.

The age-standardized prevalence rates of PD, as indicated by the Global Burden of Disease 2016 data, suggest that sub-Saharan Africa has one of the lowest rates, ranging from 30 to less than 60 cases per 100,000 population. In stark contrast, North Africa exhibits prevalence rates more akin to those observed in Europe and the Middle East, with notably higher figures (GBD 2016 Parkinson's Disease Collaborators 2018; Safiri et al. 2023;N. U. Okubadejo et al. 2006; Blanckenberg et al. 2013). Furthermore, PD prevalence demonstrates a male predominance, being about 1.4 times more frequent in males as per the 2016 GBD estimates (GBD 2016 Parkinson's Disease Collaborators 2018). This observation aligns with previous studies from Africa, which generally report a higher male-to-female ratio, with overall male:female prevalence ratios ranging between 1.32 and 1.39 (Khedr et al. 2015; El-Tallawy et al. 2013; Zirra et al. 2023). However, it's crucial to note that most of these data are derived from hospital-based studies, which might be biased due to higher hospital attendance and health-seeking behaviors for non-obstetric reasons among men, influenced by a variety of social and cultural factors.

In sub-Saharan Africa, the typical clinical presentation of PD is characterized by a delayed diagnosis, often at a more advanced stage of the disease. Interestingly, the motor and non-motor manifestations in these populations appear to be similar to those observed in other regions of the world (O. O. Ojo et al. 2020; Oluwadamilola O. Ojo et al. 2021). In a contrasting context, African Americans have been reported to exhibit higher rates of cognitive impairment compared to White populations, alongside a lower propensity for parkinsonism (Bailey et al. 2021) and reduced medication usage (Xie et al. 2021).

These findings underscore the need for more region-specific and culturally sensitive studies in Africa to better understand the epidemiology and clinical manifestations of PD. This will not only aid in developing tailored healthcare strategies but also contribute to the global understanding of PD's diverse presentations and progression patterns.

GWAS have been instrumental in unraveling the genetics of complex disorders such as PD, and have been pivotal in elucidating the hereditary aspects of PD in these populations (Leonard et al. 2024). The most comprehensive GWAS meta-analysis on PD risk to date, focusing on European ancestry, has identified 90 unique genome-wide significant risk signals, accounting for approximately 16-22% of PD's inheritable risk (Nalls et al. 2019b; Blauwendraat, Nalls, and Singleton 2020). The largest PD GWAS meta-analysis in East Asian populations identified two population-specific signals (Foo et al. 2020a), and a recent GWAS in Latin Americans proposed two potential novel loci for further exploration (Loesch et al. 2021). The first multi-ancestry PD GWAS meta-analysis nominated 11 novel loci, laying the groundwork for future research aimed at fine-mapping novel genetic regions linked to PD (J. J. Kim et al. 2022).

GWAS are also invaluable tools for improving disease prediction models and expanding our biological understanding of specific diseases (Sara Bandres-Ciga et al. 2020). Recent genetic studies in European populations suggest that nearly one-third of PD's genetic heritability can be attributed to polygenic risk scores (PRSs). Heritability estimates for PD indicate that common genetic variants account for approximately 22-27% of disease risk. GWAS loci explain 16-36% of this risk, with SNP-based narrow-sense heritability calculated using LD score regression (LDSC) estimated at about 22%. This LDSC estimate is considered conservative, while GCTA heritability estimates are slightly higher, around 27%. Twin studies further support the genetic contribution to PD, suggesting a heritability estimate of 34% (Nalls et al. 2019a). However, the extent to which PRSs explain heritability in under-researched and underserved populations remains unknown (Elsayed et al. 2021; Schumacher-Schuh et al. 2022). Despite these advances, the genetics of PD in non-European populations remains largely uncharted.

African and African admixed populations present unique opportunities for genetic studies of both monogenic and complex diseases due to their high within-population genetic variability, shorter linkage disequilibrium (LD) blocks, and the high presence of population-specific alleles (Choudhury et al. 2020). Studying these populations not only promotes scientific equity and addresses health disparities but also provides a vital platform for validating the findings from other populations. This approach could lead to the discovery of novel or unique genetic loci and enhance our understanding of genotype-phenotype correlations, shedding light on the pathophysiological mechanisms of PD.

Methods

Methodological Framework and Demographic Composition of Participants

Visual overviews of the study design and the participant selection proces are presented in **Figure 8 and Figure 9**. The data for this study was sourced from three key entities: the International Parkinson’s

Disease Genomics Consortium - Africa (IPDGCAN), the Global Parkinson's Disease Genetics Program (GP2), and GWAS summary statistics from 23andMe, Inc.

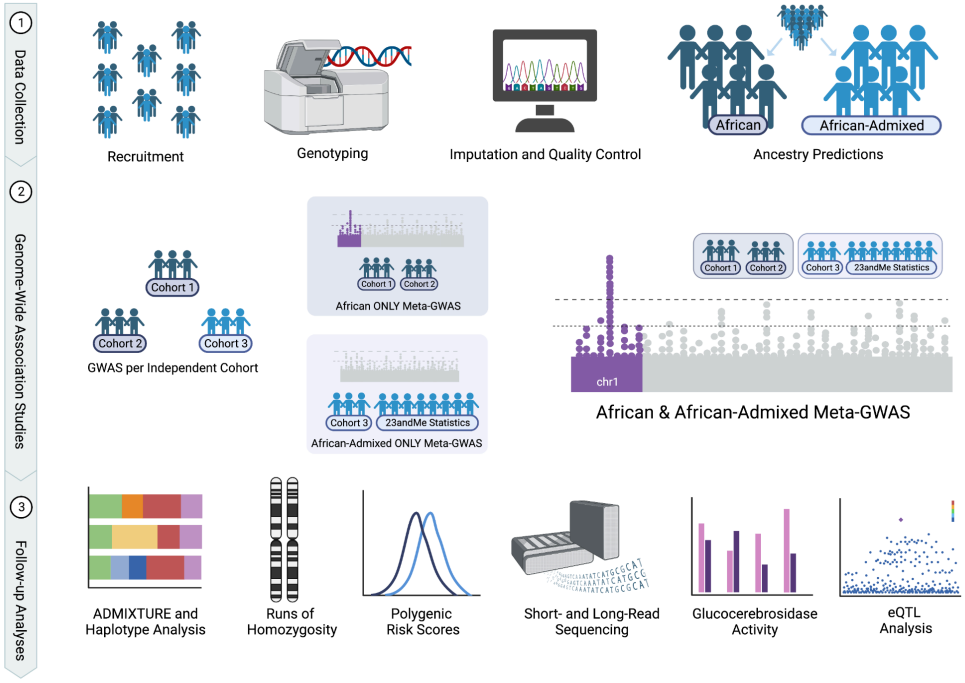


Figure 8: Analysis workflow schematic for the GWAS in African and African Admixed individuals

GWAS: Genome-wide association study; eQTL: Expression quantitative trait loci

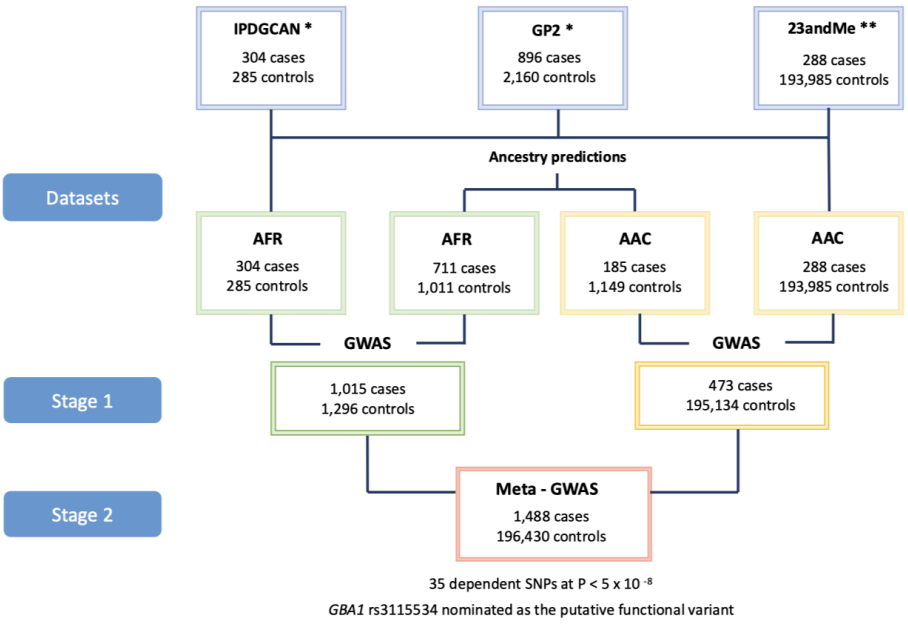


Figure 9: Workflow diagram with case/control breakdown per dataset for GWAS in African and African Admixed Individuals

IPDGCAN: International Parkinson's Disease Genomics Consortium - Africa; GP2: Global Parkinson's Genetics Program; AFR: African; AAC: African Admixed; GWAS: Genome-wide association study; SNP: single nucleotide polymorphism

* Genotyped samples underwent quality control procedures before taking part in this study

** Summary statistics were obtained via collaboration with 23andMe

IPDGCAN and GP2 Data Collection

The PD cases sourced from efforts in Africa, predominantly from West Africa and specifically Nigeria, present a limitation in terms of representativeness for the entire African continent. Nonetheless, these cases offer valuable insights into the genetic underpinnings of PD in this specific region. On the other hand, the control group in this study is largely derived from global efforts, displaying a higher degree of genetic admixture. It is important to note that while some individuals are predicted to be of African ancestry, their specific origin within Africa, particularly Nigeria, cannot be ascertained with absolute certainty. However, our principal component analysis, in comparison with the 1000 Genomes reference panel (**Figure 10**), confirms their unmistakable African heritage. For the purpose of this study, individuals of African admixed ancestry are defined based on their ancestral similarities to the African ancestry in Southwest United States of America (ASW) and African Caribbean in Barbados (ACB), as categorized in the 1000 Genomes project.

Figure 10: African cohort with 1000 Genome populations

IPDGCAN: International Parkinson's Disease Genomics Consortium - Africa

In the context of this research, cohorts are defined as groups of individuals who share similar predicted ancestry and have undergone genotyping, imputation, and processing under uniform quality control parameters. For the IPDGCAN and GP2 cohorts, ethical approval for participation in genetic studies was obtained from the respective ethical committees for medical research. Participants in these cohorts provided informed written consent. Additionally, all cohorts involved in the GP2 initiative are subject to a rigorous review of consent forms by the Operations and Compliance working group. This review ensures adherence to the ethical guidelines set by their respective institutional review boards. Consent includes participation in the initial cohorts and any subsequent studies within the constraints of local laws.

In the studies collected by IPDGCAN and GP2, all participants underwent comprehensive neurological examinations. These examinations, performed by the study's neurologists, were essential for documenting the clinical and neurological status of each participant. The diagnosis of PD in these studies hinged on meeting the criteria set forth by the United Kingdom PD Society Brain Bank (Hughes et al. 1992). This criteria was adhered to with one exception: the exclusion of participants with more than one affected relative was not a requirement for this study. Individuals who did not receive a PD diagnosis were classified as control subjects. These controls underwent assessments to identify any signs of neurological conditions. Any control subjects showing clinical indicators of neurodegenerative diseases were not included in the study cohort.

IPDGCAN and GP2 Data Generation and Processing

The genotyping process for the IPDGCAN and GP2 studies involved the use of two distinct platforms: the NeuroBooster array (version 1.0, Illumina, San Diego, CA, USA) and the NeuroChip array (version 1.0, Illumina, San Diego, CA, USA). The NeuroBooster array is notable for its comprehensive coverage, featuring a backbone of 1,914,935 variants, as well as including ancestry informative markers, IBD markers, X-chromosome SNPs for sex determination, and an additional 96,517 customized variants (Sara Bandres-Ciga et al. 2023).

For the GP2 initiative, sample genotyping was exclusively conducted using the NeuroBooster array. In contrast, the IPDGCAN initiative used both the NeuroChip array—with its 306,670 backbone variants and 179,467 customized variants (Blauwendraat et al. 2017) —and the NeuroBooster array. The raw genotype data from these arrays, containing 39,302 and 24,404 reference panel SNPs from NeuroBooster and NeuroChip arrays respectively, underwent a meticulous process involving a custom machine learning method for ancestry prediction and pruning, as part of the GenoTools pipeline (available at https://github.com/GP2code/GenoTools; Koretsky et al. 2022).

The process of ancestry estimation in this study involved a detailed and systematic approach. Initially, the samples were categorized based on ancestry estimates. These estimates were primarily determined using reference panels from three major sources: the 1000 Genomes Project (1000 Genomes Project Consortium et al. 2015), the Human Genome Diversity Project (Siva 2008), and a dataset focused on the

Ashkenazi Jewish population (Bray et al. 2010). The reference panel for ancestry estimation was comprehensive, consisting of 703 African, 601 South Asian, 585 East Asian, 534 European, 490 Latin American, 471 Ashkenazi Jewish, 190 African admixed, 183 Central Asian, 152 Middle Eastern, and 99 Finnish individuals. The reference panel was meticulously curated, excluding palindromic SNPs and filtering out variants with MAF below 5%, genotyping call rate below 99%, and Hardy-Weinberg Equilibrium (HWE) P value below 1E-4. The genotype data was then harmonized with the reference panel, with any missing genotypes imputed based on the mean value of the variant in the reference panel. For further refinement, GenoTools employed an 80/20 split for the reference panel samples into training and testing sets. PCs were derived from the overlapping SNPs, and these were transformed via Uniform Manifold Approximation and Projection (UMAP) to highlight global genetic population substructure and stochastic variation. These PCs were also used as covariates in our analyses. A linear support vector machine, was trained on the UMAP-transformed PCs. This classifier demonstrated robust performance in ancestry prediction, achieving balanced accuracies exceeding 0.95 during 5-fold cross-validation. These models were then applied to the GP2 and IPDGCAN data to estimate ancestry for all samples.

Standardized QC measures were applied to all samples. Exclusion criteria for the samples included: a call rate below 95%, discrepancies between genetically determined sex and clinical data, and excess heterozygosity indicated by |F| statistics exceeding 0.25. Further refinement involved the removal of samples exhibiting IBD across more than 12.5% of the genome, which typically indicates a relation as close as first cousins. SNP-level QC included the exclusion of SNPs failing the Hardy-Weinberg Equilibrium (HWE) with a P value less than 1E-4 in control samples, along with the pruning of variants for missingness by case-control status and haplotype, each at a threshold of P≤1E-4.

Additional filters were applied for MAF below <0.5% and HWE P value below 1E-5 before submission to the TOPMed Imputation server. The TOPMed reference panel, version r2, features data from 97,256 reference samples encompassing over 300 million genetic variants. This panel, as of October 2022, comprises approximately 180,000 participants with diverse ancestries: 29% African, 19% Latin American, 8% Asian, and 40% European. Further details about the TOPMed study, imputation Server, and Minimac imputation are documented elsewhere respectively (Taliun et al. 2021; Das et al. 2016; Fuchsberger, Abecasis, and Hinds 2015), and are accessible at https://imputation.biodatacatalyst.nhlbi.nih.gov. Following imputation, the data underwent pruning based on a MAC threshold of 10 and an imputation Rsq of 0.3, ensuring robust and reliable genetic data for subsequent analysis.

For the meta-analysis, we focused specifically on samples with African or African admixed ancestry. Using ADMIXTURE (v1.3.0; https://dalexander.github.io/admixture/binaries/admixture_linux-1.3.0.tar.gz), we ran a supervised analysis, categorizing individuals as 'African' if their African admixture was 90% or more, and as 'African admixed' if it was less than 90%. The GWAS results were annotated using ANNOVAR v2020-06-08, with samples clustering within African and African admixed ancestries represented alongside 1000 Genome populations for visualization in **Figure 10 and Figure 11**.

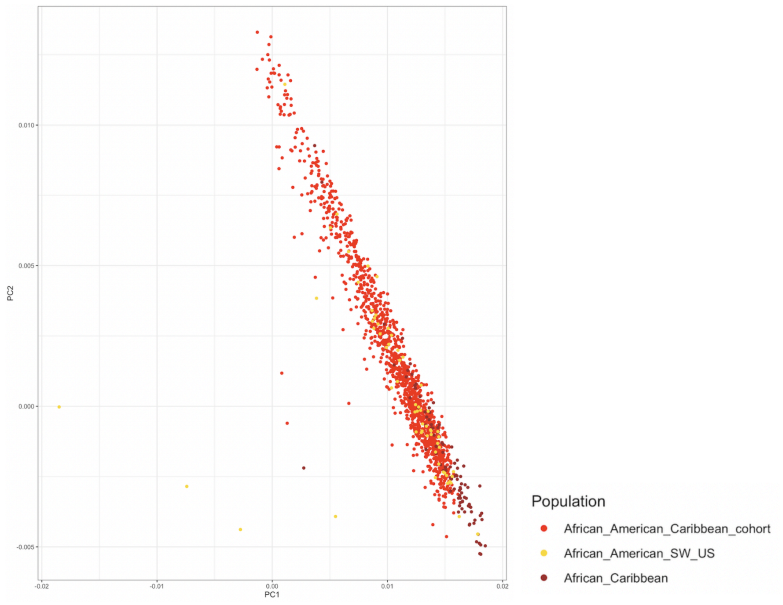


Figure 11: African Admixed cohort with 1000 Genome populations

Principal component analysis plots displaying African admixed samples under study (red) with 1000 Genome populations including Africans from the Southwest of the US (yellow) and African Caribbeans (brown)

Quality control and imputation for the 23andMe data set were performed separately and differently from the IPDGCAN and GP2 datasets. The age distribution of these cohorts is visualized in **Figure 12**.

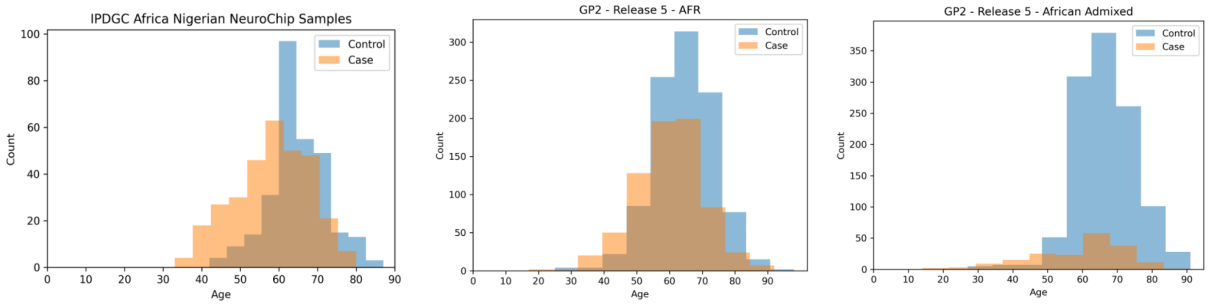


Figure 12: Age distributions of cohorts involved in studying susceptibility of risk in the African and African admixed populations

IPDGC: International Parkinson's disease Genomics Consortium; GP2: Global Parkinson's Genetics Program; AFR: African

23andMe Data Collection

The participants from 23andMe consented to take part in this research voluntarily, agreeing to the research protocol approved by the external AAHRPP-accredited IRB, Ethical & Independent (E&I) Review Services. As of 2022, this entity is a part of Salus IRB (https://wwwversiticlinicaltrialsorg/salusirb).

23andMe Data Generation and Processing

The 23andMe relied on self-reported information. This method involved recruiting PD patients through a collaborative effort with organizations such as the Michael J. Fox Foundation and The Parkinson's Institute and Clinical Center, as well as various PD patient groups and clinics. Recruitment materials, including emails and hard copy mailings, were sent to individuals affiliated with these groups who had identified themselves as PD patients. Only individuals who self-reported a PD diagnosis were considered PD cases for this study. However, those who reported either a change in their PD diagnosis or uncertainty about their diagnosis were excluded from further analysis. Prior research has demonstrated that self-reported PD status is generally as reliable as clinically diagnosed PD (Heilbron et al. 2019). Any individual with a self-reported history of atypical parkinsonism (e.g., dementia with Lewy bodies, progressive supranuclear palsy, multiple system atrophy, corticobasal degeneration) or non-parkinsonian tremor disorders was also excluded. Furthermore, a PD GWAS meta-analysis revealed a strong genetic correlation between self-reported PD cases in the 23andMe GWAS data and clinician-ascertained PD cases in non-23andMe GWAS datasets (genetic correlation from LDSC (rG) = 0.85, SE = 0.06; Nalls et al. 2019b).

23andMe participants were genotyped using one of five different platforms, each with varying amounts and types of SNPs. The platforms ranged from the Illumina HumanHap550+ BeadChip with approximately 560,000 SNPs to the Illumina Infinium Global Screening Array with around 640,000 variants. The genotyping platforms were designed to capture a wide array of genetic diversity and were standardized for research purposes. Samples that failed to reach a 98.5% call rate were re-genotyped, and participants with repeatedly failed analyses were contacted for additional samples. Across all platforms, a total of 1,522,458 variants were genotyped.

The ancestry classification for the 23andMe dataset employed an algorithm that compared the genome-wide content of 23andMe with an in-house customer reference panel consisting of 14,393 individuals. This process was carefully controlled by removing related individuals and those whose genetic ancestry did not align with their survey answers. A total of 45 ancestry populations were identified through principal component analysis using data from various genetic projects. The classifier algorithm partitioned phased genotyped data into windows and used a support vector machine to classify haplotypes into one of 45 worldwide reference populations. These classifications were then fed into a hidden Markov model (HMM) to refine the ancestry estimates, with recalibration using simulated admixed individuals. The final ancestry composition included six higher-level populations.

For African Americans and Latin Americans, whose ancestry compositions are varied, a logistic classifier was employed to differentiate between these groups based on the length of segments of European, African, and American ancestry.

In the 23andMe phasing process, they constructed a high-quality phasing panel by selecting 200,000 African American participants, each with less than 5% missingness in their genotyping data. 23andMe ensured the inclusion of only high-quality variants by focusing on SNPs, with criteria such as a MAF of at

least 0.1%, less than 5% missingness per variant, a correlation greater than 0.9 with sequence data, and MAF consistency with the gnomAD database. Additionally, 23andMe performed male heterozygosity tests on the non-pseudoautosomal regions of the X chromosome. The phasing of participant genotyping data was carried out using SHAPEIT4, tailored to each specific genotyping platform (Delaneau et al. 2019).

The imputation process at 23andMe employed two key reference panels for imputation: the Human Reference Consortium (HRC) panel, accessible from the European Genome-Phenome Archive at the European Bioinformatics Institute (accession EGAD00001002729), and the 23andMe reference panel, developed using a combinated of internal and external WGS datasets. The HRC panel, comprising 27,165 samples, were adjusted to the hg38 human genome reference sequence, with those moving to different chromosomes during this 'liftover' process being excluded. Subsequently, these variants were re-phased using SHAPEIT4, and singletons (variants appearing in only one individual) were removed. The finalized HRC panel included an array of 39,057,040 SNPs.

The 23andMe reference panel was curated from 12,217 samples, selected from both internal and external sources **(Supplementary Table 16)**. This panel's composition provides a broad and diverse genetic representation, essential for accurate imputation. Imputation was conducted using Beagle 5.31, targeting 85 million variants across the entire panel. This process was tailored according to the origin of the variants: those unique to the HRC panel were imputed using HRC samples, those exclusive to the 23andMe panel were imputed with 23andMe samples, and variants common to both panels were imputed using a combined sample set. Imputation was performed independently for each genotyping platform used by 23andMe.

To ensure the validity of their GWAS, 23andMe meticulously selected unrelated individuals using a segmental IBD estimation algorithm. Individuals were defined as related if they shared more than 700 centimorgans IBD, approximating the genetic sharing expected between first cousins in an outbred population. In selecting participants for phenotype analyses, cases were prioritized over controls to maximize the case sample size, retaining cases in instances where both a case and control were related. For the 23andMe GWAS, they conducted logistic regression analyses on both genotyped and imputed SNPs, assuming additive allelic effects. They incorporated covariates such as age, sex, the top five PCs, and genotype platform indicators to account for potential confounding factors. The association test P values reported were derived from a likelihood ratio test, with X chromosome results treated similarly, coding male genotypes as homozygous diploid. 23andMe flagged SNPs genotyped only on their older platforms (v1/v2) due to their smaller sample sizes and those on chromosomes M and Y where calling is less reliable. Using trio data, they assessed parent-offspring transmission, flagging SNPs that deviated significantly from expected inheritance patterns. They also considered the call rate, date of genotyping, sex effects, and probe reliability, flagging SNPs that failed.

For imputed results, they flagged SNPs with low imputation quality ($r2<0.5$) and those showing evidence of batch effects related to genotyping platforms. Across all their GWAS results, SNPs with insufficient sample size or problematic logistic regression results were also flagged.

Assessment of Risk, Age of Onset, and Analysis of Genetic Admixture

To assess the risk associated with PD, I analyzed imputed dosages using a logistic regression model. These imputed dosages are genotype probabilities for a variant being A/A, A/B, or B/B, ranging from 0 to 2, and they incorporate a degree of uncertainty. My analysis was adjusted for several covariates, including sex, age, and the first ten PCs. These PCs were initially fitted on the set of SNPs overlapping between our datasets and the reference panel, to represent the underlying population substructure more accurately. I used the age at onset (AAO) for PD cases and the age at recruitment for control subjects. In cases where the AAO was unavailable (which was less than 6% of individuals), I used the age at recruitment as a substitute. Additionally, for those individuals who did not provide any age information, which was less than 5% and 2% of cases and controls respectively, I imputed an average age.

For generating summary statistics, I employed PLINK 1.9 and 2.0 (Purcell et al. 2007). I ensured that only data meeting a minimum imputation quality of 0.30 and a MAF greater than 5% were included in the analysis. To explore the impact of genetic variation on the AAO of PD cases, I performed a linear regression analysis, adjusting for the same covariates. In this analysis, AAO was defined based on the self-reported date of the first motor symptom.

Furthermore, we conducted linear regression analyses to examine the correlation between potential GWAS signals and levels of genetic admixture. All these analyses were carried out on the Terra platform (https://app.terra.bio/). For the GWAS, I focused on African and African admixed ancestries independently, using PLINK and applying a Bonferroni correction threshold of 5E-8 prior to conducting meta-analysis. To synthesize the summary statistics from all sources, I used fixed-effects meta-analyses, as implemented in METAL (Willer, Li, and Abecasis 2010). For calculating pairwise LD values, I relied on data from the 1000 Genomes Project's African population, accessed through LD link (https://ldlink.nci.nih.gov/?tab=home). Finally, the colocalization of association summary statistics was visualized using LocusCompareR (version 1.0; https://github.com/boxiangliu/locuscomparer).

Logistic Regression

In my research, I focused on estimating the risk associated with PD by analyzing imputed dosages. These imputed dosages represent the genotype probabilities for a variant to be either A/A, A/B, or B/B, ranging from 0 to 2. These probabilities take into account certain levels of uncertainty. My analysis employed a logistic regression model, adjusted for covariates including sex, age, and the first ten PCs.

Logistic regression is chosen over linear regression as the foundational model in GWAS when dealing with a binary phenotype, in this study positive disease cases compared to neurologically healthy controls, to assess the relationship between the genotype and the logarithm of the odds of developing the disease. The additive model, which was the chosen approach for this study, can be articulated as follows:

$$log\frac{Pr(Y=1|X=x)}{Pr(Y=0|X=x)} = \mu + x\beta$$

In this model, μ represents the logarithm of the odds, commonly referred to as 'log-odds', for genotype 0. Meanwhile, β signifies the log of the odds ratio (logOR) between genotype 1 and genotype 0. The exponential of β (exp(β)) corresponds to the actual odds ratio (OR). Furthermore, the logOR between genotypes 2 and 0 is denoted as 2β. This model operates additively on the log-odds scale, which translates to a multiplicative function on the odds scale.

Linear Regression

Linear regression has been a fundamental tool for identifying genetic variants associated with continuous traits in GWAS. Specifically, I applied linear regression to investigate how genetic variations influence the AAO of PD cases. This analysis was carefully adjusted for critical covariates such as sex, age, and the first ten PCs. In this context, AAO was precisely defined as the self-reported date when the first motor symptom was experienced. Moreover, my research extended to conducting linear regression analyses aimed at understanding the relationship between potential GWAS signals and levels of genetic admixture.

The foundational concept of simple linear regression, which underpins these analyses, can be summarized by the equation $y = mx + b$. In this formula, y represents the response or dependent variable, which, in the case of my study, pertains to the AAO of PD. The term x denotes the predictor or independent variable, which refers to the specific genetic variations under examination. The coefficient m is the estimated slope of the regression line, indicating the rate of change in the response variable per unit change in the predictor. Lastly, b is the estimated intercept, representing the expected value of the response variable when the predictor is zero. This linear regression framework enables a systematic exploration of the genetic factors influencing the onset and progression of PD.

Power Calculations to Estimate Sample Size Requirements

Power calculations were performed to accurately forecast the number of additional PD cases required to achieve genome-wide significance with 80% power. This calculation was particularly focused on variants with minor allele frequencies and effect estimates typical of common genetic variations identified through GWAS. Additionally, these calculations were based on a disease prevalence rate of 0.6% (as per the Global Burden of Disease Study). For this purpose, we used the publicly available Genetic Association Study (GAS) power calculator from the University of Michigan (https://csg.sph.umich.edu/abecasis/cats/gas_power_calculator/). This tool employs an additive genetic model, which is appropriate for the kind of genetic analysis typical in GWAS. The additive model assumption is crucial as it aligns with the nature of the genetic variants typically observed in complex diseases like PD. Power calculations help in determining the appropriate sample size to detect an effect of a given size with a specified level of confidence. By using this tool, I was able to estimate the additional number of PD case samples needed to robustly detect genetic variants that contribute to the disease's risk. This estimation is vital for planning future studies, ensuring they are adequately powered to detect clinically and biologically relevant genetic associations.

Population Attributable Risk

I extended the analysis of PD risk to include the estimation of population attributable risk (PAR), which quantifies the proportion of disease cases in the population that can be linked to a specific genetic variant. The PAR is particularly informative as it encompasses both the prevalence of the variant, represented by the MAF, and the strength of its association with the disease, indicated by the OR. The computation of PAR is based on the following formula:

$$PAR = \frac{MAF*(OR-1)}{MAF*(OR-1)+1}$$

This equation integrates the increased risk associated with the genetic variant (OR - 1) with its frequency within the population (MAF), providing a metric that captures the impact of the variant on the disease burden in the population. In the logistic regression framework, the PAR is derived from the genotype probabilities, which are the outcomes of imputed dosages, and the covariate-adjusted ORs. The calculation of PAR thus allows me to ascertain the public health implications of genetic findings by estimating the overall disease cases that could potentially be prevented if the effects of the risk alleles were mitigated.

Conditional Analysis

Conditional analysis was performed to assess any additional association signals at the specifically nominated *GBA1* locus. This analytical approach entailed conducting an association analysis while conditioning on the primary associated SNP at the *GBA1* locus, identified as rs3115534. The objective was to determine if there existed additional SNPs at this locus that were significantly associated in a manner independent of the primary SNP.

To achieve this, I conducted conditional analysis using individual-level genotype data. However, it's important to note a key limitation in this approach: the summary statistics obtained from the 23andMe study were not incorporated into these estimations. This exclusion was a necessary consideration in the methodology, as it impacted the comprehensiveness and the scope of the secondary signal identification within the *GBA1* locus. Despite this limitation, the conditional analysis provided valuable insights into the genetic complexity and the potential independent genetic contributors at the *GBA1* locus.

Fine-mapping and Haplotype Analysis

We conducted fine-mapping analyses to prioritize potential causal variants within the identified *GBA1* risk haplotype. This process was conducted using the R package 'coloc' (available at https://CRAN.R-project.org/package=coloc). The primary goal was to pinpoint putative causal variants across the LD block that harbored the genome-wide signal. The specific methodology we employed was the "approximate Bayes Factor fine-mapping under a single causal variant assumption" provided by the 'coloc' package. This approach evaluates the posterior probabilities of each SNP being the causal variant within a given locus. For this analysis, I calculated the posterior probabilities (PP) for the region in question, adhering to the default prior probability setting of 1E-4. This setting was based on the assumption that there is a single causative variant per locus.

The rationale behind using Bayesian genetic fine-mapping in genetic studies is to identify specific causal variants within GWAS loci that are responsible for each association. This is achieved by reporting credible sets of plausible causal variants. These sets are interpreted as containing the causal variant with a pre-defined level of coverage probability. By applying this method to the LD block containing the genome-wide signal, I aimed to provide a more nuanced understanding of the genetic architecture within the *GBA1* locus, enhancing our insights into its role in disease manifestation. This approach is particularly valuable in the context of complex genetic disorders, where identifying the specific genetic variants that contribute to disease risk can be challenging.

In this study, we also compare haplotype sizes using individual-level data from African, African admixed, and European PD cases. A critical step in this process was the standardization of the three datasets, ensuring that only the same genotyped SNPs that passed identical QC steps were included in the analysis. This harmonization of data was crucial for the integrity and comparability of the results. For the actual determination of the haplotype blocks, I employed PLINK 1.9, using its default parameters. This software estimates haplotype blocks based on Haploview's interpretation of block definitions. An important aspect of this method is that it considers only pairs of variants located within a 200-kilobase (kb) range of each other. This approach is designed to ensure a focused and relevant analysis of genetic linkage. The criterion for two variants to be regarded as in strong LD within this framework is stringent. It requires that the lower bound of the 90% D-prime CI is greater than 0.70, and simultaneously, the upper bound of the CI must be at least 0.98. This dual threshold ensures a high degree of confidence in the identification of strong LD between variants.

Glucocerebrosidase Activity Assay

For this study, our team acquired patient-derived lymphoblastoid cell lines (LCLs) from the Coriell repository ([Coriell Institute for Medical Research; https://www.coriell.org/). These LCLs were maintained in suspension culture conditions, using RPMI 1640 medium (ThermoFisher Scientific, 11875093) supplemented with 2mM Glutamax (ThermoFisher Scientific, 35050061), and 15% FBS (ThermoFisher Scientific, A3160501). The cultivation was carried out at a controlled temperature of 37°C and in an atmosphere containing 5% CO_2.

For protein extraction from these LCLs, we employed a citrate-phosphate buffer (0.2 M Na2HPO4, 0.1 M citrate, protease inhibitor, pH 5.8, Millipore Sigma, 11836170001), activated with 0.25% Triton X-100. To assess glucocerebrosidase (GCase) activity, we conducted a fluorometric assay using 4-methylumbelliferone (4-MU, Sigma Aldrich, M1381). This assay was performed in quadruplicate, adhering to the protocol previously reported (Peters, Lee, and Glew 1975), with an adjusted incubation time of 2.5 hours. Each assay sample comprised 5E6 cells, with protein concentrations normalized to 0.7 mg/mL using the BCA Protein Assay (Thermo Fisher Scientific 23225).

To analyze the trends in GCase enzymatic activity among different genetic groups, I used the Welch two-sample t-test. This test was specifically implemented to compare the GCase activity between

individuals homozygous for the rs3115534-GG risk allele and those heterozygous for the rs3115534-GT allele, as well as between rs3115534-GG homozygous risk allele carriers and rs3115534-TT homozygotes for the non-risk allele. The Welch's t-test is particularly suited for this analysis as it assumes normal distribution in both groups but does not require equal variances.

The test statistic is calculated using the formula:

$$(x_1 - x_2)/\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where x_1 and x_2 are the means of the two groups, s_1^2 and s_2^2 are the variances, and n_1 and n_2 are the sample sizes.

The degrees of freedom for this test are determined by the formula:

$$\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}\right)^2 / \left\{\left[\frac{(s_1^2/n_1)^2}{n_1 - 1}\right] + \left[\frac{(s_2^2/n_2)^2}{n_2 - 1}\right]\right\}$$

This formula accounts for the differences in standard deviations (SD) between the two samples. It is noteworthy that if both samples have identical SDs, the degrees of freedom for Welch's t-test coincide with those of the Student's t-test.

Short- and Long-read Whole Genome Sequencing

To delve deeper into the novel GWAS signal identified, WGS analyses were conducted on a cohort comprising 206 individuals. This group included 141 PD cases and 65 controls, among which there were 39 *GBA1* rs3115534-GG carriers, 69 rs3115534-GT carriers, and 98 rs3115534-TT carriers. The short-read WGS DNA sequencing for this cohort was performed by Psomagen. Additionally, Oxford Nanopore Technologies (ONT) long-read whole-genome sequencing data was generated for a select group of individuals: five carrying the *GBA1* rs3115534-GG allele, two heterozygotes, and six carriers of the *GBA1* rs3115534-TT allele. The extraction of high molecular weight DNA was conducted from either frozen blood samples or cell lines.

For the short-read WGS process, Psomagen quantified the starting genomic DNA material using a fluorescence-based Picogreen assay (ThermoFisher, cat. #P7589) and verified DNA integrity via genomic DNA screen tape (Agilent Technologies, cat. # 5067-5365/5067-5366) on a TapeStation 4200 (Agilent Technologies). Approximately 0.5µg of genomic DNA (gDNA) served as the input for Truseq PCR-free library construction. The fragmentation of gDNA to an insert size of 350 bp was achieved using

LE220-plus Focused-ultrasonicator (Covaris), followed by a series of purification and validation steps, culminating in adapter ligation and final library validation using KAPA Library Quantification Kits (Roche, cat. #07960298001) on a Light cycler 480 (Roche). The validated library was then normalized and loaded onto a Novaseq6000 for sequencing.

For data processing, we employed the functional equivalence pipeline (Regier et al. 2018) at the Broad Institute for alignments and variant calls against the GRCh38DH reference genome. We adhered to the quality metrics defined by the AMP-PD initiative (https://amp-pd.org; Iwaki et al. 2021) for sample-level WGS quality control. High-quality variants were identified following the Broad Institute's joint discovery pipeline, with stringent criteria for variant quality score recalibration. We also performed specialized calling of *GBA1* variants using Gauchian (v1.0.2; Toffoli et al. 2022) and genotyped known neurological repeat expansions with STRipy (v2.2; Halman, Dolzhenko, and Oshlack 2022). Variants meeting quality control standards were annotated using ANNOVAR (Kai Wang, Li, and Hakonarson 2010). The CRAM files were visualized using the IGV web browser (Robinson et al. 2011).

For the long-read sequencing, DNA extraction from blood samples used the Kingfisher APEX instrument and the Nanobind CBB Big DNA kit (HBK-CBB-001), while frozen cell pellets underwent manual extraction using the same kit following the protocol by Kolmogorov and colleagues (Kolmogorov et al. 2023; https://dx.doi.org/10.17504/protocols.io.q26g74169gwz/v1). Post size-selection using the Circulomics Short Read Eliminator Kit (SS-100-101-01), libraries were prepared with the SQK-LSK 110 Ligation Sequencing Kit from ONT and sequenced on a PromethION R9.4.1 flow cell for 72 hours.

The raw signal data obtained in Fast5 files from minKNOW v22.10.7 (ONT) underwent super accuracy basecalling with Guppy v6.12. The Fastq files that passed quality control were mapped to the GRChg38 reference genome using winnowmap (v2.03; Jain et al. 2022), and structural variants were called using Sniffles2 (v2.0.3; Smolka et al. 2022) with default parameters and the "`--tandem-repeats`" option. This comprehensive approach allowed for an in-depth exploration of the genetic landscape associated with the *GBA1* signal in our PD samples.

Polygenic Risk Profiling

Polygenic risk score (PRS) analysis for PD was performed as follows. Briefly, a PRS was calculated incorporating effect estimates from the European meta-GWAS summary estimates for the 90 SNPs previously associated with PD risk in European populations (Nalls et al. 2019b). Risk allele dosages were counted, then summed and a genetic risk score was generated across all loci in both African and African admixed data. All SNPs were weighted by their published betas, giving greater weight to alleles with higher risk estimates. PRSs were standardized to have a mean of 0 and SD of 1. Then, a logistic regression was performed regressing disease status against PRSs using PLINK's "`--score`" option. Risk profiling analysis was adjusted for age, sex, and PCs one through 10. I repeated these steps using the African admixed effect estimates from the 23andMe summary statistics for those same 90 SNPs identified by the European meta-GWAS assessing PD risk susceptibility (Nalls et al. 2019b).

Runs of Homozygosity

Based on an LD-pruned data set (using previously described parameters), runs of homozygosity (ROHs) were defined using PLINK 1.9 to assess potential over-representation of sharing recessive regions in cases versus controls. Here, we evaluated the largest individual-level dataset (*GP2;* African individuals; **Table 9**), where samples genotyped on the NeuroBooster array included 711 African cases and 1,011 African controls. I explored ROHs containing at least 10 SNPs and a total length greater or equal to 1,000 Kb, with a rate of scanning windows of at least 0.05 (not containing >1 heterozygous call or 10 missing calls). In order to explore overall homozygosity between cases and controls, three metrics were assessed, including the number of homozygous segments spread across the genome, total kilobase distance spanned by those segments, and average segment size on autosomes only.

	African predicted ancestry		**African admixed predicted ancestry***	
	Nigerian origin (IPDGC cohort)	African, broad unspecified origin (GP2 dataset)†	African admixed origin (GP2 dataset)§	African admixed origin (23andMe dataset)
Total participants	589	1,722	1,334	194,273
Recruited from Nigerian sites	589 (100%)	1,330 (77%)	50 (4%)	N/A
Cases	304	711	185	288
Recruited from Nigerian sites	304 (100%)	672 (95%)	16 (9%)	N/A
Female	80 (26%)	206 (29%)	80 (43%)	N/A
Male	224 (74%)	505 (71%)	105 (57%)	N/A
Controls	285	1,011	1,149	193,985
Recruited from Nigerian sites	285 (100%)	658 (65%)	34 (3%)	N/A
Female	97 (34%)	448 (44%)	714 (62%)	N/A
Male	188 (66%)	563 (56%)	435 (38%)	N/A
Case age at onset, years	58.20 (9.67)	59.31 (11.37)	57.84 (14.69)	N/A
Control age at examination, years	64.4 (7.56)	65.09 (9.55)	66.34 (8.71)	N/A
Array	NeuroChip	NeuroBooster	NeuroBooster	Omni Express & GSA & 550k

Table 9: Demographic and clinical characteristics of the cohorts under study in African and African admixed GWAS

Data are n, n (%) or mean (SD). N/A=not available
*African admixed defined as individuals ancestrally similar to the following 1000 Genomes project (https://www.internationalgenome.org/) ancestry labels: African ancestry in Southwest United States of America (abbreviated as ASW in the 1000 Genomes project) and African Caribbean in Barbados (abbreviated as ACB in the 1000 Genomes project)

†GP2 cohorts with predicted African ancestry include Baylor College of Medicine (https://www.bcm.edu/), BioFIND (https://biofind.loni.usc.edu/), BLAAC PD (https://www.blaacpd.org/), Coriell (https://www.coriell.org/), Movement Disorders Genotypes and Phenotypes – King's College London (MDGAP-KINGS; further details at https://gp2.org/the-components-of-gp2s-third-data-release/), PPMI (https://www.ppmi-info.org/), PAGE (https://www.pagestudy.org/), University of Maryland (https://umd.edu/), and IPDGCAF-NG (https://www.ipdgc-africa.com/)

§GP2 cohorts with predicted African admixed ancestry include Baylor College of Medicine, BioFIND, BLAAC PD, Coriell, MDGAP-KINGS, PPMI, PAGE, University of Maryland, Systemic Synuclein Sampling Study (S4; https://pubmed.ncbi.nlm.nih.gov/28353371/), and IPDGCAF-NG

Results

Recruitment and Composition of Study Cohorts for GWAS Meta-Analysis

In this study, we used data from three distinct datasets, namely GP2, 23andMe, and IPDGCAN, to conduct a GWAS meta-analysis. This analysis encompassed a total of 1,488 individuals identified as cases and 196,430 participants classified as controls. Importantly, all these participants were of African or African admixed ancestry (detailed in **Figure 9**). To provide a comprehensive view of the cohorts, I have detailed their demographic and clinical characteristics in **Table 9**.

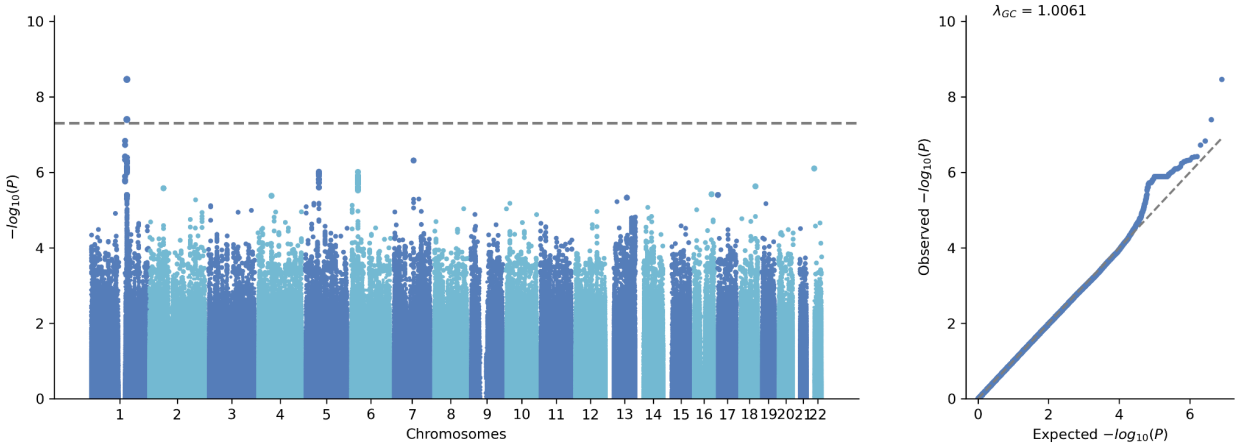


Figure 13: African Parkinson's disease risk GWAS

Manhattan plot displaying the association statistical significance as –log10(p-value) in the y-axis against chromosomes in the x-axis at a genome scale (Bonferroni correction highlighted at 5E-8). Quantile-quantile plot displaying genetic data distribution; λ GC: Lambda value representing genomic inflation

Genome-Wide Association Studies in African and African Admixed Populations Highlighting the *GBA1* Locus

My initial research involved conducting a GWAS focusing on PD risk within the African population, predominantly comprising individuals of Nigerian descent. This study included 997 PD cases and 1,294 control subjects. Of these, 693 PD cases and 1,009 controls were genotyped using the NeuroBooster array, while the remaining 304 PD cases and 285 controls were analyzed using the NeuroChip array (λ=1.01; **Figure 13**). The results highlighted a genome-wide significant SNP at the *GBA1* locus, specifically rs3115534 located in intron 8 of *GBA1*, 34 nucleotides upstream of exon 9, which emerged as the primary variant associated with increased PD risk (**Table 10 and Figure 13**; rs3115534; OR=1.58; 95% CI = 1.35 - 1.84, P=3.44E-09). Interestingly, despite examining common variation typically associated with PD risk (MAF > 5%), this variant exhibited a high OR. The study deduced that each additional G allele of rs3115534 increased the odds of developing PD by 1.58 times. Concurrently, I conducted a GWAS in the African admixed population, using data from African-American and Afro-Caribbean cohorts within the GP2 initiative, supplemented by 23andMe African-American summary statistics. This analysis involved 467 PD cases and 195,120 controls (λ=1.01; **Figure 14**), but did not nominate any genome-wide significant hits.

rsID	Allele1	Allele2	Effect	StdErr	P-value	Direction	HetISq	HetChiSq	HetDf	HetPVal	R2	TYPE	Func.refGene
rs3115534	T	G	-0.4579	0.0775	3.44E-09	--	20.9	1.264	1	0.2609	0.9706	IMPUTED	intronic
rs59025885	T	C	0.4178	0.0761	3.97E-08	++	32.8	1.488	1	0.2225	0.98373	IMPUTED	intergenic

Table 10: Genome-wide significant SNPs identified in the African only GWAS meta-analysis

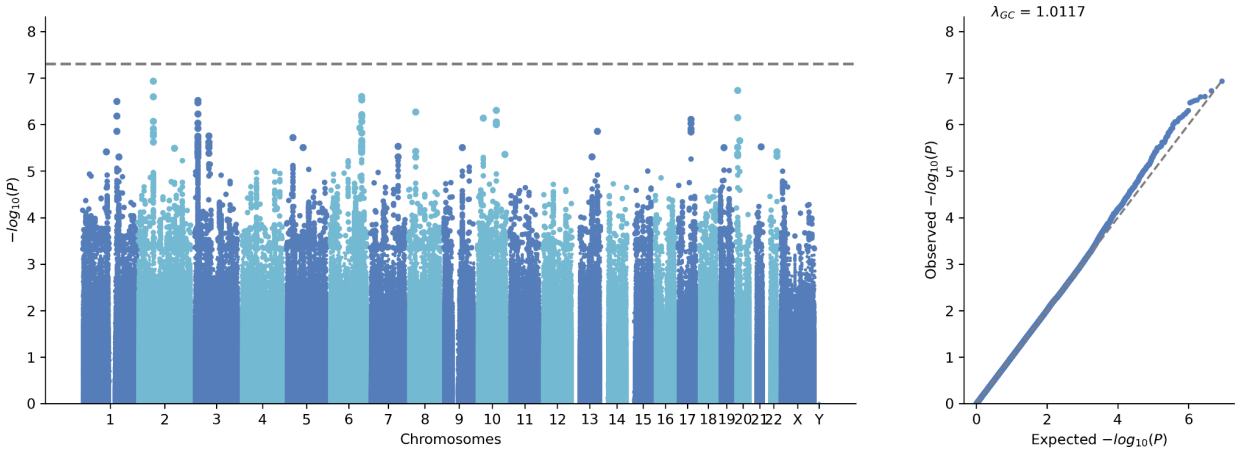


Figure 14: African Admixed Parkinson's disease risk GWAS

Manhattan plot displaying the association statistical significance as –log10(p-value) in the y-axis against chromosomes in the x-axis at a genome scale (Bonferroni correction highlighted at 5E-8). Quantile-quantile plot displaying genetic data distribution; λ GC: Lambda value representing genomic inflation

Subsequently, I performed a comprehensive GWAS meta-analysis incorporating all African and African admixed datasets (**Figure 15**), which totaled 1,488 cases and 196,430 controls. This analysis revealed that 35 SNPs in proximity to *GBA1* were significantly associated with PD risk, demonstrating consistent effect directionality. The two most distant of these SNPs were separated by approximately 639,773 base pairs (**Supplementary Table 19**). Conditional analyses focusing on the top two SNPs indicated a singular causal signal, primarily driven by rs3115534 as the lead SNP (**Figure 15**). Notably, the rs3115534-G allele was more prevalent in African and African admixed populations compared to other groups, with an allele frequency of 0.16 in gnomAD (accessed July 2023) and 0.21 according to the African 1000 Genomes panel. Similar frequencies were observed in the African (cohort MAF = 0.25; affected MAF = 0.33; unaffected MAF = 0.19) and African admixed datasets (cohort MAF = 0.14; affected MAF = 0.22; unaffected MAF = 0.13) used in this study. Furthermore, the rs3115534-G allele frequency was notably higher in Nigerian populations (**Table 11**). Linear regression analysis demonstrated a positive association between the *GBA1* rs3115534 variant and a higher percentage of African ancestry (BETA = -0.001, SE= 0.0005, P = 0.011), suggesting a potential genetic link to ancestry in PD risk.

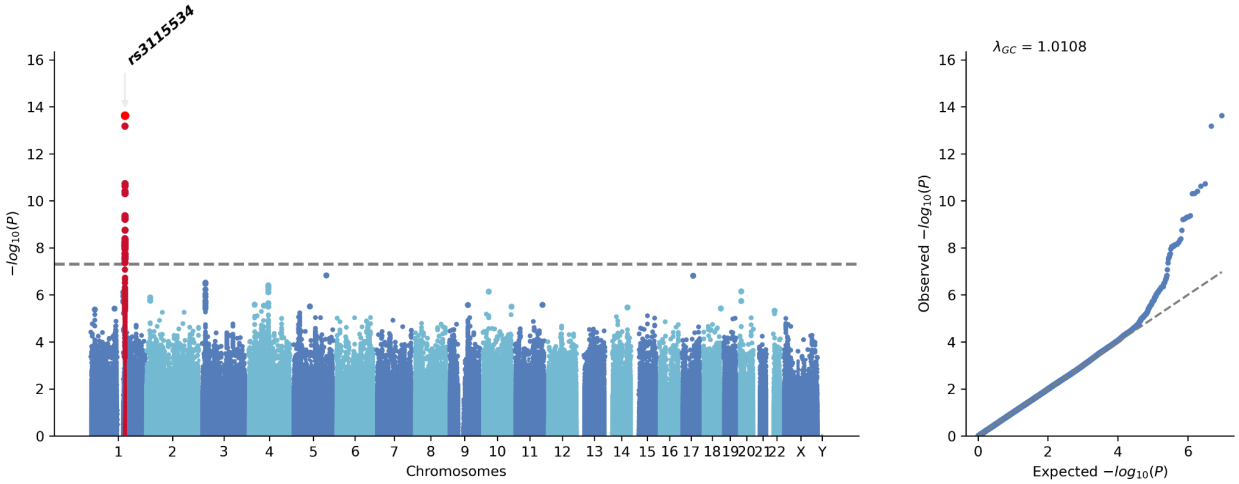


Figure 15: African and African Admixed GWAS Meta-analysis assessing Parkinson's disease risk

Manhattan plot displaying the association statistical significance as –log10(p-value) in the y-axis against chromosomes in the x-axis at a genome scale (Bonferroni correction highlighted at 5E-8). Quantile-quantile plot displaying genetic data distribution; λ GC: Lambda value representing genomic inflation

Population	Allele frequency (count)	Genotype frequency (count)
African Caribbean Black (ACB)	G: 0.198 (38) T: 0.802 (154)	G|G: 0.010 (1) G|T: 0.375 (36) T|T: 0.615 (59)
African Ancestry in Southwest USA (ASW)	G: 0.115 (14) T: 0.885 (108)	G|T: 0.230 (14) T|T: 0.770 (47)

Esan in Nigeria (ESN)	G: 0.303 (60) T: 0.697 (138)	G|G: 0.061 (6) G|T: 0.485 (48) T|T: 0.455 (45)
Gambian in Western Division (GWD)	G: 0.173 (39) T: 0.827 (187)	G|G: 0.044 (5) G|T: 0.257 (29) T|T: 0.699 (79)
Luhya in Webuye, Kenya (LWK)	G: 0.177 (35) T: 0.823 (163)	G|G: 0.061 (6) G|T: 0.232 (23) T|T: 0.707 (70)
Mende in Sierra Leone (MSL)	G: 0.171 (29) T: 0.829 (141)	G|G: 0.024 (2) G|T: 0.294 (25) T|T: 0.682 (58)
Yoruba in Ibadan (YRI)	G: 0.282 (61) T: 0.718 (155)	G|G: 0.074 (8) G|T: 0.417 (45) T|T: 0.509 (55)
Total African (AFR)	G: 0.209 (276) T: 0.791 (1046)	G|G: 0.042 (28) G|T: 0.333 (220) T|T: 0.625 (413)

Table 11: Allele frequencies for *GBA1* - rs3115534 in African and African admixed subpopulations

Variant	Base change	Functional consequence	Genetic variant	Cases with functional variant (n)	Controls with functional variant (n)	rs3115534-GG carriers (n)	rs3115534-GT carriers (n)	rs3115534-TT carriers (n)
chr1:155236249:A:C	A→C	Non-synonymous SNV	Ile320Ser	1	0	0	1	0
rs149487315	C→T	Non-synonymous SNV	Met313Ile	1	0	0	0	1
rs143222798	C→T	Synonymous SNV	Gly277Gly	6	3	0	6	3
rs61748906	A→G	Non-synonymous SNV	Trp136Arg	1	0	1	0	0
rs368786234	G→T	Non-synonymous SNV	Ser77Arg	1	0	0	1	0
rs761621516	GTA→deletion	Non-frameshift deletion	Trp75del (222_224del)	1	0	0	1	0
rs150466109	T→C	Non-synonymous SNV	Lys13Arg	12	8	0	10	10

Table 12: Functional coding variants identified by short-read whole genome sequencing in carriers of the novel *GBA1* rs3115534 variant

Analyses were done in 141 cases and 65 controls. All variants were on chromosome 1, were exonic, and were heterozygous. SNV=single nucleotide variant.

Functional Analysis of GWAS Signal through Whole Genome Sequencing and Splicing Prediction

To explore a potential functional coding variant that might be responsible for the novel GWAS signal but remained undetected through standard genotyping or imputation methods, we conducted short-read WGS analyses on a cohort comprising 206 individuals. This group included 141 PD cases and 65 controls, with 39 individuals being GBA1 rs3115534-GG carriers, 69 carrying the rs3115534-GT genotype, and 98 having the rs3115534-TT genotype. The correlation between the short-read WGS data and the imputed

genotyped data for rs3115534 was remarkably high at 96.6%, thus affirming the reliability and accuracy of the imputed data. However, no discernible differences in coding variation were noted between carriers and non-carriers of the GWAS signal (**Table 12**). To further investigate, we applied the Gauchian algorithm, a specialized variant caller designed for the *GBA1* gene, analyzing WGS BAM files. The Gauchian algorithm is particularly adept at addressing issues arising from the high sequence similarity between *GBA1* and its pseudogene paralog, *GBAP1*. Despite this advanced approach, the algorithm did not uncover any genetic rearrangements that could elucidate the GWAS signal.

Subsequently, ONT WGS long-read sequencing data was generated for a select group of individuals: five rs3115534-GG PD cases, two rs3115534-GT carriers, and six rs3115534-TT controls. This long-read data was then compared to the short-read WGS data for a known structural variant carrier, previously identified in African American populations by Tayebi and colleagues (Tayebi et al. 2000; **Supplementary Figure 10**). Despite this comprehensive analysis, no structural variants that could explain the GWAS signal were detected.

Additionally, I used splice prediction tools available at www.phenosystems.com to assess the potential impact of these variants on normal splicing processes. These tools did not predict any significant alterations in normal splicing patterns, though this requires further validation.

Evaluating the Additive Effect of the *GBA1* Risk Allele and Its Influence on Age at Onset in PD

I focused on determining the additive effects of the *GBA1* rs3115534-G risk allele and exploring its influence on the AAO of PD. To ascertain whether the risk allele's impact was additive, I calculated the frequency of homozygotes for the risk allele and compared it with that of heterozygotes in both PD cases and controls and reported the risk ratios (**Table 13**).

Risk ratios are calculated as follows:

$$RR = \frac{Risk\ of\ disease\ in\ individuals\ with\ the\ variant}{Risk\ of\ disease\ in\ individuals\ without\ the\ variant}$$

This analysis was complemented by a follow-up examination of the association between this specific *GBA1* variant and AAO. Employing linear regression analyses, I assessed 711 cases of African ancestry and 185 cases of African admixed ancestry. These analyses revealed that the GBA1 rs3115534-G allele acts as a modifier of AAO in PD (African ancestry: BETA =-2.004, SE = 0.57, P = 0.0005; African-admixed: BETA = -4.15, SE =0.58, P =0.015; Meta-analysis: BETA =-3.06, SE =0.40, P = 0.008). This implies an earlier PD onset by approximately three years for each additional risk allele (**Figure 16**).

Dataset	Phenotype	Genotype	Counts	Total	Frequency	Risk Ratio
Africans (AFR) in NeuroChip	Cases	G/G	38	265	0.143	1.944
		G/T	120	265	0.453	1.052
		T/T	107	265	0.404	1.228
	Controls	G/G	18	244	0.074	
		G/T	105	244	0.430	
		T/T	121	244	0.496	
Africans (AFR) in GP2	Cases	G/G	92	687	0.134	4.259
		G/T	278	687	0.405	1.209
		T/T	317	687	0.461	1.374
	Controls	G/G	31	986	0.031	
		G/T	330	986	0.335	
		T/T	625	986	0.634	
African Admixed (AAC) in GP2	Cases	G/G	11	183	0.060	3.807
		G/T	61	183	0.333	1.387
		T/T	111	183	0.607	1.226
	Controls	G/G	18	1140	0.016	
		G/T	274	1140	0.240	
		T/T	848	1140	0.744	
African Admixed (AAC) in 23andMe	Cases	G/G	10	288	0.035	1.904
		G/T	85	288	0.295	1.273
	Controls	G/G	3,537	193,985	0.018	
		G/T	44,967	193,985	0.232	

Table 13: rs3115534 Zygosity Information between Cases and Controls across Datasets

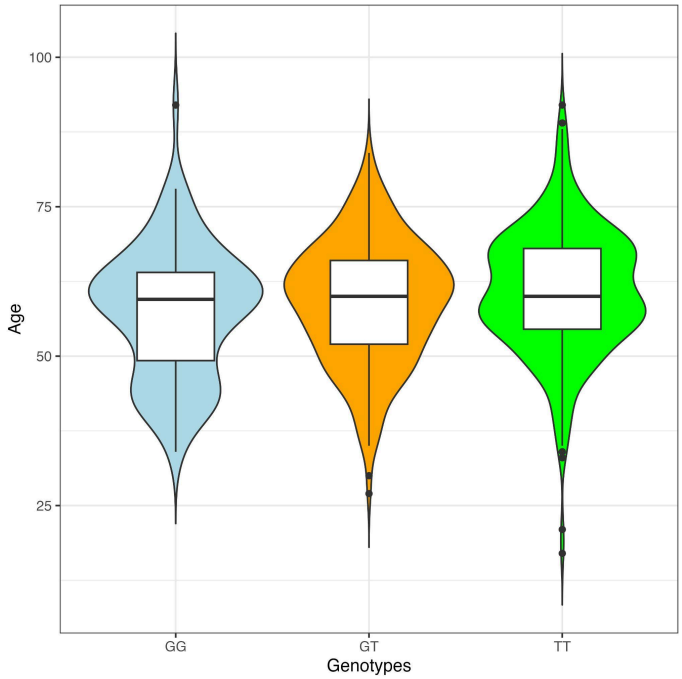


Figure 16: *GBA1* - rs3115534 Genotypes versus age at Parkinson's disease onset

rs3115534-GG versus age at onset; BETA =-1.96, SE = -0.64, P=0.002;
rs3115534-GT versus age at onset; BETA =-2.28, SE =0.85 , P=0.007

Furthermore, to investigate the allele's additive nature, I analyzed individual-level data from IPDGCAN and GP2. This revealed that the rs3115534-GG genotype was 4.26 times more prevalent in African PD cases compared to controls, and 3.81 times more prevalent in African admixed cases. Conversely, the rs3115534-GT genotype was 1.05 times more frequent in African cases and 1.39 times more frequent in African admixed cases. Zygosity analysis of 23andMe data corroborated these findings, indicating a 1.90-fold higher prevalence of rs3115534-GG in African admixed cases.

Moreover, under a dominant model, the OR for African ancestry was 1.74 and for African admixed ancestry was 1.96. Remarkably, these figures closely align with the additive model predictions (African ancestry additive model: OR =1.75; African admixed ancestry additive model: OR =1.95), suggesting an additive rather than a dominant inheritance pattern for this variant. However, caution is advised in interpreting the African-admixed estimates due to the small sample size and the low number of GG carriers. Additionally, no significant differences in AAO were observed between carriers of the *GBA1* rs3115534-GG and GBA1 rs3115534-GT alleles (T-test; P = 0.25).

Comprehensive Analysis of rs3115534-G Variant Across Diverse Ancestral Populations

In my detailed examination of the rs3115534-G variant, I used individual-level data from the GP2 initiative to understand its distribution across different ancestral groups. Interestingly, this variant was not imputed in individuals of European, Ashkenazi Jewish, South Asian, East Asian, and Central Asian ancestries, likely attributed to its low frequency in these populations. In contrast, the rs3115534-G

variant was identified within 230 PD cases and 182 controls of Amerindian and indigenous ancestry, with a MAF of 0.027 and a non-significant p-value of 0.43. My analysis also extended to examining the relationship between the rs3115534-G variant and genomic admixture. Linear regression analysis underscored a positive correlation between the presence of rs3115534-G and a higher percentage of African ancestry (BETA = 0.064, SE = 0.024, P = 0.01). This finding supports the hypothesis of an African founder effect, where the variant is more prevalent in populations with African lineage.

Further insights were gained by analyzing haplotype sizes at the *GBA1* risk locus, particularly around the rs3115534 variant. This analysis included populations of African, African admixed, and European ancestry from the GP2 initiative. Notably, the European haplotypes were longer (average length = 79.19, number of SNPs = 90), compared to the African (average length = 19.30, number of SNPs = 29) and African admixed haplotypes (average length = 15.15, number of SNPs = 22). Such variation in haplotype sizes underscores the genetic diversity across these populations. The rs3115534 variant haplotype was especially prominent within the Nigerian Esan and the Yoruba in Ibadan populations, as indicated by data from the 1000 Genomes Project (**Table 14**). This suggests that the haplotype may have originated in these Nigerian populations, aligning with the concept that founder effects typically result in larger haplotype block sizes due to decreased genetic diversity.

Population	Alleles	Allele Frequencies	Haplotype Count	Haplotype Frequency
Esan (Nigeria)	T	0.697	138	0.697
	G	0.303	60	0.303
Yoruba (Ibadan)	T	0.718	155	0.7176
	G	0.282	61	0.2824
Luhya (Webuye, Kenya)	T	0.823	163	0.8232
	G	0.177	35	0.1768
Gambian (Western Division)	T	0.827	187	0.8274
	G	0.173	39	0.1726
Mende (Sierra Leone)	T	0.829	141	0.8294
	G	0.171	29	0.1706
African Ancestry (SW USA)	T	0.885	108	0.8852
	G	0.115	14	0.1148
African (Caribbean Black)	T	0.802	154	0.8021
	G	0.198	38	0.1979

Table 14: Haplotype frequencies spanning *GBA1* - rs3115534 per African and African admixed subpopulation in 1000 Genomes

In addition to these findings, fine-mapping analyses further reinforced the significance of the rs3115534 variant. The lead SNP, rs3115534, exhibited a posterior probability (PP) of 71.4% for being the causal variant, (**Supplementary Table 17**). Typically a PP of 80%or higher is the benchmark likely to reflect a true causal relationship, meaning that this variant might not be the causal variant in the locus.

Expression Analysis of rs3115534-G

For my analysis, I used the whole blood expression quantitative trait locus (eQTL) summary statistics, as provided by Kachuri and colleagues (Kachuri et al. 2023). These statistics were derived from RNA sequencing data of 2,733 samples, primarily from individuals of African American, Puerto Rican, and Mexican ancestries. A notable eQTL signal was identified for rs3115534, situated 8,821 base pairs from the canonical transcription start site (**Figure 17**; MAF = 0.15; P= 9.99E-25, BETA = 0.238, SE = 0.022). Intriguingly, the rs3115534-G risk allele was associated with an elevation in *GBA1* gene expression levels.

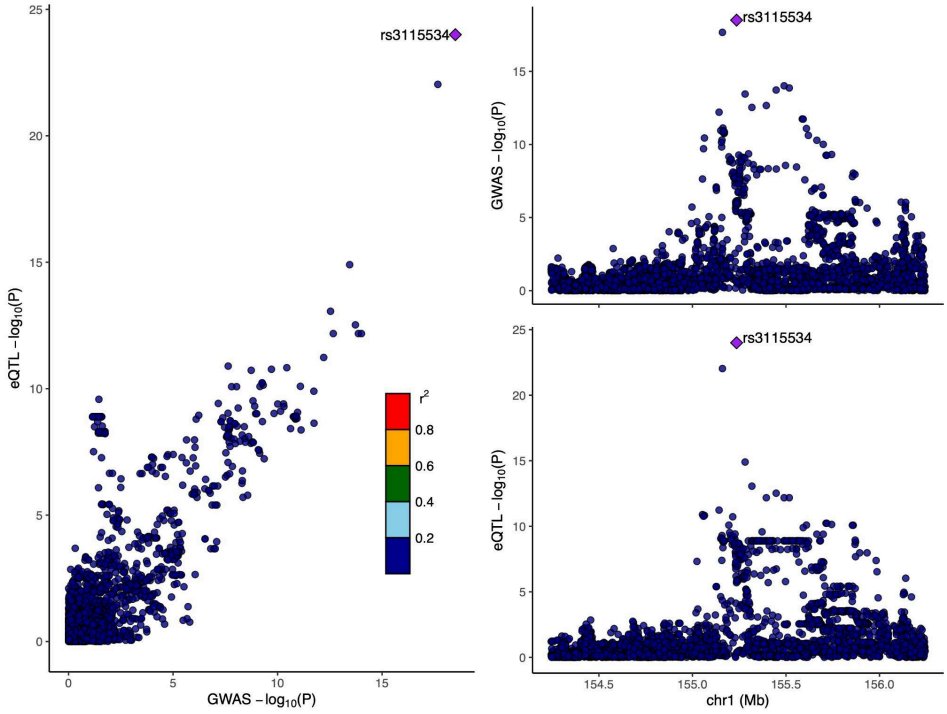


Figure 17: LocusZoom plot displaying African and African Admixed Parkinson's disease GWAS Meta-analysis

Summary statistics versus African expression quantitative trait locus summary statistics from blood (by Kachuri et al., 2023)

To further explore this association, we examined the activity estimates of GCase, the enzyme encoded by the *GBA1* gene, in relation to different genotypes of rs3115534. The analysis indicated a decline in GCase activity in individuals homozygous for the rs3115534-GG risk allele (762.50 ± 273.50 U) compared to heterozygous rs3115534-GT carriers (2743.76 ± 1960.83 U), as demonstrated through a Welch two-sample t-test (GG versus GT; t = -4.3138, df = 21.583, p-value = 0.00029). This trend was also evident when comparing homozygous non-risk allele rs3115534-TT carriers (1879.94 ± 1010.84 U) to

homozygous risk allele rs3115534-GG carriers, where the former group showed higher GCase activity (Welch two-sample t-test: GG versus TT; t = -4.7564, df = 18.363, p-value = 0.00014). These findings suggest a functional relationship between the rs3115534-G allele and *GBA1* expression, potentially influencing the enzymatic activity of GCase (**Figure 18 and Figure 19**).

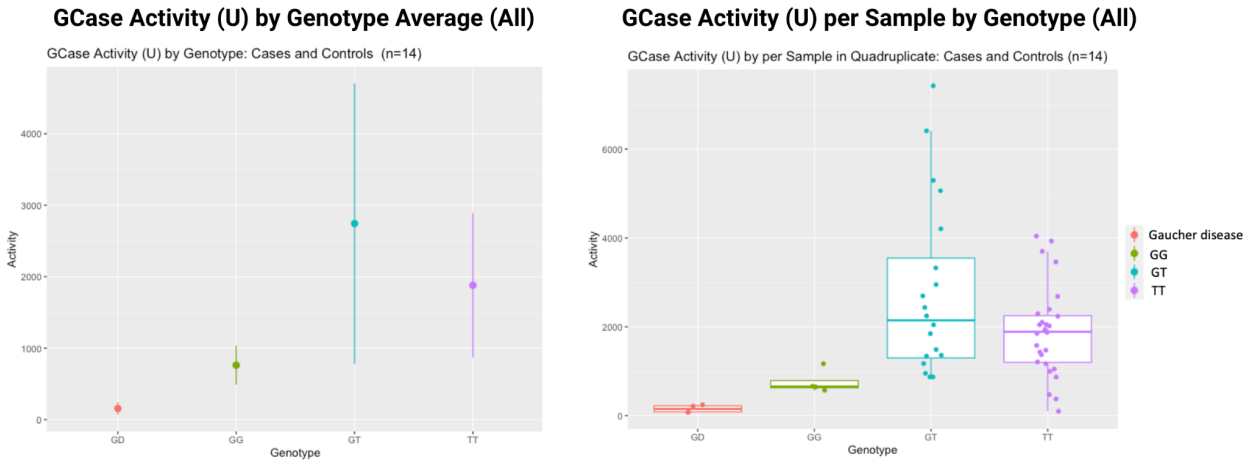


Figure 18: GCase activity analyses performed on *GBA1* - rs3115534-GG, rs3115534-GT, and rs3115534-TT carriers

A fluorometric 4-MU assay was used to measure GCase activity in 14 lymphoblastoid cell lines, including a type I Gaucher disease (GD) patient as a positive control. On the left, samples were aggregated by rs3115534 genotype and average activity. Values are represented by mean and standard deviation. On the right, all 14 samples were run in quadruplicate. Samples were screened for known *GBA1* pathogenic mutations that could bias these estimates. A total of two carriers (one heterozygous for *GBA1* p.I320S and one heterozygous for *GBA1* p.T75del) were removed from further analyses.

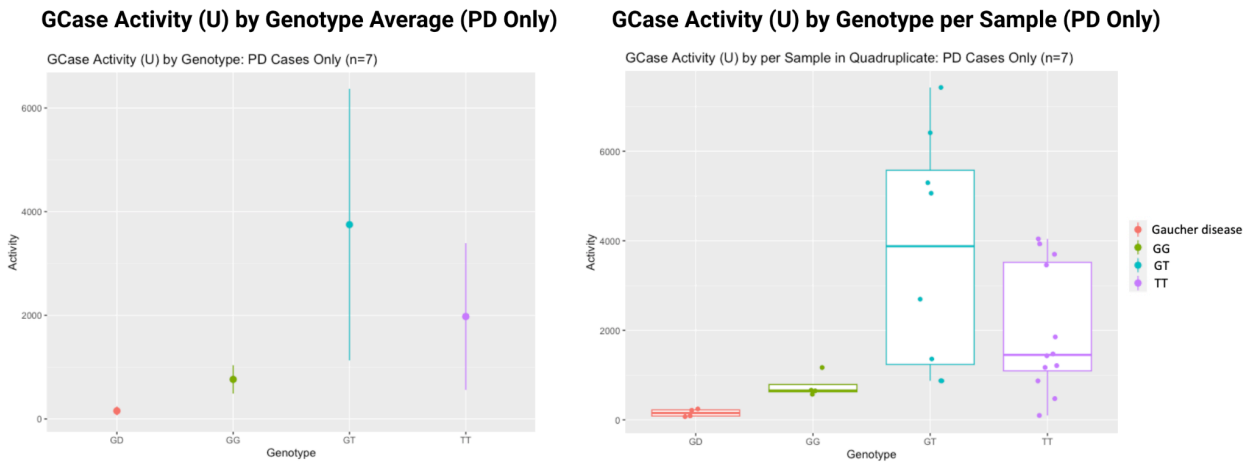


Figure 19: GCase activity analyses performed on GBA1 - rs3115534-GG, rs3115534-GT, and rs3115534-TT carriers

Samples with Parkinson's disease were pulled from the 14 samples, including a type I Gaucher disease (GD) patient as a positive control. On the left, a total of 9 samples with Parkinson's disease were aggregated by rs3115534 genotype and average activity. Values are represented by mean and standard deviation. On the right, all 9 samples with Parkinson's were run in quadruplicate (Welch Two Sample t-test: GG versus GT; t = -3.189, df = 7.3002, p-value = 0.0.01446; GG versus TT; t = -2.8158, df = 13.003, p-value = 0.01458; GT versus TT; t = 1.7509, df = 9.7545, p-value = 0.1113). A total of two carriers (one heterozygous for *GBA1* p.I320S and one heterozygous for *GBA1* p.T75del) were removed from further analyses.

Comparative Analysis of *GBA1* Locus Variants Across Diverse Populations

In my analysis aimed at unraveling the unique signal found in the *GBA1* locus, I compared effect estimates and the directionality of effects using summary statistics from the multi-ancestry PD GWAS meta-analysis which included European, Latin American, and East Asian populations (J. J. Kim et al. 2022). The rs3115534-G allele, central to my study, exhibits notably low allele frequencies in European (0.0015), East Asian (0.0005), South Asian (0.0017), and Ashkenazi Jewish populations (0.0009), as per the gnomAD database.

Further investigation into the European data from the GP2 initiative revealed that the rs3115534 variant was inadequately imputed in 13,186 samples (R^2 = 0.16, MAF = 0.009). This finding underscores a significant difference in the *GBA1* locus between African, African admixed, and European populations. In Europeans, the association with PD risk is predominantly driven by two independent signals: rs35749011 (*GBA1* E326K) and rs76763715 (*GBA1* N370S), both of which are exceedingly rare in individuals of African and African admixed ancestry (**Figure 20B**). In contrast to the African populations, the *GBA1* locus in East Asians presents a different scenario. Here, the rs3115534 variant was not imputed in the largest East Asian GWAS meta-analysis, highlighting population-specific genetic variations (**Figure 20C**). However, when examining Amerindian and indigenous populations, which possess higher levels of African admixture, the differences are less pronounced. This is evident from studies such as the Loesch et al. GWAS and the Amerindian and indigenous GWAS conducted by 23andMe. In these populations, the rs3115534-G allele demonstrated odds ratios of 1.13 (95% CI =0.41-1.86, P= 0.72) and 1.56 (95% CI = 1.55-1.88, P= 0.01) respectively, suggesting a stronger presence and potential impact of this allele compared to European and East Asian populations (**Figure 20D**).

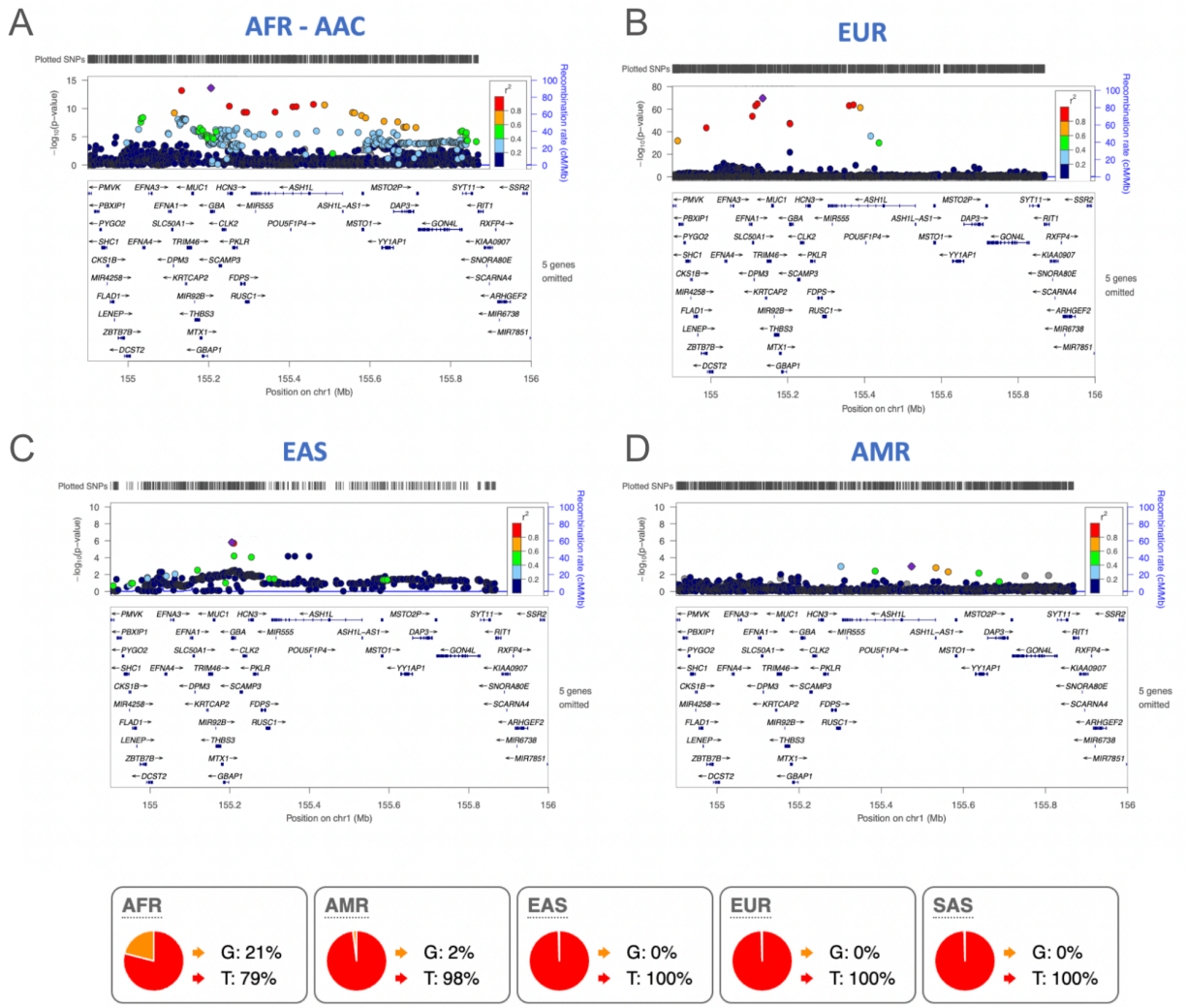


Figure 20: LocusZoom plots of *GBA1* in (A) AFR/AAC, (B) EUR, (C) EAS, (D) AMR populations

AFR/AAC: African and African admixed; EUR: European; EAS: East Asian; AMR: Amerindian and Latino Indigenous populations

The most extensive GWAS and multi-ancestry GWAS meta-analyses conducted to this date have collectively identified 104 independent significant PD risk variants (Nalls et al. 2019b; Foo et al. 2020a; (J. J. Kim et al. 2022). Among these variants, 91 passed through rigorous QC, met imputation criteria, and were present in the GWAS meta-analysis focusing on African and African admixed populations (**Figure 21 and Table 12**). Within this subset of 91 variants, 16 were found to be nominally significant ($p < 0.05$), as per the African and African admixed meta-GWAS results reported in this study (**Supplementary Table 18**). While these findings contribute valuable insights, it is crucial to acknowledge a potential limitation in my study. The data from 23andMe, which formed a part of my analysis, was also used in the multi-ancestry GWAS meta-analysis by Kim and colleagues. This overlap in data sources could potentially bias the estimates derived in my study.

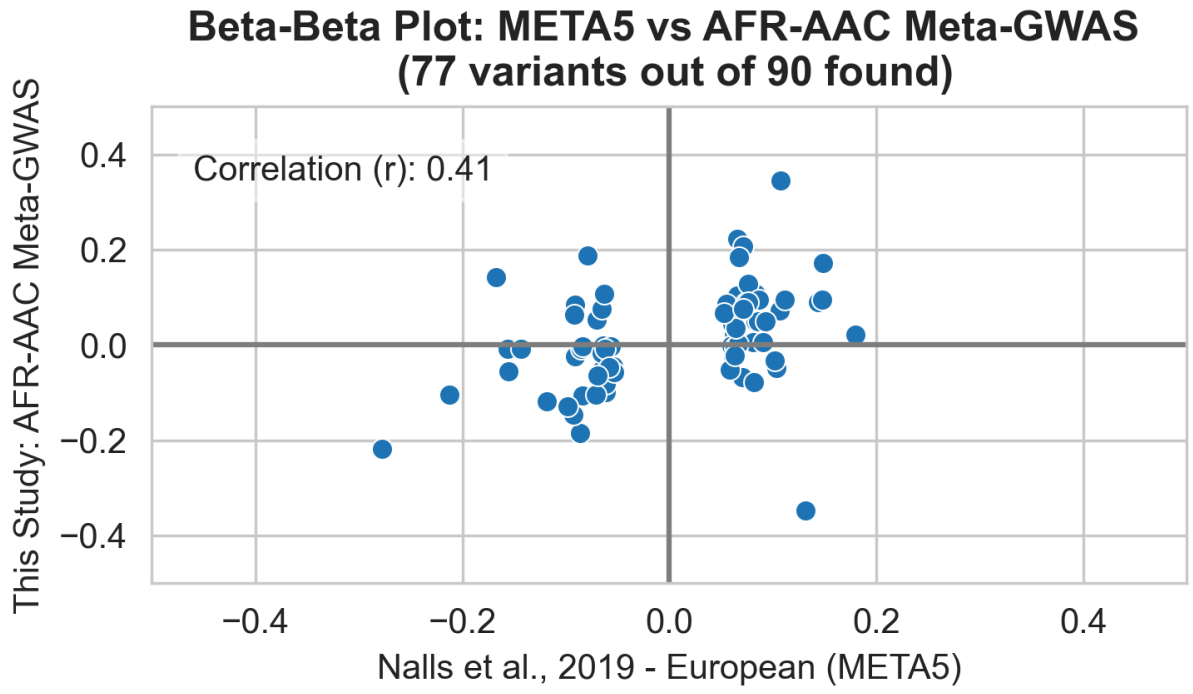


Figure 21: Beta-beta plot comparison of African versus African Admixed estimates for PD known risk loci identified in Europeans

Comparative Analysis of Polygenic Risk Scores Across African and African Admixed Populations

I conducted an in-depth analysis of 90 SNPs previously linked with PD risk in European populations (Nalls et al. 2019b; Foo et al. 2020a). Within the African admixed individual-level dataset, 86 out of these 90 risk SNPs passed the QC criteria. The PRS, based on European data, was able to predict PD disease status in individuals of African admixed ancestry with a notable odds ratio (OR=1.43; 95% CI =1.26-1.61, P=4.37E-05; **Figure 22A**). Subsequently, I calculated the PRS for the African individual-level dataset, where 79 out of the 90 risk loci passed QC. This African-derived PRS also showed predictive ability for disease status in PD versus healthy controls of African ancestry (OR=1.27; 95% CI =1.16-1.38, P=1.05E-05; **Figure 22B**). A marginally higher effect was observed in the African admixed ancestry PRS model, which could be attributed to either the greater number of SNPs passing QC or a higher proportion of European admixture within this cohort. Notably, the variant with the most significant effect size in both models was chr4:89704960:G:A_A (rs356182), located at the *SNCA* locus. Even after adjusting for rs356182, PRS differences between PD and controls remained statistically significant ($PRS_{\text{African admixed}}$ P =2.21E-05; PRS_{African} P =0.014). However, as anticipated, these effects were consistently lower compared to those observed in Europeans (PRS $OR_{\text{training dataset}}$ =3.74, 95% CI =3.35 –4.18), hinting at the possibility of additional novel genetic loci influencing PD heritability in African and African admixed populations.

Further analysis was conducted using effect estimates from the AAC 23andMe summary statistics for the same 90 SNPs. In the combined African and African admixed individual-level datasets from GP2, 77 of

the 90 risk SNPs passed QC. The PRS, based on these AAC statistics, predicted PD disease status in African admixed ancestry individuals (OR=1.26; 95% CI =1.15-1.37; P=1.89E-05; **Figure 22C**). Similarly, when calculating PRS for African individual-level data, a predictive relationship between PD and healthy controls was observed (OR=1.42; 95% CI =1.25-1.60; P=6.65E-05; **Figure 22D**). Intriguingly, when applying the AAC 23andMe summary statistics, the PRS effect size was slightly more pronounced for the African individual-level data compared to using the European summary statistics as the reference. This observation underscores the importance of considering ancestry-specific genetic factors in the assessment and understanding of PD risk.

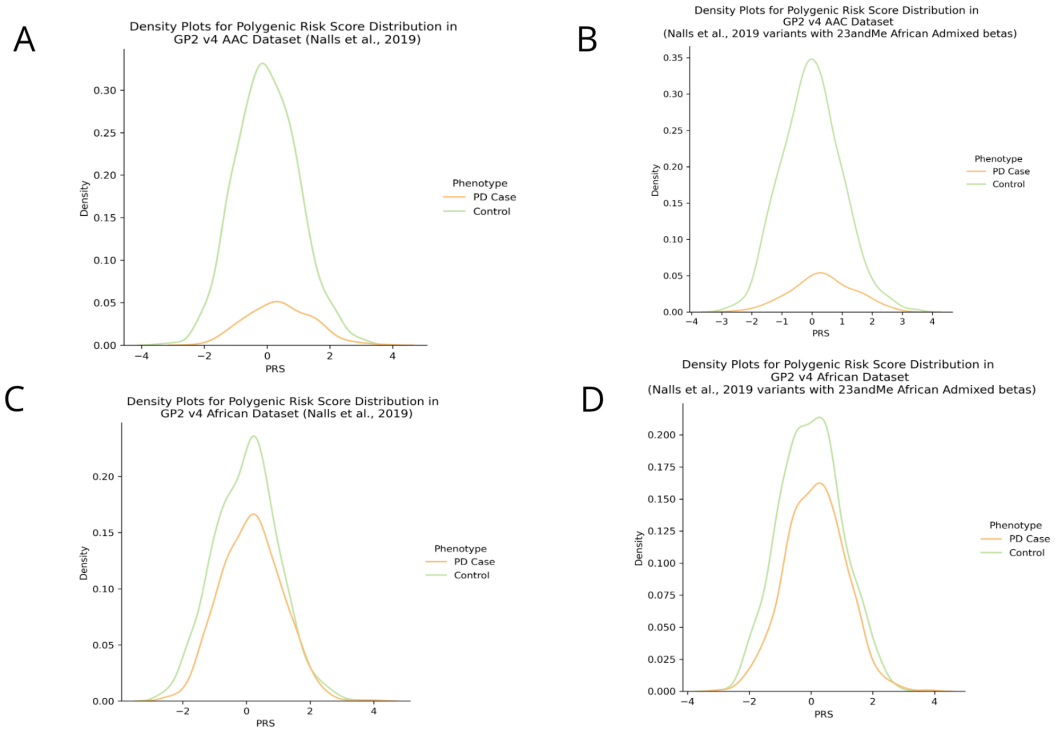


Figure 22: Density plots showing polygenic risk score distributions in the African and African Admixed individuals using the 90 Parkinson's disease risk loci

A) Nalls et al., 2019 reference estimates on African admixed individual level data; B) 23andMe African admixed reference estimates on African admixed individual level data; C) Nalls et al., 2019 reference estimates on African individual level data; D) 23andMe African admixed reference estimates on African individual level data

Runs of Homozygosity

In this research, while we did not detect any enrichment of specific ROH spanning the *GBA1* locus, a notable observation was made regarding the length of ROH at a genome-wide level. We found that PD cases exhibited longer ROH compared to controls. Specifically, an increase in the average number of ROH segments by 1Mb was associated with an increased risk of PD, with an OR of 1.08 (OR = 1.08; 95% CI = 1.066- 1.11; P = 4.63E-16). Additionally, the total size of these homozygosity regions also differed

significantly between cases and controls, albeit with a relatively small effect (BETA = 1.69E-05 per kb increase; SE = 3.531E-06; P = 1.53E-06).

These findings suggest the potential involvement of as-yet-undiscovered recessive variants in the etiology of PD in the studied populations. The observed association between longer ROH and increased PD risk implies that these regions might harbor additional genetic variants contributing to the disease's development.

Conclusions and Discussion

Expanding Research in African and African Admixed Populations

PD genetics within African and African admixed populations has historically been an underexplored area in medical research. Prior studies, often constrained by small sample sizes with typically fewer than thirty participants (Ross et al. 2010; McGuire et al. 2011; Clark et al. 2003; Gwinn-Hardy et al. 2001; N. U. Okubadejo et al. 2022). While these efforts have provided valuable insights, they have been limited in offering a comprehensive understanding of PD genetics in these specific populations due to their scale.

The current study marks a significant advancement in this field, establishing the largest collection of PD patients and controls from African and African admixed ancestries to date. This large-scale effort has enabled a thorough genome-wide assessment of PD genetics, leading to the identification of a novel African-specific GWAS signal at the *GBA1* locus. This signal, significantly associated with both PD risk and AAO, has emerged as a pivotal risk factor in these populations (**Figure 23**).

In a refined analysis, I focused exclusively on samples predicted to be of African ancestry and recruited from Nigeria. This subset comprised 244 controls and 265 cases genotyped using the NeuroChip array, along with 638 controls and 649 cases genotyped on the enhanced NeuroBooster platform. rs3115534 exhibited significant associations with disease status in both cohorts. Among Nigerian individuals genotyped on NeuroChip, the MAF was 0.2889 in controls and 0.3698 in cases (p = 0.006155). The association was even more pronounced in the NeuroBooster cohort, with MAFs of 0.2265 and 0.3374 in controls and cases, respectively (p = 4.057E-10). This discovery underscores the importance of including diverse ancestry groups in genetic studies, particularly as a substantially larger sample size was needed to identify *GBA1* as a major PD risk factor in European populations through GWAS (Simón-Sánchez et al. 2009).

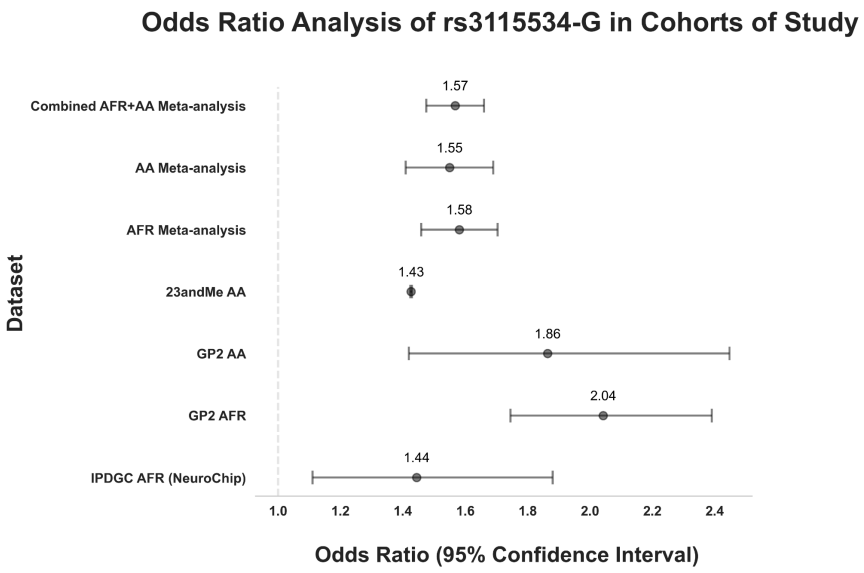


Figure 23: Comparative Odds Ratio Analysis Across Different Cohorts for rs3115534-G Variant in Parkinson's Disease Studies

AFR: African; AA: African admixed; IPDGC: International Parkinson's Disease Genomics Consortium

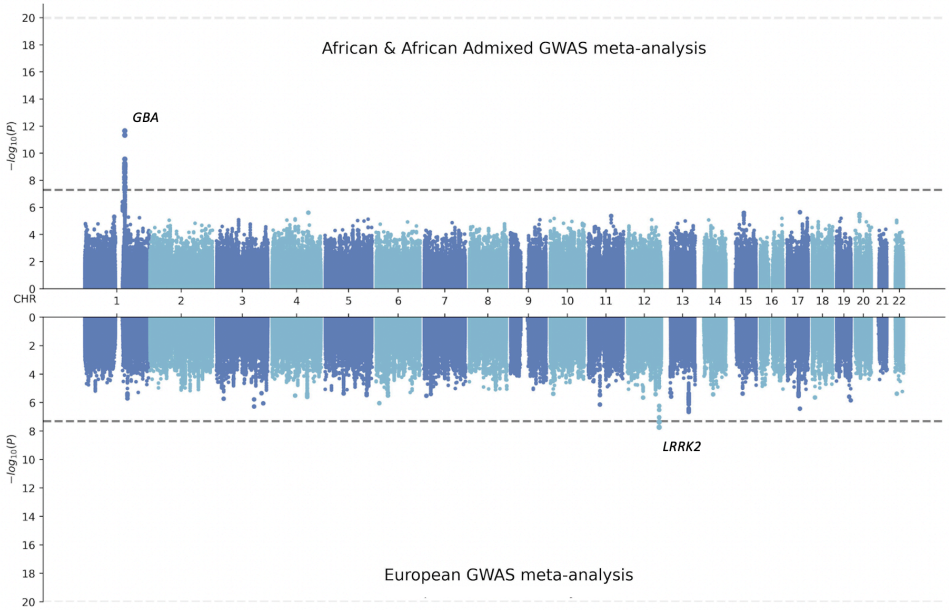
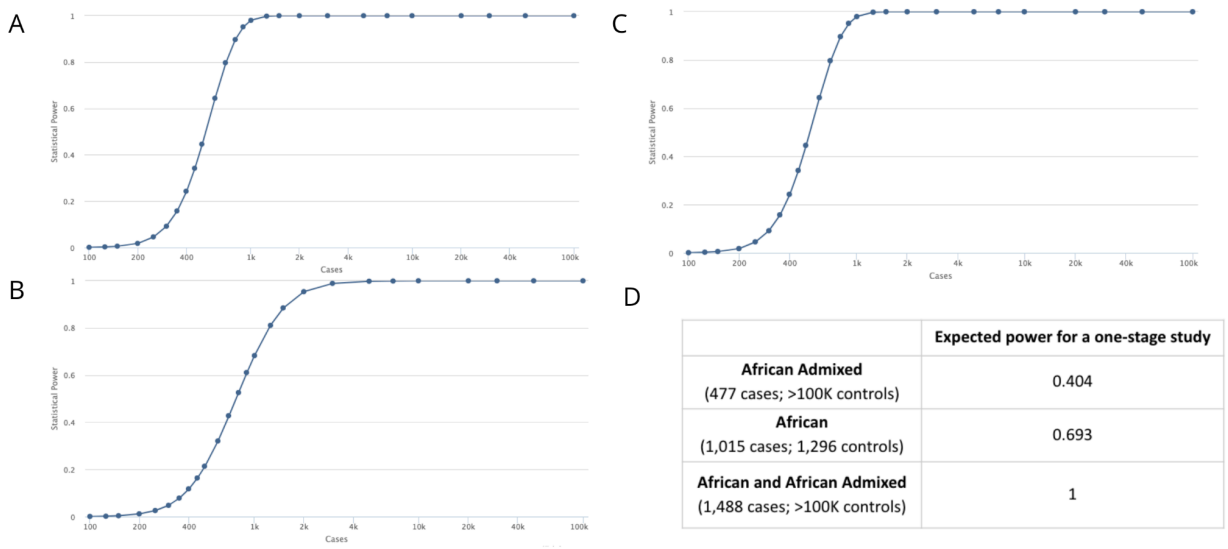


Figure 24: Miami Plot comparing European versus African and African admixed GWAS meta-analysis

Assessing Parkinson's disease risk that are similarly randomly sampled (n_{cases} = 1,200; $n_{controls}$ = 2,445). Randomly sampled 1,200 cases and 2,445 controls from European, African, and African admixed individuals for this analysis to illustrate that at this scale we are only powered to detect SNPs with the greatest association to increased disease susceptibility. The current European meta-analysis has 90 independent genome-wide significant risk signals in 78 genomic regions identified by Nalls and colleagues.

I identified an association between the variant rs3115534 and the AAO of PD. Notably, the genetic composition at the *GBA1* locus shows significant differences between African and African admixed populations compared to Europeans. In Europeans, the risk of disease is influenced by two distinct signals: rs35749011 (*GBA1* E326K) and rs76763715 (*GBA1* N370S) (**Figure 24**). However, considering this study's limited sample size, we lacked statistical power to detect common genetic variants of smaller effect sizes (**Figure 25**).



	Expected power for a one-stage study
African Admixed (477 cases; >100K controls)	0.404
African (1,015 cases; 1,296 controls)	0.693
African and African Admixed (1,488 cases; >100K controls)	1

Figure 25: Power calculations for the meta-GWAS in African and African admixed populations

Calculated using the Genetic Association Study (GAS) power calculator provided by the University of Michigan under an additive model. OR of 1.58 for rs3115534; average of 19.5% disease allele frequency, and a 0.6% estimated prevalence in the African populations (according to Global, regional, and national burden of Parkinson's disease [Lancet Neurology 2018])
(A) African admixed (477 cases; >100K controls); (B) African (1,015 cases; 1,296 controls); (C) African and African admixed (1,488 cases; >100K controls)

Novel Mechanisms and RNA Challenges at the *GBA1* Locus

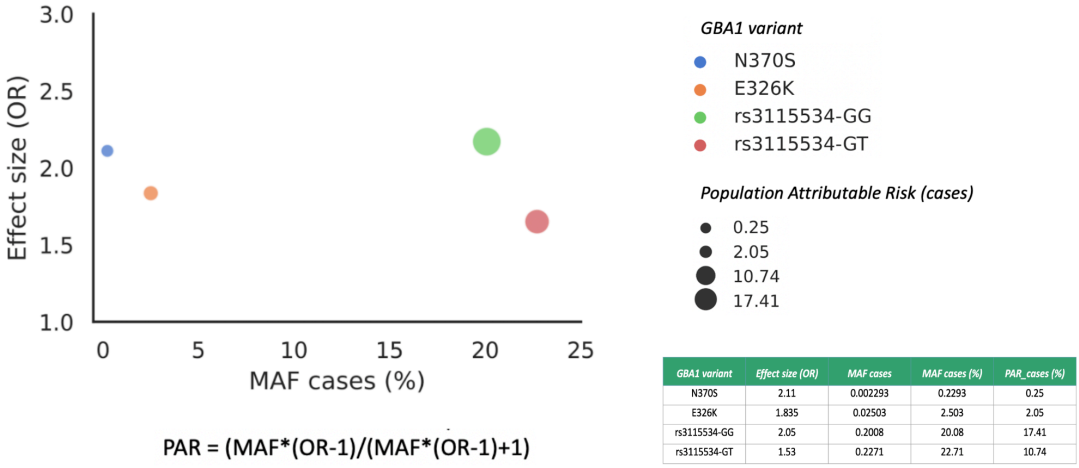
GBA1 is a genetically complex locus, exhibiting a variety of coding, structural, and non-coding variants that impact PD risk differently (A. Singleton and Hardy 2011). Despite the robust effect size of the signal identified in this study, no associations were found with previously reported or newly discovered *GBA1* coding or structural aberrations that could explain this signal (Toffoli et al. 2022; Park et al. 2001; Tayebi et al. 2000; Mahungu et al. 2020). This suggests that the mechanism underlying this association might differ from previously identified pathways. Using eQTL data primarily from African American ancestry, the study found the rs3115534-G risk allele to be associated with increased GBA1 expression levels in whole blood. However, this increase paradoxically corresponds with a trend toward decreased GCase activity, potentially due to RNA sequencing challenges at this locus or the presence of multi-mapping reads between *GBA1* and its pseudogene, *GBAP1*, which are often discarded in standard processing and do not contribute to gene-level quantification of expression in many publicly available datasets like GTEx (https://gtexportal.org/). This rs3115534-G allele might increase the expression of a non-functional

transcript, leading to decreased levels of the transcript responsible for optimal production of the protein isoform with GCase activity, a possibility supported by common transcript diversity (Gustavsson et al. 2023). This might be novel disease mechanism via expression changes consistent with a trend towards decreased GCase activity levels. The *GBA1* c.1225-34C>A (rs3115534) GWAS hit alters a non-conserved intronic nucleotide (GERP++ score = -2.04). Splice prediction tools predicted no significant impact on normal splicing, while rs3115534 has been reported to be an eQTL in several tissues (Raj et al. 2014)4).

Additionally, a large-scale pQTL study in African Americans with chronic kidney disease suggests that at the protein level the risk allele for PD in our GWAS (rs3115534-G) is associated with a reduction in the level of GCase protein in blood, as defined by the SOMAscan assay. This finding supports the concept that the risk allele leads to a partial loss of both GCase protein and GCase enzyme activity ("Identification of 969 Protein Quantitative Trait Loci in an African American Population with Kidney Disease Attributed to Hypertension" 2022). This study's large-scale genome-wide association evaluation of PD risk in individuals of African ancestry has identified the intronic *GBA1* variant, rs3115534, not previously linked with PD risk. Despite the absence of coding or structural variants in short- and long-read sequencing analyses, fine-mapping prioritized this variant with a posterior probability of over 70%. The frequent occurrence of rs3115534 in West African populations suggests a possible founder effect, highlighting the significance of ancestral diversity in genetic studies and the potential for novel RNA-based therapeutic interventions targeting lifetime PD risk reduction.

Implications of the rs3115534-G Allele and Ancestral Genetic Diversity

The high population frequency of the rs3115534 signal and the characteristics of homozygous carriers suggest that this variant does not cause Gaucher disease. Its rarity in non-AfricanaAfrican admixed populations suggests an African founder effect, indicating distinct genetic architectures of the *GBA1* locus across populations. Comparative analyses with the largest PD GWAS meta-analyses of European (Nalls et al. 2019b), Latin American (Loesch et al. 2021), and East Asian (Foo et al. 2020a) populations revealed the rs3115534-G allele to be exceedingly rare, further emphasizing the importance of considering ancestry-specific genetic factors in PD research. Additionally, the PAR comparison for *GBA1* known coding variants in the European and Ashkenazi Jewish populations compared to the novel *GBA1* intronic variant identified in the African population in this study demonstrates that while p.N370S has the highest effect size on PD risk, the rs3115534-GG variant, while less frequent, contributes a larger PAR (**Figure 26**).



GBA1 variant	Effect size (OR)	MAF cases	MAF cases (%)	PAR_cases (%)
N370S	2.11	0.002293	0.2293	0.25
E326K	1.835	0.02503	2.503	2.05
rs3115534-GG	2.05	0.2008	20.08	17.41
rs3115534-GT	1.53	0.2271	22.71	10.74

Figure 26: Population attributable risk comparison for *GBA1* known coding variants in the EUR population versus the novel *GBA1* intronic variant in the AFR population

Polygenic Risk Scores and Genetic Contributors to PD

PRSs were used to predict disease status between PD cases and healthy controls of African and African admixed ancestry. The PRSs showed slightly better prediction using African admixed summary statistics on African individual-level data. rs356182, the lead risk variant at the *SNCA* locus in European and Amerindian and indigenous ancestry studies, was identified as a major genetic contributor in both models. This aligns with the partial overlapping directionality of effect observed between PD known risk alleles predisposing to disease in European and African populations (Nalls et al. 2019b; Loesch et al. 2021).

Limitations and Future Perspectives

Despite the insights provided, the study faces several limitations, including the need for a larger cohort to explore susceptibility genetic risk and phenotypic relationships comprehensively. The absence of well-powered, African-specific RNA sequencing datasets and the challenge posed by multi-mapping reads between *GBA* and *GBAP1* complicate the full elucidation of the potential novel functional mechanism suggested by the study's findings. Additionally, while I used ancestry-matched controls for the analyses, the predominance of Nigerian cases does not fully represent the genetic diversity across the African continent.

In conclusion, this study is the first GWAS-based insights into the genetics of PD in African and African admixed populations. This comprehensive genome-wide assessment examines PD risk and age at onset, delving into ancestry-specific risk profiling, haplotype structures, and genetic admixture. Using the unique genetic makeup of these populations, I have uncovered a novel association signal in the *GBA1* gene, which encodes the lysosomal enzyme GCase. This association has led to the revelation of a novel disease mechanism linked to decreased glucocerebrosidase activity levels, which correlates with an increased risk of PD. This study offers crucial insights into the construction of African ancestral

haplotypes and potential novel pathogenic mechanisms underlying PD etiology. The GP2 initiative, under which this study was conducted, has generated key data to understand PD's molecular mechanisms. These findings could pave the way for future clinical trials and therapeutic interventions targeting *GBA1* and other PD-related genes. Future studies should aim to expand cohort sizes, integrate *omics data, and evaluate population-specific associations to enhance our understanding of PD genetics and its clinical implications.

Chapter 5: Integrating Multi-Omics with Machine Learning for Early Diagnosis Predictions

Overview and Broader Relevance

This study addresses the critical need for global, scalable, early, and precise PD diagnosis, which is most effective when applied during the disease's early stages. Using a data-driven approach, this study employs cost-effective methodologies to diagnose PD proactively, often before patients manifest noticeable symptoms. Central to this work is the use of multi-modal PD datasets and the creation of a custom Python package, GenoML, both integral to the development of predictive models for PD diagnosis.

The study leverages data from the first release of the Accelerating Medicines Program - Parkinson's Disease (AMP-PD) and uses GenoML's open-source automated ML software for analysis. This approach ensures reproducibility, transparency, and adherence to open science principles. GenoML's capabilities are harnessed to create and refine peri-diagnostic models that effectively predict PD risk. These models nominate key features that are then used to construct unbiased genetic networks. These networks not only elucidate the biological pathways involved in PD onset but also highlight potential therapeutic targets.

The resultant model from this study marks a significant improvement over previous efforts in this domain. It demonstrates enhanced performance metrics, both in current cross-validation on withheld samples and, in some instances, surpasses the training phase metrics of prior studies. This advancement in PD risk prediction models underscores the potential of data-driven, machine learning approaches in medical diagnostics.

This work has been published here:

> **Makarious MB**, Leonard HL, Vitale D, Iwaki H, Sargent L, Dadu A, Violich I, Hutchins E, Saffo D, Bandrés-Ciga S, Kim JJ, Song Y, Maleknia M, Bookman M, Nojopranoto W, Campbell RH, Hashemi SH, Botia JA, Carter JF, Craig DW, Keuren-Jensen KV, Morris HR, Hardy JA, Blauwendraat C, Singleton AB, Faghri F, Nalls MA: "**Multi-Modality Machine Learning Predicting Parkinson's Disease**" *npj Parkinson's disease (2022)*; https://doi.org/10.1038/s41531-022-00288-w

The accompanying package created and distributed for this work, GenoML, is available on arXiv:

> **Makarious MB**, Leonard HL, Vitale D, Iwaki H, Saffo D, …, Singleton AB, Nalls MA, Faghri F: **"GenoML: Automated Machine Learning for Genomics"** *arXiv (2021); https://arxiv.org/abs/2103.03221*

GenoML is available on GitHub, which, at time of writing, has ~16,000 downloads at time of writing:

> https://github.com/GenoML/genoml2

Introduction

Within the challenging landscape of progressive neurodegenerative disorders, the critical need for timely and precise diagnosis is particularly pronounced in PD. This urgency is driven by the imperative to develop and apply novel therapeutic interventions at a juncture where the disease is most susceptible to treatment. Our research significantly contributes to this goal through the adoption of economical and data-centric methodologies for early detection. This approach is fundamentally designed to detect, analyze, and manage the disease proactively, ideally before patients exhibit noticeable symptoms (Nalls et al. 2015; Paulsen et al. 2013).

In this endeavor, I employed GenoML, an innovative, open-source automated machine learning (auto-ML) tool I developed, to perform extensive analyses of multimodal genomic and clinical data, thereby making such analyses more accessible and widespread (Makarious et al. 2021). Aligning with the National Human Genome Research Institute's strategic vision, our work contributes to the anticipated routine integration of epigenetics and transcriptomics into models that predict phenotypic outcomes from genotypic data by 2030 (Green et al. 2020). This initiative is further propelled by remarkable strides in the automation of ML processes and the proliferation of extensive datasets encompassing clinical, demographic, and genomic information. These developments are pivotal in enhancing early disease detection and remotely pinpointing individuals at elevated risk of PD (Sudlow et al. 2015).

It is noteworthy that the accuracy of initial clinical diagnoses in PD is around 80%, even when confirmed pathologically (Rizzo et al. 2016). Traditional approaches in biomarker research for neurodegenerative diseases have predominantly used conventional statistical techniques and linear models. However, current research has branched into diverse methods using ML, investigating modalities like cerebrospinal fluid biomarkers, imaging techniques, RNA analysis, and movement metrics, including data derived from wearable sensors (Prashanth et al. 2016; D. A. Lee et al. 2021; Mei et al. 2021; Chen-Plotkin 2018; Uehara et al. 2021; Noyce et al. 2014; Palmerini et al. 2017). Our research seeks to push the boundaries in this area, focusing on developing predictive models grounded in nonlinear machine learning techniques. These models use cost-effective and readily available data, either remotely collected or derived from existing biobanked sources, thereby eliminating the need for further clinic visits or expensive protocols.

In our study, we used a variety of publicly accessible, multimodal PD datasets alongside GenoML. By comparing the top ML algorithms as outlined in the 2020 Kaggle executive summary, I developed an effective peri-diagnostic model for assessing PD risk. This methodology also aided in constructing unbiased gene networks, shedding light on biological pathways and identifying potential therapeutic targets linked to the onset of PD. Our models, rigorously validated against publicly available datasets, have demonstrated enhanced performance metrics in comparison to prior studies, thereby making a significant contribution to the realms of clinical trial recruitment and drug development in PD research (Nalls et al. 2015; Makarious et al. 2021).

In conclusion, by harnessing the extensive availability of varied datasets and leveraging advancements in ML, our research introduces a sophisticated and cost-effective model for the early detection of PD. This model not only enhances the accuracy of disease risk prediction but also provides vital biological insights, highlighting its substantial potential to influence PD research and improve patient care strategies.

Methods

GenoML

I have been closely involved in the development of GenoML, a Python open-source autoML platform. This innovative tool, which I have had the privilege to design and expand on during my studentship, is specifically designed to streamline and enhance the analysis of multifaceted genomics and clinical data. My team and I aimed to make GenoML not only powerful but also accessible, democratizing the complex field of genomics analysis. We drew upon a carefully selected list of the top dozen ML algorithms, as identified in the 2020 executive summary by Kaggle (Makarious et al. 2021). This list, refined through expert review, ensures a diverse and effective set of methods for supervised prediction in biomedical research. We have made GenoML available on its official website and GitHub, offering detailed documentation and insights into its functionalities.

In my work, I have emphasized GenoML's role as more than just an autoML Python package. It is a solution designed to navigate the complexities of genomics, integrating genetics with multi-omics data analysis under an open science framework. Our mission with GenoML has always been to make machine learning for genomics and clinical data accessible even to those without extensive expertise in the field. We have achieved this by automating the entire development, evaluation, and deployment process, thereby simplifying tasks that traditionally require deep knowledge in both machine learning and genomics. The GenoML package encapsulates not just data processing and analysis, but also the tuning, validation, and interpretation of results, taking into account the unique biological aspects and limitations of the underlying data collection, protocols, and technology.

Our aim was to bridge the growing gap between the demand for machine learning expertise and the current supply, especially in the specialized area of genomics. In GenoML, I combine the use of software geneticists are familiar with such as PLINK, with Python’s scikit-learn package, and implement custom Python scripts for the pre-processing of the data. PLINK is commonly used in the field of genetic data analysis for its efficiency and versatility in managing large-scale genomic data. It allows users to perform a wide range of complex genetic analyses, including but not limited to: association studies, population stratification, and haplotype block identification (Purcell et al. 2007). By integrating PLINK, GenoML is equipped to handle the intricacies of genetic data in a similar manner that a geneticist would in their own analyses, providing defaults for new users. Complementing PLINK's functionality, I incorporated Python’s scikit-learn package, a powerful tool for machine learning. Scikit-learn provides a comprehensive array of algorithms and models, including classification, regression, clustering, and

dimensionality reduction, among others (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, and Others 2011). Its integration into GenoML allows the application to leverage and auto-compete advanced machine learning techniques, thereby enhancing its capability to identify patterns and make predictions based on genetic data. This gap has been previously addressed by other user-friendly software, yet they often fall short when it comes to genomics data, which require significant domain expertise for tasks such as data cleaning, preprocessing, and quality control. By providing an end-to-end framework for genomic datasets, GenoML facilitates not only the more straightforward aspects of ML workflows, like model training and tuning, but also the more intricate processes such as data preprocessing and cleaning.

In conclusion, GenoML represents an advancement in making machine learning workflows for genomics and multi-omics more accessible and efficient. By relying on other open-source packages, it integrates smoothly into existing systems and analytical protocols, fostering broader adoption and community contributions. Our vision for GenoML extends beyond its current capabilities as an autoML tool, anticipating its evolution into a more expansive framework of machine learning tools tailored for genomics and beyond. GenoML can be accessed through the website (https://genoml.com/) or directly on GitHub (https://github.com/GenoML/genoml2). An overview of the current functionality included in GenoML can be found in the arXiv preprint (Makarious et al. 2021), as well as **Figure 27**.

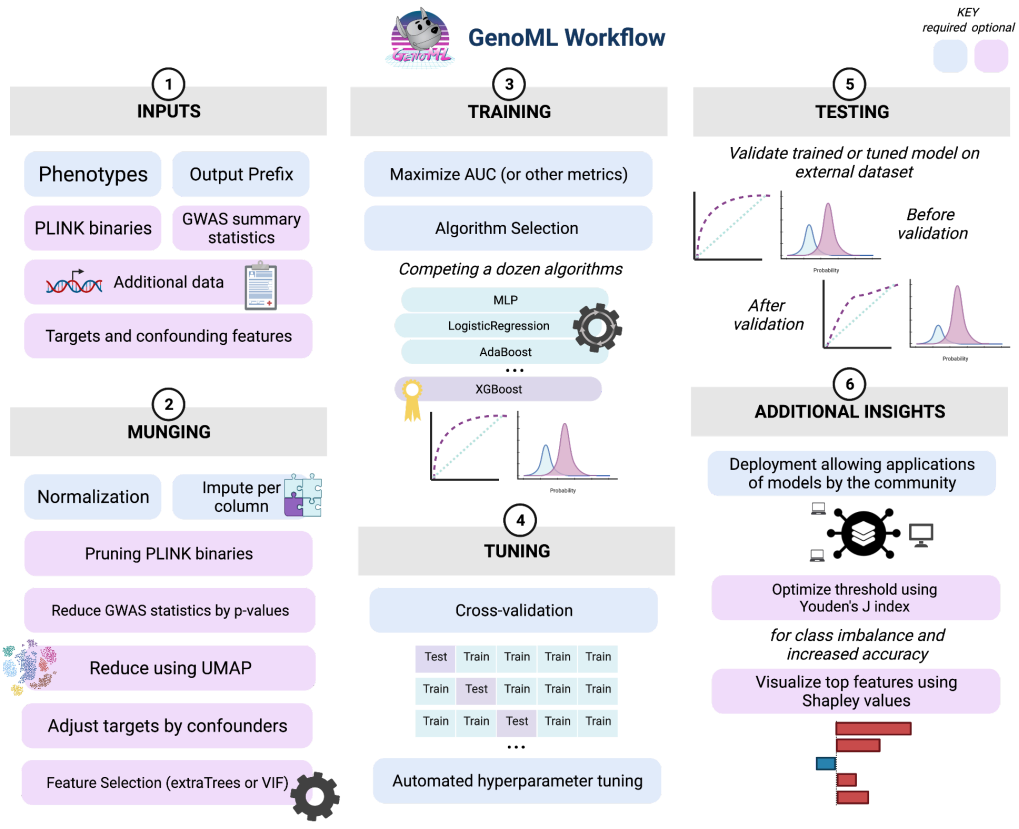


Figure 27: GenoML Workflow

All current functionality in the GenoML open-source Python package. Blue indicates required inputs and features, while purple indicates optional inputs and features. GWAS: Genome-wide association study; AUC: area under the curve; UMAP: Uniform manifold approximation and projection; VIF: variance inflation factor; MLP: multi-layer perceptron networks; XGBoost: an extreme gradient boosting.

Overview of Supervised Machine Learning

Supervised ML algorithms learn from labeled training data, allowing them to predict outcomes or categorize data. In the context of this multi-modal PD risk prediction study, the dataset comprises patient samples, each labeled as either "has PD" or "does not have PD". Supervised ML algorithms trained in this study aim to discern the patterns in the features that distinguish PD patients from neurologically-healthy controls.

General ML Approach

1. **Preprocessing or Munging:** This stage involves cleaning and organizing the data. For example, handling missing values or normalizing data ranges.
2. **Feature Selection:** Here, the algorithm selects the most important features for the model, reducing complexity and improving performance by eliminating irrelevant data.

3. **Training:** Here, the algorithm learns from the labeled dataset. It involves optimizing parameters to best fit the data, usually by minimizing a loss function.
4. **Testing and Validation:** This phase tests the trained model's performance on unseen data, helping to assess its generalization ability.
5. **Cross-validation and Tuning:** Cross-validation aids in model generalization by using different data subsets for training and testing, while tuning adjusts hyperparameters to optimize model performance.
6. **Presentation and Interpretation of Results:** The outcomes of the ML model are interpreted and presented in a meaningful way, often involving statistical analysis and visualizations.
7. **Replication and Follow-up Investigation of Results:** This step involves verifying the results through repeated experiments or extended studies.

Machine Learning Metrics

AUC

The Area Under the Curve (AUC) metric, particularly the Receiver Operating Characteristic (ROC) AUC, is an essential metric for assessing the performance of classification models. It gauges how effectively a model can differentiate between various classes. A higher AUC value suggests that the model is more adept at accurately identifying true positives while reducing false positives. The AUC value varies between 0 and 1, where a score of 0.5 signifies a model's inability to discriminate between classes.

Accuracy

Accuracy reflects the proportion of correctly predicted instances compared to the total number of observations. It offers an immediate and simple assessment of a model's overall effectiveness, especially in datasets with balanced class distributions. However, in cases of imbalanced datasets, where one class predominates over another, relying on accuracy as an indicator can be unreliable.

Balanced Accuracy

Balanced accuracy compensates for dataset imbalances by averaging the sensitivity (true positive rate) and specificity (true negative rate). This metric offers a more detailed perspective on a model's effectiveness across different classes, particularly useful in situations with uneven class distribution.

Log Loss

Log loss, also known as logistic loss, evaluates a classification model's accuracy by assigning a penalty to incorrect predictions. A lower log loss value signifies superior model performance, with an ideal model achieving a log loss of 0. This measure is especially important when evaluating models that yield probability predictions instead of discrete classes.

Sensitivity

Sensitivity, often referred to as the true positive rate or recall, assesses the percentage of actual positive cases that are correctly identified. This metric is particularly important in situations where failing to

detect a positive case (like a disease) has more severe consequences than mistakenly identifying a false positive.

Specificity

Specificity, also known as the true negative rate, measures the percentage of actual negative cases that are accurately identified by a model. High specificity implies that the model is effective at correctly rejecting non-relevant cases, thereby minimizing the risk of unnecessary interventions. In contexts where the costs or implications of a false positive are high, prioritizing specificity ensures more reliable and safe decision-making based on the model's predictions.

PPV

The Positive Predictive Value (PPV), or precision, calculates the percentage of correct positive predictions out of all positive identifications made by a model. This metric is relevant in contexts where the repercussions of false positives are significant. PPV is essential for ensuring the reliability of a model in accurately identifying true positive cases.

NPV

The Negative Predictive Value (NPV) measures the accuracy of a model in correctly identifying negative cases among all the identified negative instances. This metric is relevant in scenarios where correctly determining the non-existence of a condition or characteristic is critical. NPV is essential for ensuring the model's effectiveness in accurately confirming true negative outcomes.

Supervised Machine Learning Algorithms Implemented in GenoML

The majority of the development of GenoML has centered around classification. Classification aims to categorize objects into predefined labels. In this research, it involves categorizing individuals as either having PD or not.

Logistic Regression

Logistic regression is a statistical model for estimating the probability of a binary outcome. It is expressed by the logistic function:

$$p(y = 1) = \frac{1}{1+e^{-(\beta_0+\beta_1+x_1+..+\beta_n+x_n)}}$$

Where $p(y = 1)$ is the probability of the instance being in class 1, $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients, and $x_1, \ldots, x_n$ are the feature values. The benefit of using logistic regression is the simplicity and interpretability. However, it assumes that there is a linear relationship between the independent variables and the log odds of the dependent variable, which limits its use when trying to identify complex relationships between the features.

Decision Trees

Decision Trees are a non-parametric supervised learning method used for classification and regression. The model predicts the value of a target variable by learning simple decision rules inferred from the data features. Decision Trees are easy to understand and interpret, and can handle both numerical and categorical data. However, they are prone to overfitting, especially if the tree is allowed to grow complex without pruning.

Random Forest Classification

Random forest classification is an ensemble of decision trees, improving classification accuracy and control over-fitting. It is an ensemble learning method for classification that constructs a multitude of decision trees at training time. It outputs the class that is the mode of the classes of the individual trees. Random forests correct for decision trees' habit of overfitting to their training set. However, they can be slow to create predictions once trained. Their complexity can also make them harder to interpret than a single decision tree.

AdaBoost Classification

AdaBoost (Adaptive Boosting) is an ensemble technique that combines multiple weak classifiers to form a strong classifier. It sequentially focuses on misclassified instances by increasing their weights, guiding subsequent weak learners to concentrate on harder cases. The final model is a weighted sum of these weak classifiers. Ada Boost is sensitive to noisy data and outliers; however, it is less prone to overfitting compared to some other algorithms.

Gradient Boosting Classification

Gradient Boosting builds models sequentially, with each new model correcting errors made by the previous ones. It optimizes a loss function, and each new tree helps to correct errors made by previously trained trees. While gradient boosting often produces highly predictive models, a major drawback is its tendency to overfit and its sensitivity to noisy data. Moreover, training generally takes longer because trees are built sequentially.

Stochastic Gradient Descent (SGD Classification)

Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to fitting linear classifiers and regressors under convex loss functions such as linear Support Vector Machines and Logistic Regression. Even though SGD has been successfully applied to large-scale and sparse machine learning problems often encountered in text classification and natural language processing, the main disadvantage of SGD is that it requires a number of hyperparameters such as the regularization parameter and the number of iterations.

Support Vector Machine Classification (SVC)

Support Vector Machine (SVC) is a powerful and versatile classification method. It works by finding the hyperplane that best divides a dataset into classes. It is effective in high-dimensional spaces and for cases where the number of dimensions is greater than the number of samples. However, SVC can be

inefficient on large datasets and does not perform well when the data set has more noise i.e., target classes are overlapping.

MLP Classification

Multi-layer Perceptron (MLP) is a class of feedforward artificial neural network (ANN). An MLP consists of at least three layers of nodes: an input layer, a hidden layer, and an output layer. MLP uses backpropagation for training the network. MLPs are well-suited for complex pattern recognition tasks, but they require a large amount of data, and their complex architecture can make them prone to overfitting. Also, determining the right number of layers and the number of nodes in each layer can be challenging.

KNeighbors Classification

K-Nearest Neighbors (KNN) classification uses feature similarity to predict the values of new datapoints, which means that new data points are assigned a value based on how closely they resemble the points in the training set. While KNN is simple and easy to implement, its performance can significantly degrade with high-dimensional data due to the curse of dimensionality. It is also computationally expensive for large datasets.

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a method used in statistics, pattern recognition, and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events. It is simple, easy to implement, and interpretable. However, LDA assumes that the variables are normally distributed and have the same variance-covariance matrix across groups, which might not always be the case.

Quadratic Discriminant Analysis

Quadratic Discriminant Analysis (QDA) is similar to LDA but allows for the covariance matrices to differ between classes. This means it can model a wider range of datasets compared to LDA. The downside is that QDA can be more prone to overfitting, especially when the number of training samples is small compared to the number of features.

Bagging Classification

Bagging (Bootstrap Aggregating) is an ensemble machine learning algorithm that fits multiple models (typically of the same type) on different sub-samples of the dataset and then combines their predictions. While it reduces variance and helps to avoid overfitting, bagging can be computationally expensive and less interpretable compared to a single model.

XGBoost Classification

eXtreme Gradient Boosting (XGBoost) is an efficient and scalable implementation of gradient boosting. It has proven to be a highly effective ML algorithm, widely used in data science competitions and

real-world applications. XGBoost is known for its performance and speed, but like other gradient boosting methods, it can be prone to overfitting and requires careful tuning of its parameters.

Additional Features Implemented in GenoML

Additional Pre-processing Options

One of the key features I employed was leveraging PLINK and it's pruning algorithms, which is a method to reduce the complexity of genomic data by removing correlated variants, thereby minimizing redundancy and potential bias in the dataset.

To optimize the performance of models generated by GenoML, I implemented the option for users to choose their performance metrics. The tool allowed me to focus on maximizing either the AUC or balanced accuracy, depending on my specific research needs. Balanced accuracy, importantly, is a metric that considers both sensitivity and specificity, providing a more nuanced measure of model performance especially in imbalanced datasets. Additionally, I conducted basic ML preparations including imputation for missing values. GenoML facilitates this by offering options to replace missing data using either the median or mean of the respective feature, which helped in maintaining the integrity of the dataset. Users can also focus on specific metrics like maximizing sensitivity or specificity. Implementing this feature was particularly useful in scenarios where detecting true positives (sensitivity) or true negatives (specificity) is more critical, allowing for tailored model optimization based on the unique requirements of user's research. Furthermore, I developed GenoTools to internally to drop columns with zero variance, which are features that do not change across the dataset and thus do not contribute to the predictive power of the model.

Another critical aspect of GenoML was developing a way during data pre-processing that allowed for the adjusting and normalizing the features using PCA. This step was crucial to account for genetic population substructure, which can be a significant confounder in genetic association studies. By incorporating PCA when users use genetic data ensures that the models are robust and less likely to be biased by population stratification.

Feature Selection with ExtraTrees Classification

Feature Selection with ExtraTrees Classifier is a type of ensemble machine learning algorithm that uses randomized decision trees. By selecting important features, it can improve the interpretability of the model, reduce overfitting, and enhance model performance. However, the randomness in tree construction can lead to variability in the importance assigned to the features, and it can be computationally intensive with large datasets and many trees.

Hyperparameter Tuning and Optimization

To further refine the performance of GenoML models, I incorporated hyperparameter tuning and optimization strategies. Hyperparameters, otherwise known as the external configurations of the model, significantly influence the learning process and the resultant model performance. I employed a grid search approach, systematically working through multiple combinations of hyperparameters and using

cross-validation to evaluate their effectiveness. This exhaustive search ensured that the model was not just fitted to the training data but also generalized well to new, unseen data. By leveraging scikit-learn's optimization of hyperparameters such as the learning rate, number of trees in ensemble methods, and regularization techniques, GenoML is able to significantly auto-enhance the model's accuracy and efficiency.

Interpretability with Shapley values

I used Shapley values, a concept from cooperative game theory, to enhance the interpretability of my GenoML models. My work involved analyzing complex datasets that comprised various genotypes, gene expression levels, and some clinico-demographic data and Shapley values helped in understanding the contribution of each individual feature to the model's predictions. Shapley values adapt their game theory role to measure each feature's contribution to model predictions. This method treats the prediction task as a "game", where features are "players" and the prediction output resembles the game's "payout". Shapley values assess every possible combination of features, determining the impact of each on the prediction. This process then allocates equitable "credit" or "blame" among features based on their influence on the final decision.

While useful for aiding interpretability of otherwise "black box" ML models, in high-dimensional spaces like genomics, this calculation is resource-intensive, especially in models with numerous features, due to the need for retraining with every feature combination. The result is a set of values for each feature, indicating its average impact on predictive accuracy. I used SHAP to approximate a feature's influence on the model, focusing on the bi-directional changes of that feature relative to all others in the model (Lundberg et al. 2018, 2020). To further validate the model's efficacy, I trained a surrogate XGBoost model on 70% of the data and subsequently tested on the remaining 30% of withheld data. Conceptually, Shapley values can be likened to the relative importance measures derived from standard regression in terms of their interpretative value.

Study Participants

The data referenced in this study were sourced from the AMP PD Knowledge Platform. Updated details about the study can be found at https://www.amp-pd.org. Each participant in the respective cohorts gave their written informed consent for participation. The study's design and the data sharing agreement were jointly developed by AMP PD and the NIH, adhering to the NIH's standard ethical approval guidelines. Furthermore, all individuals participating in the AMP PD initiative consented to the use of their data in research pertinent to their respective cohorts.

In my research, I used clinical, demographic, and genome-wide DNA and RNA sequencing data from baseline visits of participants in the Parkinson's Progression Marker Initiative (PPMI) and the Parkinson's Disease Biomarkers Program (PDBP). These participants included individuals with PD and controls without neurological diseases. I focused on features present in at least 80% of the training and validation cohorts on the AMP PD platform, aligning with the inclusion criteria from previous modeling efforts and prioritizing *omic data that could be passively collected.

For this analysis, I chose RNA sequencing data from baseline visits, as these represent the time point closest to diagnosis, given that RNA signatures vary with disease stages. PPMI, with its recruitment of unmedicated individuals within one year of diagnosis, served as the training cohort. My retrospective model specifically analyzed refined PD diagnoses, excluding samples with conflicting diagnostic data over a decade of follow-up. This exclusion criterion also applied to cases with additional neurological diagnoses or retracted PD diagnoses, as well as controls who developed PD or other neurodegenerative diseases post-enrollment.

Furthermore, I omitted a subset of PD cases and controls from PPMI, specifically those from targeted recruitment efforts focusing on carriers of known genetic risk mutations (*LRRK2* and *GBA1*). The analysis was limited to unrelated individuals of European ancestry. For the AMP PD sample PRS, I used weights based on the latest European ancestry GWAS data. I identified participants for inclusion in my study based on their clinical, demographic, and genomic data, including DNA and RNA sequencing. Those with excessive missing data (over 15% per feature) were excluded. Each study involved adhered to the ethical guidelines set by their respective institutional review boards, and all participants had given informed consent for their inclusion in the original cohorts as well as in subsequent studies.

For this study, I collected clinical and demographic data, including the age at diagnosis for PD cases and the age at baseline visit for controls. I also included family history as a feature, specifically if a first or second-degree relative was diagnosed with PD, and inferred if individuals were of Ashkenazi Jewish ancestry. The latter was inferred using PCA, comparing our samples to a genetic reference series from GSE23636 at the Gene Expression Omnibus. Additionally, I ascertained sex clinically and confirmed it using X chromosome heterozygosity rates. The University of Pennsylvania Smell Inventory Test (UPSIT) was incorporated into the modeling. In total, the training cohort from PPMI consisted of 427 cases and 171 controls, and the validation cohort from PDBP consisted of 804 cases and 442 controls. For a detailed summary of the basic clinical and demographic features, see **Table 15**.

Study	Status	Age at Baseline (mean, SD)	UPSIT Score (mean, SD)	Male (%)	Positive Family History of PD (%)	Inferred Ashkenazi Ancestry (%)
PPMI	Case	61.75 (9.69)	23.48 (8.35)	65.57	25.53	6.09
	Control	60.61 (10.43)	34.18 (4.71)	63.74	5.85	11.11
PDBP	Case	64.59 (8.99)	19.65 (8.01)	64.18	24.88	3.61
	Control	62.87 (10.96)	32.52 (5.98)	45.25	8.14	4.07

Table 15: Descriptive statistics of studies included for multi-modal predictions from AMP-PD Release 1

All data included in this analysis following quality control using the first release of data made available by AMP-PD. AMP-PD: accelerating medicines partnership in Parkinson's disease; PPMI: Parkinson's progression marker initiative; PDBP: Parkinson's disease biomarker program; PD: Parkinson's disease; SD: standard deviation; UPSIT: University of Pennsylvania smell identification test.

Training Cohort - PPMI

PPMI, launched in 2010, represents a significant effort in the study of PD, sponsored by The Michael J. Fox Foundation and backed by various industry partners. This longitudinal study, involving over 1,500 participants at 33 clinical sites worldwide, aims to discover clinical, imaging, and biological markers crucial for the progression of PD. Participants, followed for durations ranging from five to thirteen years, contribute extensive clinical data, imaging results, and biological samples. The study's primary goal is to create a comprehensive dataset and biological samples to analyze the mean rates and variability of clinical, imaging, and *omic outcomes in early PD patients, those at prodromal stages, and subjects with specific genetic mutations in *LRRK2*, *GBA1*, or *SNCA*.

For PD subjects, inclusion criteria encompass having at least two primary PD symptoms (such as resting tremor, bradykinesia, or rigidity), a PD diagnosis for no more than two years, and being in the early stages (Hoehn and Yahr stage I or II) at baseline. These subjects must not be expected to need PD medication within six months from the baseline. Exclusion criteria are extensive, including current or recent use of specific PD medications, treatments that interfere with dopamine transporter imaging, and conditions that contraindicate safe lumbar puncture procedures. In addition to PD subjects, the study includes healthy controls, SWEDD subjects, prodromal subjects, and specific genetic cohorts, each with their own set of detailed inclusion and exclusion criteria.

PD subjects in the genetic cohort must meet similar criteria to the general PD cohort but also require confirmation of mutations in *LRRK2*, *GBA1*, or *SNCA*. They are eligible if diagnosed with PD for less than seven years and are in the earlier stages of the disease (Hoehn and Yahr stage less than 4). Unaffected individuals in the genetic cohort include those aged 45 or older with *LRRK2* or *GBA1* mutations or those aged 30 or older with *SNCA* mutations. These participants may also be first-degree relatives of mutation carriers. Both affected and unaffected individuals in these cohorts must be willing to undergo genetic testing and, if necessary, abstain from certain medications before imaging procedures. Exclusion criteria for these groups mirror those of the PD subjects, focusing on precluding factors for lumbar puncture and the use of certain medications. For detailed PPMI recruitment and biorepository information, please see the official study website https://www.ppmi-info.org/.

Validation Cohort - PDBP

The National Institute of Neurological Disorders and Stroke (NINDS) launched the PDBP to expedite the identification of novel diagnostic and progression biomarkers for PD. PDBP supports a range of research activities, from basic and translational studies to clinical research including hypothesis testing, target and pathway discovery, biomarker development, and disease modeling. PDBP shares data and biospecimens generated under the program, encompassing demographics, family history, medication history, and various clinical assessments and genetic information from 1,604 participants who met the AMP PD minimum clinical data criteria.

PDBP establishes clear inclusion and exclusion criteria for all participants. Individuals aged 21 or older, capable of consenting (or having an appropriate surrogate), and able to participate in study activities

without hardship or adverse health effects are included. Exclusions are primarily based on medical conditions or treatments that might interfere with the safe performance of routine procedures like lumbar punctures, history of certain mental health conditions, or use of investigational drugs shortly before the baseline visit. For PD cases, the inclusion criteria specifically require a clinical diagnosis of PD, while exclusions apply if the diagnosis is uncertain at enrollment. Control participants, on the other hand, are excluded if they have a significant neurological disorder or a family history of neurodegenerative diseases. For detailed PDBP recruitment and biorepository information, please see the official study website https://pdbp.ninds.nih.gov/. Age distribution comparisons between PPMI and PDBP, the training and validation datasets respectively, can be found in **Figure 28**.

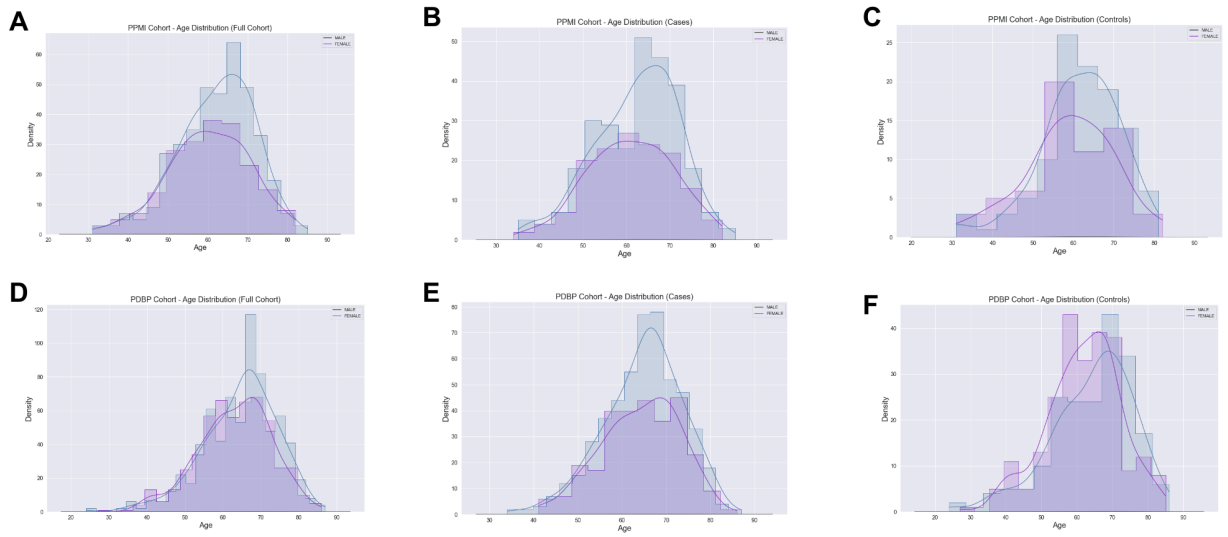


Figure 28: PPMI and PDBP Age Distribution

PDBP: Parkinson's Disease Biomarker Program; PPMI: Parkinson's Progression Markers Initiative

Data Pre-processing

DNA and RNA Sequencing

In this study, I used the generated DNA sequencing data using the standard short-read technology from Illumina. AMP-PD aligned this data using the Broad Institute's functional equivalence pipeline, and conducted joint genotyping of the sequencing data using the standard GATK pipeline (Regier et al., n.d.). These processes have been documented, from sample preparation to variant calling, in a comprehensive manuscript that details the AMP-PD whole-genome DNA sequencing effort (Iwaki et al. 2021).

For quality control of the samples, we set stringent inclusion criteria based on the genetic data output from the pipeline. These criteria included matching the genetically determined gender with the clinically ascertained one, achieving a call rate greater than 95% at both the sample and variant levels, maintaining a heterozygosity rate below 15%, and ensuring a freemix estimated contamination rate less than 3%. Additionally, we required a transition:transversion ratio above 2, no genetic relation closer than

a first cousin (IBD less than 12.5%), and confirmation of European ancestry through comparison with the HapMap project. I included WGS data in our study only if the variants passed the basic quality control checks of the initial sequencing (indicated by the PASS flag from the joint genotyping pipeline) and met specific criteria including non-palindromic alleles, missingness by case-control status with a P-value over 1E-4, a Hardy-Weinberg p-value above 1E-4, and a minor allele frequency in cases over 5%, as reported in the largest European PD meta-GWAS (Nalls et al. 2019b). Additionally, for our genetic modeling, we incorporated PRS from this meta-GWAS, ensuring that our testing and training samples were excluded from the weights (Nalls et al. 2019b).

In parallel, I also used the generated RNA sequencing data made available on AMP-PD from whole blood samples, a process carried out by the Translational Genomics Research Institute using standard protocols for Illumina NovaSeq technology (Hutchins et al. 2021). Our focus was specifically on blood samples collected at baseline. I adjusted variance stabilized counts for experimental covariates using the established limma pipelines (Ritchie et al. 2015). Our analysis centered on gene expression counts for protein-coding genes, and calculated differential expression p-values between PD cases and controls using logistic regression, adjusting for sex, plate, age, ten PCs, and the percentage of usable bases. Although the effects of non-coding RNA on PD have been explored in other studies (Lv et al. 2020), we chose to focus on protein-coding genes due to their established data quality, relevance in drug development, and the robust annotations available.

Data Preparation and Principle Component Analysis

The methodological overview and statistical analysis approach of this project are shown in **Figure 29**, which outlines the workflow and data. The initial phase involved data pre-processing at the baseline visit of individual-level data. Concentrating on baseline data was crucial to align the PDBP closely with individuals enrolled in PPMI, given that PPMI encompasses newly diagnosed and drug-naïve patients, whereas PDBP includes participants at various stages of PD.

The initial phase of the study concentrated on data preparation, involving both DNA and RNA sequencing datasets. For DNA sequencing, a specific subset of 10,000 variants was selected after LD pruning. This process involved retaining variants with an $r2 < 0.1$ within a 1MB range, thereby excluding areas known for extensive LD (Abraham, Qiu, and Inouye, 2017). Notably, these variants were not chosen based on their p-values in recent GWA studies, but to avoid regions with known high LD, such as the HLA region. This refinement process removed SNPs in high LD tracts while preserving those with known genetic risks within these regions. In RNA sequencing, PCA was employed using read counts for all protein-coding genes from each sample, creating a distinct set of ten PCs.

Subsequent to the preparation phase, features representing genetic variations from sequencing data were adjusted using the PCs derived from DNA sequencing. This adjustment aimed to mitigate the impact of identifiable European population substructures on the genetic features. This method aligns with the approach of adjusting for PCs in common variant regression models used in GWAS. Similarly, RNA sequencing data underwent parallel adjustments using RNA-derived PCs. The primary aim of these

adjustments was to statistically account for latent population substructures and experimental variables at the feature level, thus enhancing the applicability of the study across diverse datasets.

Post-adjustment, all continuous features underwent Z-transformation to ensure standardization with a mean of zero and a standard deviation of one, facilitating consistency across numerical scales. This standardization was crucial for preparing the data for subsequent phases of analysis.

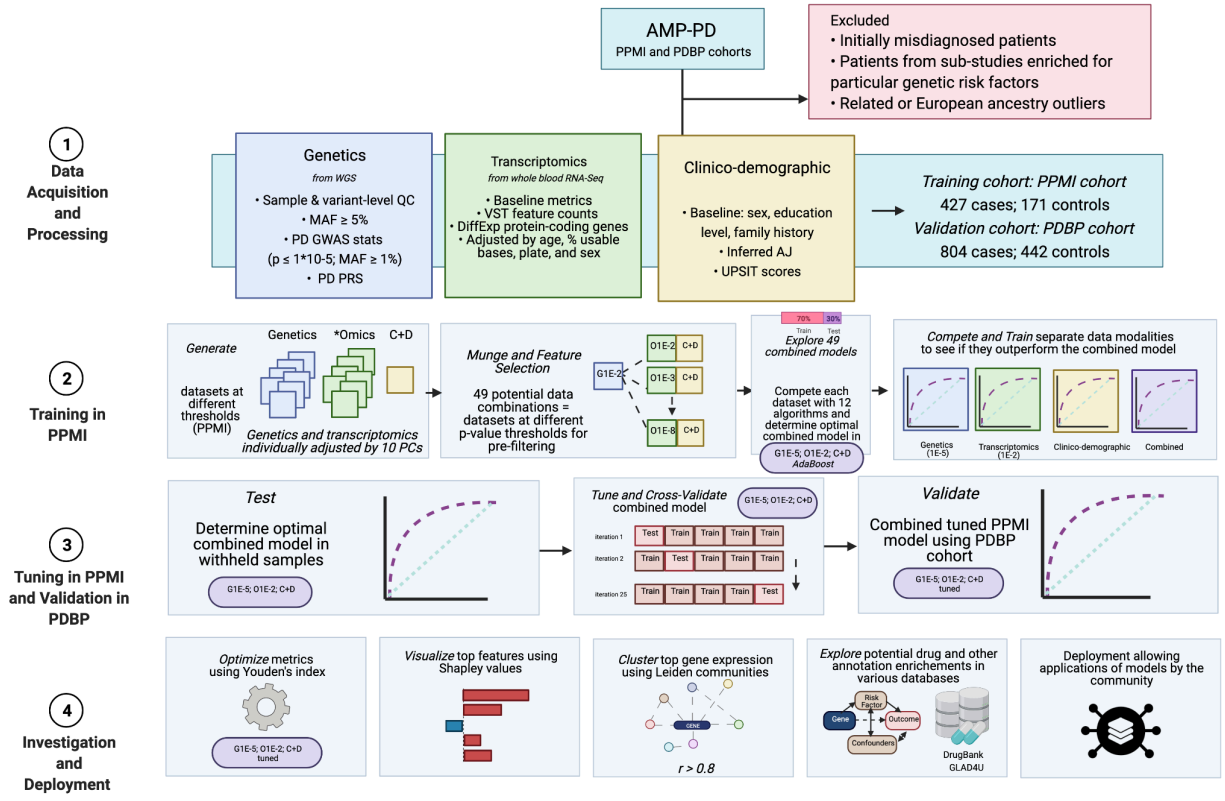


Figure 29: Workflow and Data Summary for Multi-omics of PD Risk Prediction using ML Study

Scientific notation in the workflow diagram denotes minimum p-values from reference GWAS or differential expression studies as a pre-screen for feature inclusion. Blue indicates subsets of genetics data (also denoted as "G"), green indicates subsets of transcriptomics data (also denoted as *omics or "O"), yellow indicates clinico-demographic data (also denoted as C+D), and purple indicates combined data modalities.

PD: Parkinson's Disease; AMP-PD: Accelerating Medicines Partnership in Parkinson's Disease; PPMI: Parkinson's Progression Marker Initiative; PDBP: Parkinson's Disease Biomarker Program; WGS: whole genome sequencing; GWAS: Genome-wide association study; QC: quality control; MAF: minor allele frequency; PRS: polygenic risk score

Algorithm Training, Validation, and Application

The data from PPMI was then partitioned into a 70% training set and a 30% testing set, randomly. This separation allowed for the distinct phases of algorithm training on the training set and subsequent

validation on the testing set. We evaluated 12 machine learning algorithms known for their robust performance, aiming to ascertain which could maximize the AUC for the classification of cases and controls. The selection of these algorithms was influenced by their demonstrated success in various domains, compatibility with Python's scikit-learn package, and their capability to output probability-based predictions ("Extremely Randomized Trees" 2010; Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, et al. 2011). This facilitated more straightforward training, testing, and interpretation of the model. The algorithms included 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). These methods represented a comprehensive range of techniques, from conventional linear models typically used in genetic prediction analyses to more complex tree-based, kernel-based, and deep learning approaches.

Feature selection was re-emphasized using the extremely randomized trees classifier algorithm (extraTrees) to minimize redundant features that could contribute to model overfitting. Following the feature selection process, the algorithm that demonstrated the highest AUC and balanced accuracy in the testing set (30% of PPMI data) was chosen for further tuning and cross-validation. This leading algorithm, AdaBoostClassifier in this case, was subjected to intensive hyperparameter tuning across the entire PPMI dataset. The tuning involved 25 random iterations of 5-fold cross-validation for each iteration, adjusting the number of potential predictors (estimators) from 1 to 1000, to ensure the robustness and reliability of the model.

This procedure was replicated 49 times, with varying thresholds of p-value-based feature inclusion. The thresholds were iteratively applied across all combinations of p-value thresholds [1E-2, 1E-3, 1E-4, 1E-5, 1E-6, 1E-7, 1E-8], derived from genetic data from the latest European meta-GWAS and transcriptomic data from differential expression studies (Nalls et al. 2019b). Due to structural and analytical differences between genetic and transcriptomic data, each modality underwent separate feature selection phases. At each of the 49 threshold combinations, the model maximizing the AUC metric was selected. The final model was chosen based on a blend of metrics including balanced accuracy, sensitivity, and specificity, to appropriately address case-control imbalance.

Prior to training, p-values were filtered using the largest meta-analysis of 17 datasets from PD GWAS in European ancestry samples (Nalls et al. 2019b). The training set from PPMI represents a small fraction of the most recent GWAS study, and the feature selection approach described above was considerably conservative, thus minimizing potential data leakage. This study primarily focuses on a model with a maximum p-value of 1E-5 for genetic data inclusion and 1E-2 for transcriptomic data inclusion, reflecting a careful balance between inclusivity and specificity in feature selection.

Feature and Model Selection and Refinement

Once the features were adjusted and normalized, internal feature selection was conducted using decision trees (extraTreesClassifier) within the PPMI training dataset. This process aimed to identify features to ensure optimal information extraction from the features (Lopez et al. 2020), while mitigating the risk of overfitting. Overfitting, defined as a model's tendency to overperform in training yet underperform in validation datasets, is often due to the inclusion of correlated or insignificant features ("Extremely Randomized Trees" 2010; Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, et al. 2011). The implementation of decision trees in feature selection played a crucial role in eliminating redundant and low-impact features, refining the feature set for more efficient modeling. This selection process was applied to combined data modalities to prevent potential inflation of model accuracy due to redundant feature contributions. The extraTrees classifier further provided estimates of features in order of their likely contribution to the final model, thereby streamlining the model construction process.

Subsequently, I used GenoML for algorithm competition and feature selection within the PPMI dataset, adhering to a 70:30 training-to-testing data split. Among the top 5% of features identified in the Shapley analysis, the mean feature correlation was less than 5%, with a maximum of 36% (**Figure 30**). By employing correlation-based pruning and an extraTrees classifier in the data munging stage, this limited the potential for overfitting but also ensured a more conservative model approach.

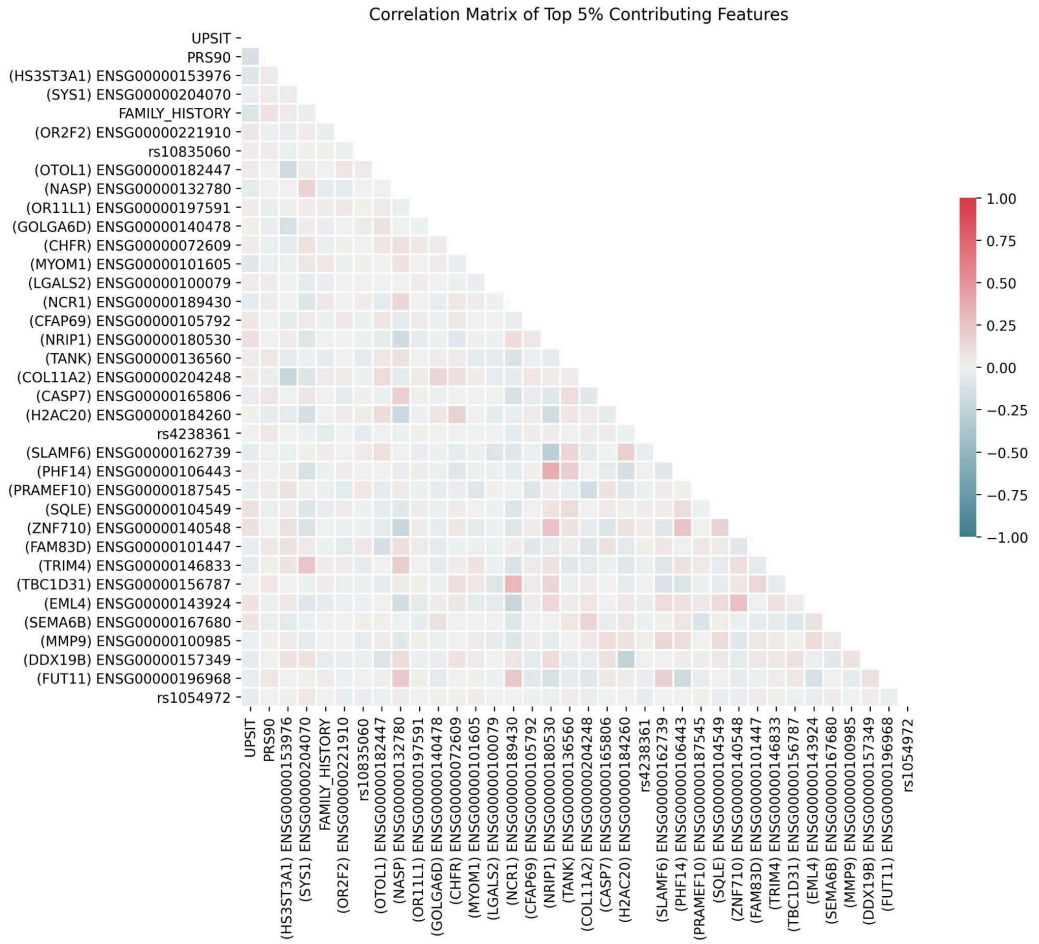


Figure 30: Correlation matrix of top 5% contributing features in ML model

Model Improvements and Feature Interpretation

Class Imbalance Post-hoc Optimization

I performed a post-hoc optimization process addressing class imbalance. This refitting uses an optimized threshold for case probability, derived from Youden's J statistic, to enhance the model's ability to manage case-control imbalance, consequently improving balanced accuracy and associated metrics (Ruopp et al. 2008).

Specifically, for PDBP, case probabilities are generated using the same model, features, and parameters as those used in the training and tuning phases of PPMI. However, a critical distinction is made in the PDBP cohort; the probability threshold for identifying a case is uniquely adjusted to address the specific imbalance found in PDBP. This adjustment represents a targeted post-hoc probabilistic optimization that deliberately avoids any reweighting or prioritization of features in both datasets, thus preserving the integrity of a standalone validation phase.

Network Communities

I explored constructing gene networks potentially relevant to PD, using features identified by the top ML classifiers. The primary objective was to determine whether any drug-related annotations were disproportionately represented in my network communities, based on correlated gene expression in cases, compared to other protein-coding genes identified as potential classifiers of case-control status.

The initial step involved extracting RNA feature counts for 597 genes, recognized as significant RNA sequencing-derived features during the training phase. I then refined this dataset to include only case samples and computed the correlation between gene-level transcriptomic data for these nominated genes. This process led to the construction of a network graph, setting a minimum correlation coefficient (r) threshold of 0.8 for connections between gene nodes. Following this, I employed the Leiden algorithm with default settings to cluster the genes into related communities within the larger network. The quality of these network clusters was assessed by calculating a modularity score (Traag, Aldecoa, and Delvenne 2015). To investigate potential therapeutic connections within these network communities, I used webGestalt's R package, specifically its over-representation analysis function (https://www.webgestalt.org/). This tool helped in exploring the enrichment of druggable targets for network genes within two drug databases: DrugBank and GLAD4U (Y. Liao et al. 2019; Wishart et al. 2008; Jourquin et al. 2012). The analysis was conducted under default settings and were accessed on February 23rd, 2021.

Initially, the 300 genes forming our network communities, derived from the background of all 598 genes identified in the initial feature selection phase, were queried. This comparison aimed to discern the enrichment of genes within my network communities, which show high correlation in cases, against genes implicated in case-control differences. Additionally, a similar analysis was conducted comparing all 597 genes differentiating cases and controls with over 18,000 protein-coding genes. This step was repeated for our 300 network community genes to investigate the over-representation of druggable targets against the entire set of protein-coding genes.

Results

Integrating Various Modalities Results in More Accurate Predictions

In this study, I have demonstrated that employing a combination of various data modalities significantly enhances the accuracy of predicting PD diagnosis, especially when compared to models relying on a single modality. Specifically, our multi-modality model, incorporating diverse data types, achieved a notably higher AUC of 89.72%. This is an improvement over models based solely on clinico-demographic data (87.52% AUC), genetics-only approaches using genome sequencing data and PRS (70.66% AUC), and transcriptomics-only methods employing genome-wide whole blood RNA sequencing data (79.73% AUC). These results were evident in withheld samples from PPMI (summarized in **Table 16** and **Figure 31**).

Data Modality	Stage	Algorithm	AUC (%)	Accuracy (%)	Balanced accuracy (%)	Log Loss	Sensitivity	Specificity	PPV	NPV	Number of Features
Genetics (P<1E-5)	Training in PPMI (70:30)	MLPClassifier	70.66	70	60.64	0.83	0.83	0.38	0.77	0.48	233
Clinico-demographic	Training in PPMI (70:30)	Logistic Regression	87.52	79.44	75.27	0.39	0.85	0.65	0.86	0.64	6
Transcriptomics (P<1E-2)	Training in PPMI (70:30)	SVC	79.73	73.89	54.6	0.48	0.97	0.12	0.75	0.6	678
Combined	Training in PPMI (70:30)	AdaBoost Classifier	89.72	85.56	82.41	0.63	0.89	0.76	0.91	0.73	713

Table 16: Performance metric summaries comparing training in withheld samples in PPMI

PPMI: Parkinson's progression marker initiative; PDBP: Parkinson's disease biomarker program; AUC: Area under the curve; Log loss: Logarithmic loss; PPV: Positive Predictive Value; NPV: Negative Predictive Value; MLPClassifier: Multi-Layer Perceptron Classifier; SVC: Support Vector Classifier; AdaBoost Classifier: AdaBoost (Adaptive Boosting) Classifier.

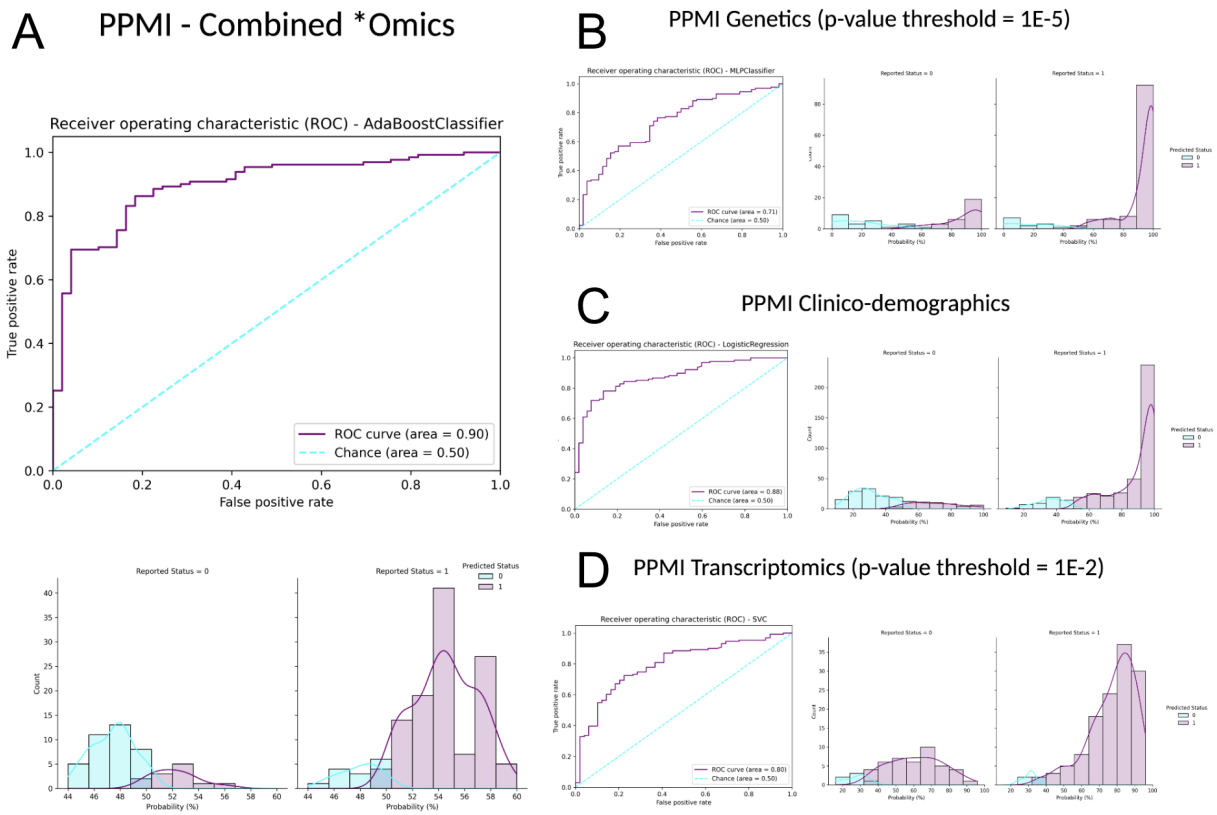


Figure 31: Receiver operating characteristic curves and case probability density plots in withheld training samples at default thresholds comparing performance metrics in different data modalities from the PPMI dataset

P-values mentioned indicate the threshold of significance used per data type, except for the inclusion of all clinico-demographic features. A) PPMI combined *omics dataset (genetics p-value threshold = 1E-5, transcriptomics p-value threshold = 1E-2, and clinico-demographic information); B) PPMI genetics only dataset (p-value threshold = 1E-5); C) PPMI clinico-demographics only dataset; D) PPMI transcriptomics only dataset (p-value threshold = 1E-2). Note that x-axis limits may vary as some models produce less extreme probability distributions than others inherently based on fit to the input data and the algorithm used. PPMI: Parkinson's Progression Marker Initiative; ROC: Receiver Operating Characteristic curve.

Further fine-tuning of the model yielded better outcomes. Initially, the untuned model in the PPMI dataset had a mean AUC of 80.75, with a standard deviation of 8.84, ranging from 69.44 to 88.51. Post-tuning, the mean AUC in PPMI improved to 82.17, with a standard deviation of 8.96 and a range of 70.93 to 90.17 (**Table 17**). This improvement was also mirrored in validations using the PDBP dataset, where the AUC of the combined modality model was 83.84% before tuning (summarized in **Table 18** and **Figure 32**).

Data Modality	Genetics (P<1E-5)	Clinico-demographic	Transcriptomics (P<1E-2)	Combined
Stage	Tuning in PPMI	Tuning in PPMI	Tuning in PPMI	Tuning in PPMI
Algorithm	MLP Classifier	Logistic Regression	SVC	AdaBoost Classifier
AUC at training (%)	70.66	87.52	79.73	89.72
Mean, AUC during CV for baseline model (%)	69.44	88.51	78.05	86.99
Standard deviation, AUC during CV for baseline model (%)	4.46	2.17	4.27	2.3
Min, AUC during CV for baseline model (%)	62.45	86.19	71.49	84.27
Max, AUC during CV for baseline model (%)	75.73	91.98	82.62	90.7
Mean, AUC during CV for tuned model (%)	70.93	88.55	79.01	90.17
Standard deviation, AUC during CV for tuned model (%)	5.39	2.2	4.71	1.64
Min, AUC during CV for tuned model (%)	61.29	86.33	70.88	88.06
Max, AUC during CV for tuned model (%)	76.71	92.15	84.01	92.73
Variance, AUC during CV for baseline model (%)	19.89	4.73	18.2	5.29
Variance, AUC during CV for tuned model (%)	29.03	4.82	22.18	2.7

Table 17: Performance metric summaries comparing at tuned cross-validation in withheld samples in PPMI

PPMI: Parkinson's progression marker initiative; Min: minimum; Max: maximum; AUC: Area under the curve; CV: cross-validation; PPV: Positive Predictive Value; NPV: Negative Predictive Value; MLPClassifier: Multi-Layer Perceptron Classifier; SVC: Support Vector Classifier; AdaBoost Classifier: AdaBoost (Adaptive Boosting) Classifier.

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	68.09	0.67	0.93	0.43	0.74	0.78

Table 18: Performance metric summaries comparing combined tuned and untuned model performance on PDBP validation dataset

PPMI: Parkinson's progression marker initiative; PDBP: Parkinson's disease biomarker program; AUC: Area under the curve; Log loss: Logarithmic loss; PPV: Positive Predictive Value; NPV: Negative Predictive Value; AdaBoost Classifier: AdaBoost (Adaptive Boosting) Classifier.

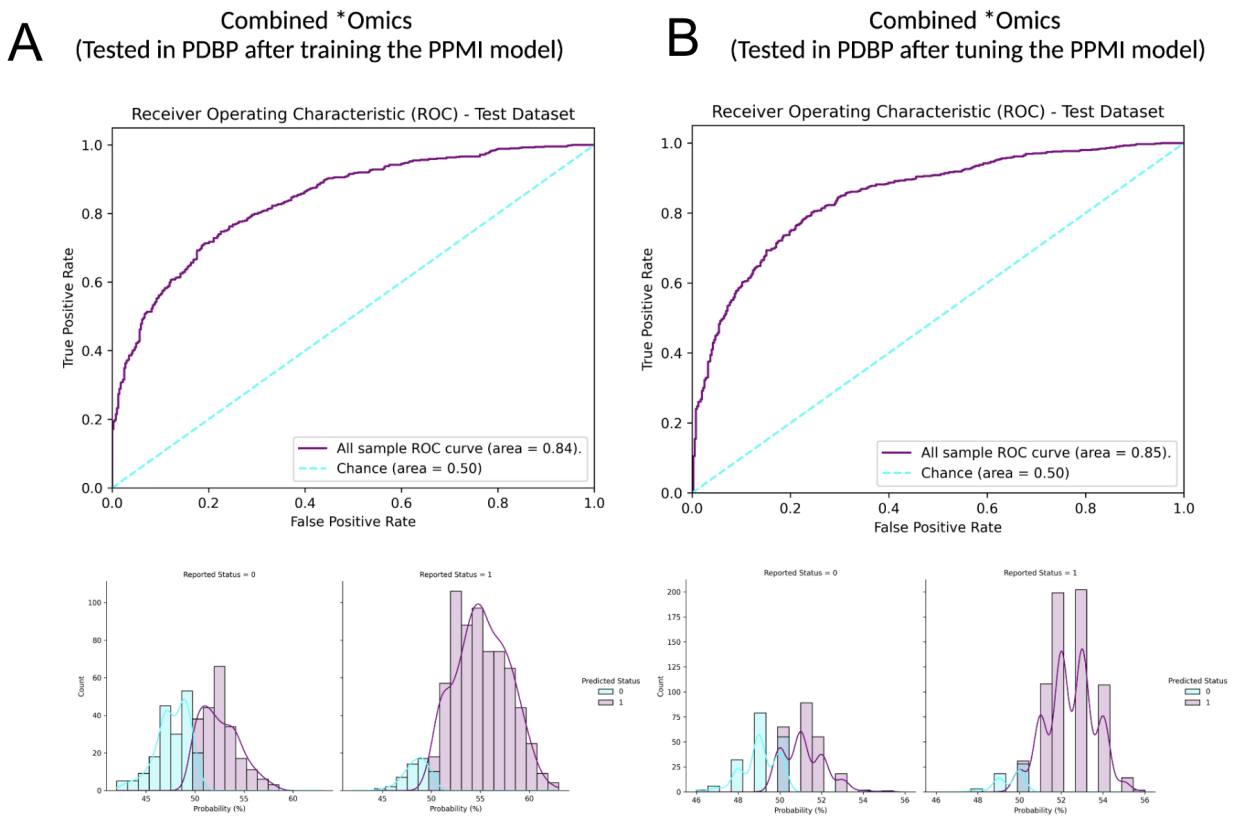


Figure 32: Receiver operating characteristic and case probability density plots in the external dataset (PDBP) at validation for the trained and then tuned models at default thresholds

Probabilities are predicted case status (r1), 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. A) Testing in PDBP the combined *omics model (genetics p-value threshold = 1E-5, transcriptomics p-value threshold = 1E-2, and clinico-demographic information) developed in PPMI prior to tuning the hyperparameters of the model; B) Testing in PDBP the combined *omics model (genetics p-value threshold = 1E-5, transcriptomics p-value threshold = 1E-2, and clinico-demographic information) developed in PPMI after tuning the hyperparameters of the model. PPMI: Parkinson's Progression Marker Initiative; PDBP: Parkinson's Disease Biomarker Program; ROC: Receiver Operating Characteristic curve

Importantly, the multi-modal model exhibited the lowest rates of false positives and negatives when compared to models focusing on a single modality. This was consistent across both the withheld test set in PPMI and the PDBP validation set. Thus, transitioning from a single data modality to a multi-modal approach not only enhanced the AUC but also improved other performance metrics, underscoring the effectiveness of integrating multiple data types in predictive modeling for PD diagnosis.

Comparison of the Refined Multi-Modal Model with Earlier Models

Previous work conducted by Nalls and colleagues had developed a strong classifier using the UPSIT-only model within the same PPMI training set (Nalls et al. 2015). Their model achieved a high AUC of 90.1%. However, according to DeLong's test, a statistical method used to compare the areas under two correlated ROC curves often used in medical research when comparing the diagnostic accuracy of two

tests or models, an integrative model they proposed proved to be more informative. While the UPSIT-based model showed high predictive power, it had inherent limitations. Specifically, a decline in smell identification is not exclusively indicative of PD; it can also signal general neurodegeneration, aging, or environmental impacts.

The advantage of employing a multi-modal approach lies in the varied predictive strengths of different modalities. In this context, leveraging a diversity of data types enhanced both the sensitivity and specificity of our predictions. The final multi-modal model constructed in this study, applied to withheld PPMI data, demonstrated superior performance over single-modality models. It achieved higher accuracy and balanced accuracy rates of 85.56% and 82.41% respectively, along with a sensitivity of 89.31% and a specificity of 75.51% (**Table 18**), complete performance metrics for best combined method comparing training in withheld samples in PPMI can be found in **Supplementary Table 20**.

To further quantify the effectiveness of our approach, I compared the distribution of AUCs across all cross-validation iterations using a T-test. This comparison revealed that our combined model consistently outperformed the clinico-demographic model in the PPMI dataset, with a significant statistic of |10.23| and a p-value of 9.95E-23. Improved balanced accuracy is particularly crucial in binary classifiers, especially when predicting classes that are rarer in the general population, like PD.

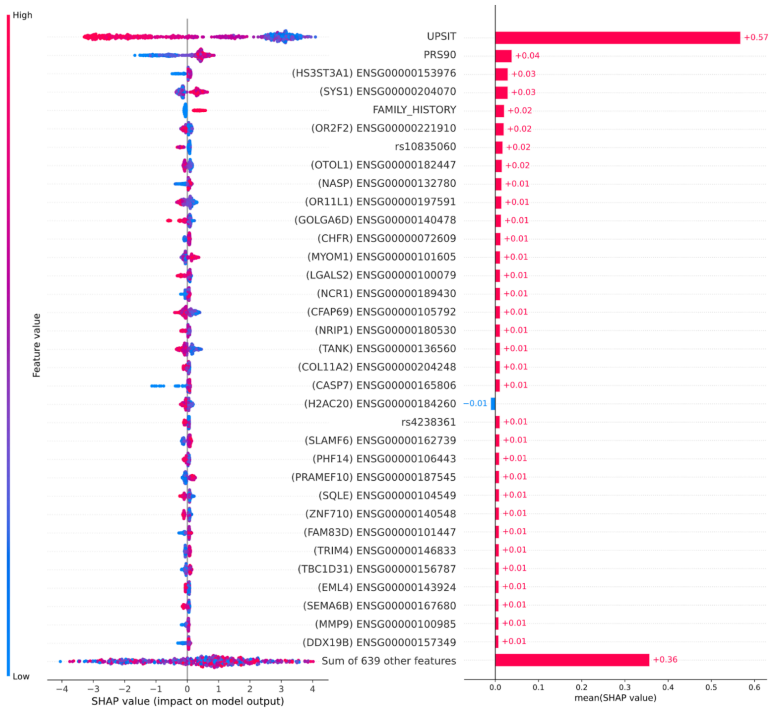


Figure 33: Feature importance plots for top 5% of features in XGBoost surrogate of best combined *omics model

Plot on the left have lower values indicated by the color blue, while higher values are indicated in red compared to the baseline risk estimate. Plot on the right indicates directionality, with features predicting for cases indicated in red, while features better predicting controls are indicated in blue.
SHAP: Shapley values; UPSIT: University of Pennsylvania Smell Identification Test; PRS: Polygenic risk score

Also in this study, special emphasis was placed on validating our model. This included interpreting and visualizing the top features that assist in classification prediction using Shapley values (**Figure 33**). We also focused on tuning the model prior to developing hypothesis-free transcriptomic communities and exploring potential drug-gene interactions.

Advantages of Using Machine Learning for Developing Multi-Modal Prediction Models

I leveraged GenoML to construct and fine-tune model parameters and handle non-linear associations, offering a distinct advantage over traditional regression-based methods in disease prediction. This approach is particularly evident in the performance of our best model, which used the strengths of the AdaBoost classifier. By integrating all available data, this model exhibited an AUC ranging from 88.06% to 92.70% during 5-fold cross-validation, with an average of 90.20% and a standard deviation of 2.3%, as observed in the PPMI dataset (**Table 17**).

Further validation of this model in the PDBP dataset revealed an AUC of 85.03%, along with a high sensitivity of 93.12% and specificity of 43.07%. Notably, these metrics showed additional improvement following post-hoc optimization of case probability thresholds.

To address the sample imbalance between case and control groups, I opted to calculate Youden's J statistic instead of traditional methods like upsampling or downsampling. By using Youden's J to find the optimal threshold, I adjusted the case-control threshold from the default 50% to a more tailored 51%. This threshold adjustment enhanced the specificity for the PDBP cohort during validation. Applying this optimized threshold to the withheld PPMI samples, the model's balanced accuracy in the training phase was elevated to 83.95%. Similarly, the balanced accuracy improved to 77.97% when applying the model to the PDBP validation data (see **Table 19** for an overview of other related metrics and a contrast between optimized and default thresholds). Overall, this optimization of thresholds strategy generally enhanced classifier performance with only a slight increase in computational expense.

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	0.8	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

Table 19: Optimizing the AUC threshold in withheld training samples and in the validation data

Optimization in this scenario is referring to using Youden's J statistic to account for case/control imbalance and more appropriately define the threshold. PPMI: Parkinson's progression marker initiative; PDBP: Parkinson's disease biomarker program; Log loss: Logarithmic loss; PPV: Positive Predictive Value; NPV: Negative Predictive Value.

Predictive Performance is Primarily driven by UPSIT and PRS

In developing the combined predictive model, I included a comprehensive range of features: 51 SNPs, 418 protein-coding transcripts, along with key demographics, age, family history, olfactory function, and previous genome-wide significant PRS (Nalls et al. 2019b). The importance of these features was analyzed using Shapley explanations and visualized in **Figure 33**, which was based on withheld training data.

A closer examination of the SHAP values, both in training and testing phases, revealed that the UPSIT scores and PRS were the most influential in driving the model's predictive power. However, the accuracy of the model was further enhanced by the inclusion of numerous smaller effect transcripts and risk SNPs. The SHAP plots also indicated a clear trend: lower UPSIT scores, shown in blue, were strongly associated with a higher likelihood of PD, corroborating findings from earlier studies that used the smell identification test for PD diagnosis (Doty et al. 1984; Morley et al. 2018; Picillo et al. 2014).

An interesting observation from the SHAP analysis was the varied directionality of genetic features. This variation implies that the overexpression of certain genes is associated with healthy controls, while for others, it indicates the opposite. Regarding sex, coded as "MALE" in the dataset, it was initially included but was not a determining factor in the final model due to a balanced distribution of sexes in the PPMI training dataset. This resulted in sex being excluded during feature selection, leading to a SHAP value of zero for this feature. Detailed SHAP values for all predictive features in the top-performing model are available in **Supplementary Table 21**. Moreover, the inclusion of SNPs beyond those in the PRS hints at potential compensatory or risk-modifying interactions with the PRS, as further detailed in **Figure 30**.

QTL analysis was conducted on each SNP and transcript combination among the top features in the final model, using linear regression on the adjusted dataset. This analysis revealed that no SNP and transcript combination passed multiple test corrections, highlighting the effectiveness of our feature selection process in eliminating redundant or correlated features (**Supplemtary Table 22**). Additionally, an analysis of the independence of these top features showed correlations ranging from a minimum of 1.40E-05 to a maximum of 0.364, with an average of 0.045 and a standard deviation of 0.049. These results, fully detailed in **Figure 30** confirm that the features in the top-performing model are largely independent of each other.

Predictive Performance is Primarily driven by UPSIT and PRS

After choosing the best multi-modal model, I focused on constructing some preliminary gene expression network communities, which were built using RNA sequencing data obtained from positive PD cases. The genes included in this analysis were specifically selected through a feature selection process. These communities represent unique PD-specific networks, identified within whole blood RNA sequencing data. Conceptually, these networks can be likened to biological pathways. They comprise genes whose expression is highly correlated within the case-only transcriptomics dataset, and the communities are essentially subgroups of these genes that are closely interrelated within the larger network. I identified 13 such network communities, encompassing 300 genes, characterized by an Erdős–Rényi modularity score of 0.794. This score being closer to 1, which is a measure used to assess the strength of division of a network into modules or communities, indicates a strong model fit. A graphical representation of these findings are detailed in **Figure 34**.

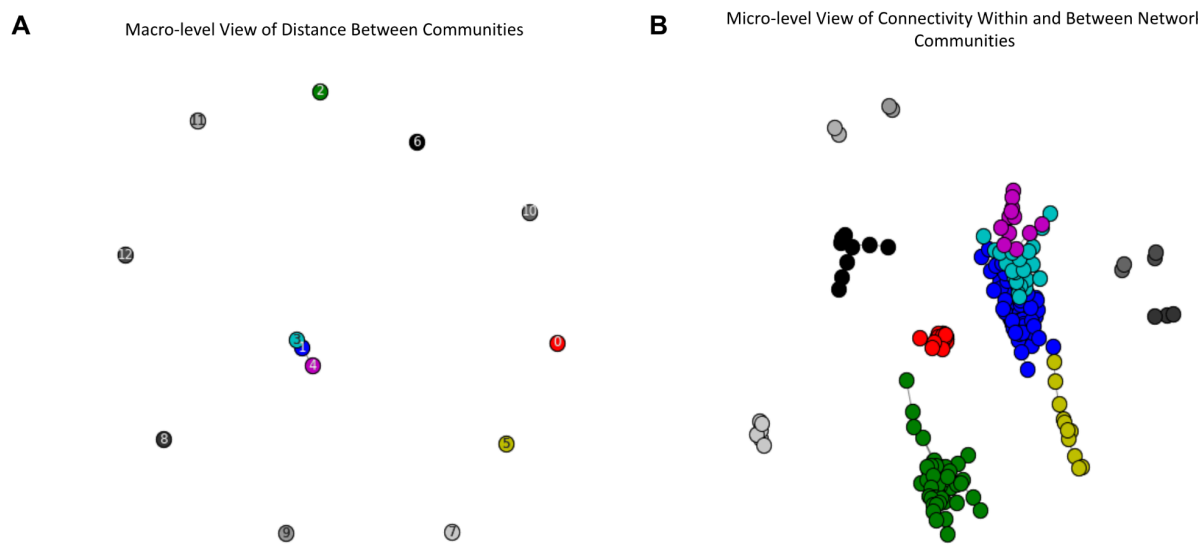


Figure 34: Network plot of nominated genes following best combined multi-model ML prediction model

Panel A provides a macro-level view of the distance between communities. 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.

In analyzing the genes from these network communities, I also explored their potential overrepresentation as known drug target genes. A comparative analysis between the genes in the network communities (300 genes) and those included as features in the case-control model build (598 genes) revealed significant enrichments. Specifically, genes related to fostamatinib showed a notable enrichment (FDR adjusted p-value of 2.21E-4 for *MYLK*, *EPHA8*, *HCK*, *DYRK1B*, and *BUB1B-PAK6*) as per DrugBank annotations. Fostamatinib is known for its role in modulating immune responses and inflammation, with the enrichment of these genes suggests potential therapeutic targets for reducing inflammation in PD (M. W. Kim et al. 2022).

Similarly, genes associated with copper were also overrepresented (FDR adjusted p-value of 0.0286 for *HSP90AA*, *CBX5*, and *HSPD1*). An additional query in the GLAD4U database indicated a significant overrepresentation of L-lysine annotated genes (FDR adjusted p-value of 0.0057 for *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*). L-lysine, an essential amino acid, enrichment in these networks might possibly be related to protein aggregation or neurotransmitter synthesis, suggesting new pathways or mechanisms in PD pathology (R. Wang et al. 2020).

However, other variations of our drug target enrichment analyses did not yield any significant overrepresentation of drugs after adjusting for multiple tests, although gamma hydroxybutyric acid consistently emerged as the top unadjusted result (p-values ranging from 0.0056 to 0.0001 for *SLC16A7*, *SLC16A3*, and *GABBR1*). Gamma hydroxybutyric acid, known for its impact on sleep and mood, could have neuroprotective qualities, and could be a therapeutic avenue to explore when managing non-motor symptoms like sleep disturbances in PD (Dornbierer et al. 2019).

Conclusions and Discussion

Integrating Diverse Data Modalities for Enhanced Parkinson's Disease Prediction

In the current landscape of research, where there is an increase in genomic and clinico-demographic data, the construction of multi-modal models leveraging these varied data modalities, will continue to enhance performance and scope.

Numerous studies have explored PD risk and onset using varied data types, such as gait analysis (Palmerini et al. 2017), fall detection techniques (Silva de Lima et al. 2017), other motor data (Noyce et al. 2014), and sleep behavior analysis (Campabadal et al. 2021). This study takes a unique approach, integrating adjusted transcriptomics, genetics, and clinical data into a single predictive model. This model, framed within ML, seeks to complement existing studies, such as those focusing on imaging for

predicting REM sleep behavior disorder (Mei et al. 2021; D. A. Lee et al. 2021) and cognitive deficits (Rahayel et al. 2018). My approach offers the advantage of lower resource requirements for model implementation.

I see potential in merging this model with existing studies through transfer or ensemble learning techniques, which could be highly beneficial. In fact, this work has already been replicated elsewhere (J. Zhang et al. 2023). However, I faced limitations in this study due to data scarcity and the availability of potential features in public datasets. Previous research identified UPSIT as a critical predictive feature in PD, as demonstrated by Prashanth and colleagues, who trained a model on CSF biomarkers (Prashanth et al. 2016). Building on this, my study aims to develop a model based on data that are remotely accessible or commonly found in biobanked samples, avoiding the need for costly and logistically challenging specialist clinical visits.

A comprehensive review in this space examined studies employing ML in PD diagnosis, highlighting a key distinction of this study: the focus on specific modalities and commitment to training, tuning, and validating using publicly available cohorts. This ensures transparency, reproducibility, and applicability in transfer learning (Mei, Desrosiers, and Frasnelli 2021).

This study underscores the effectiveness of integrating diverse data modalities in predictive modeling, with broad implications for healthcare data utility. This approach is pivotal in areas like clinical trial enrollment and stratification. We describe a methodology that not only enables accurate, early PD diagnosis but is also potentially cost-effective for biobanks and established healthcare systems. The model-building process has successfully developed robust models for peri-diagnostic PD, the period around the time of diagnosis of disease, simultaneously generating novel gene network communities correlated with PD that could inform therapeutic development. A key strength of this multi-modal approach is its compensatory mechanism, where various modalities balance each other out, some excelling in predicting case status, others in classifying controls.

Machine Learning to Develop Adjunct Screening Models

While this study introduces a novel approach to PD diagnosis, it should be emphasized that the model's role is a supplementary tool rather than a replacement for current diagnostic methods. The primary aim is to assist in identifying individuals at high risk, particularly useful in large-scale biobank or study recruitment settings. To fully understand the model's efficacy in early PD detection, further studies are necessary, especially to differentiate early PD cases from other diseases in high-risk groups.

Regarding the prevalence of PD in an aging population, approximately 2% are affected (Alves et al., 2008). Within this context, our optimized model, based on the PPMI data (**Table 19**), demonstrates a PPV of 8.75% and a NPV of 99.66%. However, the model exhibits a high false discovery rate (FDR) of 91.26% and a low false omission rate (FOR) of 0.34%. This low FOR implies that in a scenario of screening 1000 individuals deemed as healthy, 3 to 4 might actually have PD. The model, therefore, is more effective for identifying broader groups for monitoring within health registries or biobanks, rather than pinpointing

individual cases. Despite its lower specificity, the model's high sensitivity makes it valuable for large-scale biobank contributions and trial recruitments, targeting individuals at risk.

My study's strength lies in its high balanced accuracy in differentiating between PD cases and controls. The PPMI dataset, which focuses on patients close to diagnosis and pre-treatment, differs from the PDBP in aspects such as recruitment style and patient medication status. For instance, PPMI involves unmedicated patients diagnosed within a year (confirmed by DatScan), whereas PDBP includes patients diagnosed within five years, regardless of medication status and without DatScan confirmation. These differences are reflected in the mean age and UPSIT scores between the cohorts, affecting model performance (**Table 15 and Table 20**). Notably, the clinico-demographic model, with UPSIT as a top predictor, performs well in the PDBP cohort. However, the multi-modal approach, trained on PPMI data, is more suited for identifying high-risk individuals at the point of diagnosis.

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
Genetics (P<1E-5)	Validation in PDBP	MLPClassifier	53.67	60.27	52.20	1.16	0.80	0.24	0.66	0.40
Clinico-demographic	Validation in PDBP	LogisticRegression	87.65	80.26	75.74	0.44	0.91	0.60	0.81	0.79
Transcriptomics (P<1E-2)	Validation in PDBP	SVC	63.62	65.50	53.32	0.65	0.97	0.09	0.65	0.67
Combined	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

Table 20: Performance metric summaries comparing best model in training in withheld samples in PPMI on PDBP validation dataset

PPMI: Parkinson's progression marker initiative; PDBP: Parkinson's disease biomarker program; AUC: Area under the curve; Log loss: Logarithmic loss; PPV: Positive Predictive Value; NPV: Negative Predictive Value; MLPClassifier: Multi-Layer Perceptron Classifier; SVC: Support Vector Classifier; AdaBoost Classifier: AdaBoost (Adaptive Boosting) Classifier.

Furthermore, chromosomes X and Y were not included in the AMP PD version 1 release, and recent research suggests no correlation between autosomal risk factors and PD sex differences (Blauwendraat et al. 2021). The combined model, incorporating both dynamic (age, UPSIT, RNA sequencing) and static

(family history, genetics) features, is most accurate at diagnosis, having been trained on newly diagnosed, imaging-confirmed, unmedicated PPMI cases. This precision makes it an ideal tool for healthcare systems to identify at-risk individuals for monitoring and potentially nominating them for low-cost preventative interventions or prodromal clinical trial enrollments. Additionally, combining clinical and algorithmic insights to refine participant group phenotypes could enhance trial recruitment efficacy. Since this model is designed to target early-stage PD, it holds promise in facilitating timely interventions or treatments, potentially preventing irreversible damage. By identifying large pools of at-risk individuals, this approach enables proactive follow-up and monitoring, anticipating symptom onset and aiding in early intervention (Leonard et al. 2020).

Exploring the Top Predictive Features

Our team have developed an interactive web-based application to facilitate the exploration of key factors in this PD prediction model, which incorporates multiple data modalities. Some of these metrics, abbreviated in **Supplementary Table 20**, provides users with the flexibility to explore various model variations, including transcriptomics-only models and models excluding clinico-demographic features. A significant feature of this application is its ability to generate decision plots. These plots are particularly insightful, as they allow users to understand the reasoning behind the classification of individuals, especially those who were challenging to categorize as either PD cases or controls.

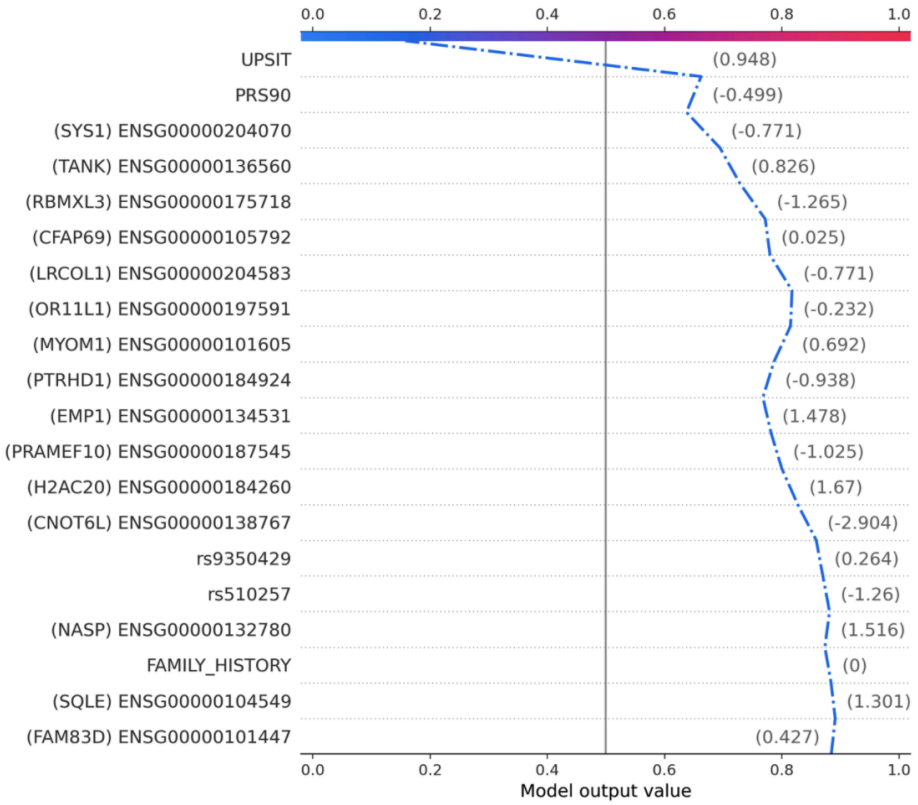


Figure 35: Misclassified case as a healthy control using the best multi-modal model

UPSIT: University of Pennsylvania Smell Identification Test; PRS: Polygenic risk score

My analysis highlights that the UPSIT score is a predominant factor in determining whether an individual is classified as a PD case or a healthy control (**Figure 33**). However, it's important to note that while UPSIT is a general indicator of neurodegeneration, incorporating diverse data types, especially genetic data, enhances disease specificity. For instance, one decision plot illustrates an individual clinically diagnosed with PD being initially classified as a case by the model, but an unexpectedly high UPSIT score led to a misclassification as a healthy control (**Figure 35**). This exemplifies the complex role UPSIT plays in this model, being both advantageous and challenging in terms of model performance. My analysis also highlights the impact of genes and variants on the model's performance and potentially on PD biology. Many top features are transcriptomic, likely due to the PRS capturing a significant portion of the genetic aspects of PD risk. Genetics data, being static and unchanging over time, contrasts with the dynamic nature of clinical or transcriptomic data. These genetic contributions are numerous but individually have smaller effects compared to the few but large effects seen in time-varying clinical data.

The genetics and PRS data in this model, derived from well-established knowledge of PD, may result in some degree of overfitting. However, the top-ranked features from the transcriptomics data offer intriguing biological insights. For example, the expression of *HS3ST3A1*, implicated in alpha-synuclein aggregation in PD cellular models, has recently been associated with white matter hyperintensity and cognitive decline in the elderly (Lehri-Boufala et al. 2015; "Genome-Wide Association Study of White Matter Hyperintensity Volume in Elderly Persons without Dementia" 2020). *OTOL1*, another top feature, is a potential genetic modifier of familial PD (Hill-Burns et al. 2016). *CHFR* has been associated with rotenone-related PD risk (Cabeza-Arvelaiz and Schiestl 2012) and *CASP7*'s involvement in apoptosis and neuroprotection, along with its link to late-onset familial Alzheimer's disease (X. Zhang et al. 2019; Magalingam et al. 2015), are also notable.

Furthermore, the genetic variant rs4238361 near the coding gene *VPS13C*, known for harboring PD variants (Lesage et al. 2016; Rudakou et al. 2020), and the gene *PHF14*, which is suggested to be downregulated in neurodegenerative diseases (Ibáñez et al. 2014), are of interest. Recent studies connecting *SQLE* with dopamine stress responses in PD ("Single-Cell Transcriptomics of Parkinson's Disease Human In Vitro Models Reveals Dopamine Neuron-Specific Stress Responses" 2020) and the suggested association of *MMP9* overexpression with neuronal cell death in neurodegeneration (He et al. 2013) indicate plausible biological relevance of these findings.

Leveraging Network Analysis in Therapeutic and Biomarker Research

In this study, I have found that incorporating feature selection from model building into network community analyses results in the creation of networks with relatively low bias. This is in contrast to networks derived from literature and text mining, which often contain inherent biases (Gillis, Ballouz, and Pavlidis 2014; Haynes, Tomczak, and Khatri 2018). Notably, the networks formed through this method reveal nodes that indicate potential shared effects in genetically targeted drugs. This insight is crucial for drug development, as drugs linked to genetic or genomic data generally demonstrate higher success rates in clinical trials (Nelson et al. 2015; King, Davis, and Degner 2019). In addition to building classifiers, these modeling techniques aid in pinpointing drug targets during feature selection and

network construction phases. The network community we constructed from case-only expression data included only two quantitative trait loci identified in blood from Mendelian randomization in previously published PD GWAS (**Supplementary Table 23**). These two genes are *ZBTB4* and *FCF1*. Moreover, our analysis of transcriptomic data revealed an enrichment of drug targets within the genes nominated. These networks, based on genes highly correlated in PD cases, naturally led to clusters of genes with similar expression patterns among these cases.

Of note were the overrepresentation of genes targeted by existing drugs within these network communities. For instance, we identified an interaction between gamma hydroxybutyric acid and the genes *SLC16A7*, *SLC16A3*, and *GABBR1*. The *SLC16A7* and *SLC16A3* genes belong to a family of drug transporter genes known as monocarboxylate transporters, which are involved in various stages, including the absorption, distribution, and elimination of drug molecules (Halestrap 2013). The *GABBR1* gene, which encodes a receptor for gamma-aminobutyric acid (*GABA*) and is implicated in several neurobehavioral disorders, interacts with gamma hydroxybutyric acid, a GABA-B receptor agonist (Halestrap 2013; Ngo and Vo 2019). This interaction suggests potential pathways for therapeutic intervention and optimization.

Overcoming Challenges and Embracing Innovations

In conducting this research, the primary limitation was the lack of diversity in our sample series. The effectiveness of genetic predictive models is known to vary significantly across different genetic ancestry groups, as indicated by current research (A. R. Martin et al., n.d.). To address this in future studies, I plan to leverage the resources of the GP2 and federated learning (Global Parkinson's Genetics Program 2021; (Riley, AU Riley, and Schekman 2021). This collaboration aims to expand our modeling efforts to include a wider array of genetic ancestry groups and generally larger sample series.

Furthermore, I acknowledge that finding an optimal dataset for validating our findings is challenging due to the specific design of PPMI, which focuses on unmedicated, recently diagnosed PD cases. Additionally, significant predictors such as constipation and REM sleep behavior disorder (RBD) were not included in our analysis. In previous studies, constipation did not meet the criteria for feature selection and was consequently omitted (Nalls et al., 2015). Although RBD is a recognized predictor of PD, comparable to the smell identification test, its data was available only for a subset of samples, which could have led to issues with sparse data and data harmonization challenges. The RBD questionnaire has also been deemed insufficient for screening idiopathic PD (Halsband et al. 2018). However, the ongoing expansion of the PPMI study is expected to facilitate further advancements in this area.

While advanced computational techniques may promise improved accuracy, they also introduce increased complexity, reduced interpretability, and potential implementation challenges. The scientific and medical communities value parsimony - the idea that simpler models should be preferred when they adequately describe the phenomenon at hand. Therefore, the marginal benefits of ML in this setting should be carefully weighed against the added complexity and resource requirements. Nonetheless, these findings do not negate ML's potential utility in clinical settings. Rather, they emphasize that ML's

practical value in healthcare depends on achieving a balance of improved predictive accuracy, user-friendliness, computational efficiency, and cost-effectiveness. Optimal clinical implementation of ML should enhance decision-making while seamlessly integrating into existing workflows without imposing undue burdens on healthcare providers or systems.

Despite these challenges, I believe this work provides a conceptual and practical framework building on previous efforts. The combined multi-modal classifier exhibits improved performance, broader applicability, and high reproducibility. Moreover, the transparency of our approach and the inclusion of various data types move us towards more clear and interpretable models.

This research represents a study that is part of the shift away from the traditional focus on single biomarkers in biomedical research. By using and competing several ML algorithms to explore complex relationships between features, I have attempted to maximize data value across multiple modalities. These models in the future that are developed with growing diverse datasets will be ethical and generalizable to the population, enhancing risk prediction in PD for precision medicine.

Chapter 6: Conclusions and Horizons in PD Genetics Research

PD is a complex neurodegenerative disorder expected to affect 12-17 million by 2040 (The Lancet 2024), and presents a significant challenge in healthcare due to its varied manifestations and unclear etiology. Through comprehensive analysis, I aimed to uncover the genetic landscape of PD and its implications for diagnosis and treatment. My dissertation evaluates the genetic underpinnings of PD, focusing on the examination of rare variants, structural variants, common variants, and culminates in the development of advanced diagnostic model leveraging ML and multi-omics data.

A pivotal aspect of understanding PD's genetic landscape is the study of specific genes implicated in the disease. One such gene is *SNCA*, known for its significant role in PD pathology (Morris et al. 2024). In Chapter 2, I investigated the frequency and nature of *SNCA* structural variants, such as duplications, deletions, and complex rearrangements, using data from the UK Biobank. Despite *SNCA*'s known links to familial and sporadic PD, there remains an unclear impact of these variants in the general population. Notably, none of the participants with known pathogenic *SNCA* missense mutations or CNVs had PD, though some had prodromal symptoms or a noted family history of the disease. This could be due to early sampling bias of the UK biobank, as it does not have targeted disease recruitment.

I also reported potential mosaic events in *SNCA*. While these have been linked to PD in other studies, their role in disease progression remains uncertain. Interestingly, a significant number of individuals with mosaic *SNCA* events also had blood-based cancers, showing the connection between large genetic changes in DNA derived from blood and cancer risk. However, the effects of these mosaic events on gene and protein expression, especially in relation to PD, are still not fully understood. These findings highlight the need for further research to fully grasp the implications of *SNCA* variation and mosaicism in PD. This could involve monitoring individuals with these genetic alterations over time to see if PD symptoms develop or including them in clinical trials for preventative therapies. Another limitation of the study was relying on EHRs for PD diagnosis and the potential of overlooking smaller or more complex genetic variations. EHRs, particularly those based on ICD-10 coding, might not capture every aspect of a patient's health history, especially more subtle symptoms or conditions that are not routinely coded. This could lead to underreporting or misclassification of PD cases. Another major challenge was the lack of immediate access to DNA samples, which restricted our ability to use Multiplex Ligation-dependent Probe Amplification (MLPA) for direct validation of identified CNVs. Instead, we relied on exome sequencing data, which, while informative, is not as targeted for CNV detection as MLPA. This meant that the validation of genomic events was only partial. Furthermore, my methodological approach, which included manual interpretation of visualized genetic data, raised the risk of missing certain types of genetic variations. Smaller deletions or complex genomic rearrangements, which might be crucial in understanding PD but less apparent in the data, could have been inadvertently overlooked.

Building on the limitations identified in my study, there are several areas for future work and improvements. Daida and colleagues highlight the importance of long-read sequencing in identifying

complex structural variants, especially in the *PRKN* gene related to young-onset PD. Their discovery of a novel inversion in *PRKN* in monozygotic twins with dystonia-parkinsonism illustrates the crucial role of advanced sequencing techniques in resolving intricate genetic variations in PD (Daida et al. 2023). There is a need to develop more advanced screening techniques that can accurately detect structural variants in *SNCA*, and other genes (such as *PRKN*) across diverse populations. Longitudinal studies are also crucial. By conducting studies that track the progression of PD in individuals with identified *SNCA* variants over time, we can gain a deeper understanding of how these genetic alterations influence the development and progression of the disease. This would enhance our ability to identify individuals at risk for PD more efficiently and inclusively and eventually prioritize them for targeted therapies for more effective treatment.

Looking forward, it would be beneficial to replicate this *SNCA* analysis using the newly released UK Biobank data as well as the AllofUs data, which might provide additional insights given the larger sample size and confirm previous findings. Relevant work by Billingsley and colleagues provide a significant contribution to this field. Their study focused on identifying structural variants associated with PD risk using whole-genome sequencing samples. They discovered three novel deletion SVs linked to PD, including a significant deletion within *LRRN4 (Billingsley et al. 2023)*. This research underscores the value of genome-wide analysis in understanding PD genetic risk and complements my study's findings. There is also the potential in leveraging automated, machine-learning-powered algorithms for enhanced CNV detection, prioritization, and nomination across different ancestries in a large-scale manner across all genes, an active project we have at GP2. This approach could address the current limitation of potentially missing smaller or complex variations.

Moving from the study of structural variants, I shift my focus to the impact of rare genetic variants on PD. Beyond the focus on specific genes like *SNCA*, I conduct a comprehensive genetic analysis on the impact of rare variants in all genes, which might hold pivotal insights into the multifaceted nature of PD.

In Chapter 3, I analyzed a large sample comprising 7,184 PD cases, 6,701 proxy-cases, and 51,650 controls to conduct rare variant gene burden tests for PD. My meta-analysis reinforced the association of rare variants in *GBA1* and *LRRK2* with PD risk in individuals of European ancestry. Additionally, I identified several novel PD-associated genes, including *B3GNT3*, *AUNIP*, *ADH5*, *TUBA1B*, *OR1G1*, *CAPN10*, and *TREML1*, though their significance varied across datasets. A key finding was the strong evidence of a novel rare variant association in *B3GNT3*, primarily driven by the Genentech and UK Biobank parent proxy datasets. However, these variants were absent in the AMP-PD and NIH genomes, suggesting the need for more data to confirm their PD risk association. The rarity of these variants and their absence in other datasets indicate the challenges in studying rare genetic variants in PD and underscore the need for further validation.

I also assessed rare variants in previously suggested PD GWAS loci, but found no significant associations. This aligns with similar findings in East Asian populations and raises questions about the mechanisms at these loci. Notably, genes commonly associated with autosomal recessive early-onset PD, such as *PINK1* and *PRKN*, were absent in my study, likely due to the algorithms used being more attuned to autosomal

dominant and high-risk variants. The rarity of certain disease-causing variants, like pathogenic missense variants in *SNCA*, also presented a challenge, as these mutations are scarce in the broader population and likely underrepresented in my datasets. My results led me to investigate the role of immune response and microtubule defects in PD, highlighting genes like *B3GNT3*, *TUBA1B*, and *TREML1*. *B3GNT3*'s role in lymphocyte homing and *TUBA1B*'s involvement in microtubule function suggest their potential connection to PD pathology. *TREML1*'s increasing implication in neurodegenerative disorders and *ADH5*'s controversial association with PD risk also warrant further exploration. However, the connections between PD and the functions of *AUNIP*, *OR1G1*, and *CAPN10* remain unclear, necessitating additional studies with genetic support and functional data.

Despite being the largest effort to identify rare genetic variants in PD, my study had limitations, primarily its focus on individuals of European ancestry. Future research should include diverse populations and various age-at-onset ranges. The study's limitation to four variant classes may overlook other disease mechanisms, such as gain-of-function mutations, as those are more difficult to detect. Additionally, the study's power to detect associations in genes with a small percentage of functional variants was limited. The absence of certain genes in my analysis, notably *PINK1* and *PRKN*, anticipated due to the nature of the burden testing algorithms, highlights the need for more comprehensive approaches. The clinical heterogeneity within PD cases also necessitates further validation of the pathogenicity of rare or ultra-rare variants. In conclusion, while my study reaffirms the role of *GBA1* and *LRRK2* in PD and nominates several new genes, it emphasizes the need for more research, particularly focusing on familial PD cases and diverse populations, to deepen our understanding of PD genetics.

In the future, the enhancement of imputation of rarer variants with greater confidence would improve a future iteration of this study. This will involve broadening genetic screening to encompass a global perspective on PD, recognizing that rare genetic variants may be relevant across diverse populations. Incorporating detailed clinical phenotyping with genomic data regarding rare variants is a critical step, as that will establish the foundation for scalable precision medicine strategies and potential for personalized treatment. This integration will enable a better understanding of genotype-phenotype correlations, aiding in the development of therapeutics targeting pathways affected by identified rare variants.

Extending the analysis of rare variants to include underrepresented populations once enough samples are collected will be crucial. This expansion is vital to uncover novel variants and associations that might be specific to these groups. Advanced computational predictors, like AlphaMissense (Cheng et al. 2023) and AlphaFold (Jumper et al. 2021), will be employed to predict the immediate functional impact of rare genetic variants in tandem with existing prediction tools. Improved imputation of rare variants, scalable and ancestry-aware, will power statistical tests and enhance their accuracy. The use of population-specific reference panels, as demonstrated in studies with Estonian individuals, shows significant promise in improving imputation confidence and accuracy for low-frequency and rare variants (Mitt et al. 2017). Additionally, it is worth exploring advanced statistical methods like the aggregated Cauchy association test (ACAT) for its robust power in sequencing studies, especially in situations where only a small number of variants are causal (Liu et al. 2019).

The work in this study has also been replicated by Pitz and colleagues, where they assessed over 27,590 PD cases and 3 million controls, illustrating the complexities of studying rare variants in large-scale cohorts. They successfully replicated five significant variants, including those in *GBA1* and *LRRK2*, and identified eight strong candidate variants for association with PD (Pitz et al. 2024). This approach underscores the importance of large databases in confirming and discovering new variants linked to PD. By building upon these methodologies and findings, my future research aims to expand our understanding of rare variants in PD, overcoming the current limitations and paving the way for more effective diagnosis and treatment strategies.

I extend the genetic investigation from rare to common variants, addressing the gap in genetic diversity, crucial for a holistic understanding of PD. The multi-ancestry meta-analysis and fine-mapping in Alzheimer's Disease (Lake et al. 2023) and the large-scale multi-ancestry meta-analysis in PD (J. J. Kim et al. 2024) both highlight the importance of incorporating diverse genetic backgrounds into neurodegenerative disease research. These studies have identified novel loci and fine-mapped causal variants, demonstrating the value of including diverse ancestries in genetic studies of PD and Alzheimer's disease. I explored the genetics of PD in African and African admixed populations, a relatively underexplored area. In Chapter 4, I conducted the first extensive genome-wide assessment in these populations, analyzing data from 197,918 individuals from various cohorts, including GP2 and 23andMe. I focused on identifying ancestry-specific genetic risk factors and exploring coding and structural variations. My study revealed a novel common risk factor at the *GBA1* locus, unique to African ancestry populations, that was significantly associated with both PD risk and age at onset. This discovery underscores the importance of large-scale studies in diverse populations, as larger sample sizes were needed to identify *GBA1* as a major PD risk factor in European populations. *GBA1* is a complex locus with various coding, structural, and non-coding variants affecting PD risk. Despite the robust effect size of the signal identified, no associations were found with known or newly discovered *GBA1* coding or structural variation. Instead, eQTL data suggested that the rs3115534-G risk allele is associated with increased *GBA1* expression but paradoxically corresponds to decreased GCase activity. This work highlights the importance of including diverse ancestries in genetic research and opens new avenues for RNA-based therapeutic strategies and inclusive clinical trial designs. As a follow-up to this study, the study by Ojo and colleagues on the *GBA1* variant and REM sleep behavior disorder (RBD) in the Nigerian population highlights the importance of investigating specific non-coding variants in African populations. Their findings suggest a significant association between the *GBA1* rs3115534 variant and possible RBD symptoms, even in the absence of PD, underscoring the need for more in-depth research in this area (Oluwadamilola Omolara Ojo et al. 2024).

Despite its contributions, my study faces limitations such as the need for larger cohorts and the challenges posed by multi-mapping reads between *GBA1* and its pseudogene, *GBAP1*. These limitations highlight the necessity for future research to comprehensively explore susceptibility genetic risk and phenotypic relationships. In conclusion, this research provides vital GWAS-based insights into PD genetics in African and African admixed populations and sets the stage for future clinical trials and therapeutic interventions targeting *GBA1* and other PD-related genes. Future studies should expand

cohort sizes, integrate omics data, and evaluate population-specific associations to enhance our understanding of PD genetics and its clinical implications.

Building on the current work, my future research will focus on expanding the understanding of local ancestry for each individual, offering insights into the unique genetic variations that occur due to the historical admixture of different populations. Additionally, GWAS in various subpopulations across Africa as well as extending these GWAS to other global populations will identify novel loci and enhance our comprehension of PD's genetic landscape. With the identification of ancestry-specific genetic risk factors, there is an opportunity to create scalable precision medicine interventions that are effective across different genetic backgrounds. This approach could lead to more personalized and effective treatments for PD patients worldwide.

In my future work, I plan to build upon the foundational efforts of GP2, as outlined in their publication (Global Parkinson's Genetics Program 2021). GP2 aims to enhance global PD genetic research, emphasizing data democratization and diversity, which aligns with my goal of expanding understanding in underrepresented populations, as they have a deep commitment to integrate cohorts into GP2 for global PD research, detailed in their protocol developed by Towns and colleagues (Towns et al. 2023).
GP2 uses the Illumina NeuroBooster array, developed for the GP2 and the Center for Alzheimer's and Related Dementias, as described by Bandrés-Ciga and colleagues (Sara Bandres-Ciga et al. 2023). This tool is designed to address disparities in representing diverse genetic variations in neurological research and will be instrumental in identifying low-frequency variants and imputing common variants across various populations.

Building on the technological and methodological advancements facilitated by GP2, the need for further refining our understanding of genetic risk factors across diverse ancestries becomes evident. This necessity is highlighted by the work of Saffie-Awad and colleagues, who evaluated PRS models across different populations. Their findings reveal significant disparities in risk estimates and allele patterns, emphasizing the limitations of current PRS models that are predominantly biased towards European populations (Saffie-Awad et al. 2023). This underscores the urgent need for larger, more diverse cohorts to enhance the precision and inclusivity of genetic predictions in PD research across all ancestries.

Building upon the genetic findings, I explored the application of these insights in developing generalizable ML models. This transition highlights the practical implementation of genetic research in PD, moving from theoretical understanding to actionable diagnostic tools. In Chapter 5, I address the urgent need for global, scalable, early, and precise PD diagnosis, focusing on the disease's early stages. Central to my approach is the use of multi-modal PD datasets and the development of GenoML, a custom Python package integral to creating innovative diagnostic models. Using data from AMP-PD and employing GenoML's open-source automated machine learning software, I ensure reproducibility, transparency, and adherence to open science principles. The models I developed effectively predict PD risk and identify key features for constructing unbiased genetic networks, illuminating biological pathways involved in PD onset and potential therapeutic targets. The model shows a significant improvement over previous efforts, demonstrating enhanced performance metrics in current

cross-validation and even surpassing the training phase metrics of some prior studies. In my study, I integrated adjusted transcriptomics, genetics, and clinical data into a single predictive model. This model complements existing studies that focus on imaging for predicting REM sleep behavior disorder and cognitive deficits. My study underscores the effectiveness of integrating diverse data modalities in predictive modeling, with broad implications for healthcare data utility. However, I encountered limitations due to data scarcity and the availability of potential features in public datasets. My goal is to develop a model based on data that are remotely accessible or commonly found in biobanked samples, thus avoiding the need for costly and logistically challenging specialist clinical visits. The methodology I describe not only enables accurate, early PD diagnosis but is also potentially cost-effective for biobanks and established healthcare systems. The multi-modal approach balances out various modalities, some excelling in predicting case status, others in classifying controls, making it valuable for large-scale biobank contributions and trial recruitments. Despite a high false discovery rate, the model's high sensitivity and balanced accuracy in differentiating between PD cases and controls make it an effective tool for identifying high-risk individuals for monitoring and early intervention. While the methodology demonstrates significant improvements in PD risk prediction and highlights key genetic networks involved in PD onset, it is not without limitations that suggest future research directions.

One major limitation lies in the availability and diversity of data. Despite integrating transcriptomics, genetics, and clinical data, the study faces challenges due to data scarcity and a limited range of features in public datasets. Future studies could expand the dataset to include more diverse and extensive features, potentially overcoming this limitation. Additionally, integrating other forms of data, such as environmental or lifestyle factors, could enhance the model’s predictive power. Addressing the high false discovery rate identified in the study is also crucial. Future iterations of the model should aim to reduce this rate to improve the accuracy and reliability of PD predictions. Advanced ML techniques and a more comprehensive dataset might help in achieving this.

Another direction for future research involves the model's application and integration with existing healthcare systems and studies. While the current model is less resource-intensive and shows potential for merging with other studies through techniques like transfer or ensemble learning, further research is needed to fully realize this integration. The model's usage in clinical settings warrants further exploration. While it shows promise in identifying high-risk individuals for early intervention, assessing its practicality and effectiveness in real-world healthcare settings is essential.

The next phase of this research will concentrate on several innovative methodologies and applications to enhance the precision and applicability of PD diagnosis. The use of newer and improved releases of the AMP-PD data will be crucial in this endeavor. This will ensure that the models are developed using the most up-to-date and comprehensive data available, enhancing their predictive accuracy and relevance. The development of GenoML will be expanded to include multi-class predictions, aimed at enhancing GenoML's capabilities, making it a more versatile tool in PD research and improving predictions beyond PD case and control.

In response to the need for more generalizable models, there will be an exploration of generating models with fewer features. This approach aims to increase the models' applicability across diverse populations, addressing one of the key limitations of current predictive models. Furthermore, the application of federated ML will be pivotal for analyzing locally-restricted samples and data silos, offering a solution to privacy and logistic challenges in genomic studies by continuously updating a central model with weights and redistributing the model to the silos.

Alongside this, network community clustering with transcriptomics will be employed to gain deeper insights into the pathways associated with PD. This method will help in identifying key genetic networks and the effects of their potential perturbations. The use of *in silico* network perturbations will offer a novel approach to understanding the impact of genetic changes on PD. This technique will simulate the effects of various genetic alterations, providing valuable insights into potential therapeutic targets. Additionally, a focus on proteomics and the longitudinal progression of PD across different ancestries will help in understanding the disease's progression and developing more effective treatment strategies.

Regarding personalized treatment strategies, leveraging ML to develop tailored approaches based on individual genetic and clinical profiles will be a major goal. This personalized approach promises to revolutionize PD treatment, making it more effective and reducing the likelihood of adverse effects. The integration of additional data types, such as proteomics and metabolomics, will enhance the predictive power and accuracy of the models. A global perspective will be maintained, ensuring that these advanced diagnostic tools are accessible worldwide, particularly in regions with limited resources. This holistic approach aims not just at early and precise diagnosis but also at developing targeted, effective treatment strategies for individuals suffering from PD.

The landscape of PD research is evolving rapidly, with the future pointing towards the development of novel diagnostic and therapeutic methods. Central to this progression is the focus on non-invasive diagnostic tools, like blood tests that can be carried out with relative ease. These innovative methods are targeted at detecting early PD biomarkers, potentially before the manifestation of clinical symptoms. The importance of this early detection cannot be overstated, as it allows individuals access to timely interventions that could improve their outcome.

An active area of research is understanding the interaction between environmental and lifestyle factors with genetic variants associated with PD. Studies are increasingly exploring how exposure to certain toxins, dietary habits, or levels of physical activity and how any one of these factors influence the risk and progression of PD. Quantifying and accurately predicting these factors have the potential to inform public health strategies and individual lifestyle changes, mitigating PD risk or decelerate its progression. Epidemiological studies, particularly in underrepresented populations, provide essential insights into the prevalence and incidence of PD across various ethnic and geographic groups. This research is key to uncovering population-specific risk factors and disease patterns, which in turn can inform targeted public health interventions and inform healthcare policy-making, appropriate resource allocation, and healthcare planning.

In the therapeutic domain, gene therapy emerges as a promising approach, aiming to correct genetic variants, has the potential to offer targeted and possibly curative treatments. However, its efficacy and safety require thorough investigation. Complementing this, drug repurposing offers an efficient framework to discover new PD treatments. By exploring the potential of existing drugs to modulate biological pathways implicated in PD, researchers can expedite the development of new therapeutic applications. Efforts such as OmicSynth, an open multi-omic community resource developed by Alvarado and colleagues, allows for evidence-based identification of therapeutic targets for neurodegenerative diseases. In PD, 46 target genes that pass multiple test corrections were identified, presenting new avenues for drug discovery and development. These targets are categorized based on their druggability and existing approved therapeutics, highlighting 41 novel targets, 3 known targets, and 115 difficult targets, with a significant portion expressed in disease-relevant cell types (Alvarado et al. 2024).

Another significant stride in PD research is the implementation of patient stratification in clinical trials based on genetic profiles. A study performed by Leonard and colleagues identified that genetic diversity within clinical trials that do not match participants genetically could obscure the actual effects of the therapy, as anticipated. Clinical trials should incorporate genetic considerations before starting, or minimum conduct post-trial genetic adjustments and analyses, especially in cases where the trials do not meet their intended outcomes (Leonard et al. 2020). This strategy enhances the personalization and efficacy of treatments, potentially leading to more effective and safer therapeutic options. By aligning treatments with patients' genetic backgrounds, researchers can identify which individuals are more likely to benefit from specific treatments.

Finally, global health initiatives are critical for enhancing PD care and management in low-resource settings. By leveraging genetic insights, these initiatives aim to reduce the global PD burden, facilitating the exchange of knowledge and best practices, and ultimately improving health outcomes for PD patients worldwide. This holistic approach, encompassing innovative research and global collaboration, paves the way for a future where PD management is more effective, equitable, personalized, and accessible across the globe.

A significant emphasis must be placed on advancing precision medicine for NDDs, particularly PD. Precision medicine's potential to tailor treatments based on individual genetic profiles and disease manifestations offers a promising avenue to enhance treatment efficacy and patient outcomes. By integrating patient-specific genetic data, biomarker discovery, and the elucidation of disease heterogeneity, we can move towards developing targeted interventions, finding the right patient at the right time and using the right therapeutic intervention (Leonard et al. 2024). This approach not only aligns with the goal of addressing the unique aspects of PD in diverse populations but also sets the stage for the next phase of research focused on implementing precision medicine frameworks. Such strategies will be critical in overcoming the current limitations of PD treatments and in paving the way for more personalized, effective therapeutic options.

In summary, this dissertation presents a multifaceted exploration of PD, from its genetic underpinnings to the innovative use of ML in diagnostics. It emphasizes the need for global and scalable approaches in

PD research, paving the way for precision medicine and personalized treatment strategies, contributing to a deeper understanding of the disease and opening new avenues for research and clinical application.

Manuscripts

Pre-prints and Published Works

Work directly conducted for this dissertation, in chronological order:

1. Blauwendraat C*, **Makarious MB***, Leonard HL, Bandrés-Ciga S, Iwaki H, Nalls MA, Noyce AJ, and Singleton AB: "**A Population-Scale Analysis of Rare *SNCA* Variation in the UK Biobank**" *Neurobiology of Disease (2021); https://doi.org/10.1016/j.nbd.2020.105182*
2. Lake J*, Storm CS*, **Makarious MB***, Bandrés-Ciga S*: "**Genetic and Transcriptomic Biomarkers in Neurodegenerative Diseases: Current Situation and the Road Ahead**" *Cells (2021); https://doi.org/10.3390/cells10051030*
3. **Makarious MB**, Leonard HL, Vitale D, Iwaki H, Saffo D, Sargent L, Dadi A, Salmerón Castaño E, Carter JF, Maleknia M, Botia JA, Blauwendraat C, Campbell RH, Hashemi SH, Singleton AB, Nalls MA, Faghri F: "**GenoML: Automated Machine Learning for Genomics**" *arXiv (2021); https://arxiv.org/abs/2103.03221*
4. **Makarious MB**, Leonard HL, Vitale D, Iwaki H, Sargent L, Dadu A, Violich I, Hutchins E, Saffo D, Bandrés-Ciga S, Kim JJ, Song Y, Maleknia M, Bookman M, Nojopranoto W, Campbell RH, Hashemi SH, Botia JA, Carter JF, Craig DW, Keuren-Jensen KV, Morris HR, Hardy JA, Blauwendraat C, Singleton AB, Faghri F, Nalls MA: "**Multi-Modality Machine Learning Predicting Parkinson's Disease**" *npj Parkinson's disease (2022); https://doi.org/10.1038/s41531-022-00288-w*
5. **Makarious MB**, Lake J, Pitz V, Fu AY, …, Beach TG, Serrano GA, Real R, Morris HR, Ding J, Gibbs RG, Singleton AB, Nalls MA, Bhangale T, Blauwendraat C: "**Large-scale Rare Variant Burden Testing in Parkinson's Disease**" *Brain (2023); https://doi.org/10.1093/brain/awad214*
6. Rizig M*, Bandrés-Ciga S*, **Makarious MB***, Oluwadamilola O, Wild Crea P, Abiodun O, Levine KS, …, Nigeria Parkinson Disease Research Network, International Parkinson's Disease Genomics Consortium - Africa, Black and African American Connections to Parkinson's Disease (BLAAC PD) Study Group, the 23andMe Research Team, Blauwendraat C, Houlden H, Singleton AB, Okubadejo N, Global Parkinson's Genetics Program: "**Genome-wide Association Identifies Novel Etiological Insights Associated with Parkinson's Disease in African and African Admixed Populations**" *Lancet Neurology* (2023); https://doi.org/10.1016/S1474-4422(23)00283-1

Related Pre-prints and Published Works

Work conducted during and related to this dissertation, in chronological order:

7. The Global Parkinson's Genetics Program (GP2): "**GP2: The Global Parkinson's Genetics Program**" *Movement Disorders (2021); https://doi.org/10.1002/mds.28494*
8. Iwaki H, Leonard HL, **Makarious MB**, Bookman M, …, AMP PD Whole Genome Sequencing Working Group on behalf of the AMP PD Consortium: "**Accelerating Medicines Partnership -**

Parkinson's Disease: Genetic Resource" *Movement Disorders (2021);* *https://doi.org/10.1002/mds.28549*

9. Billingsley KJ, Ding J, Alvarez Jerez P, Illarionova A, Grenn FP, **Makarious MB**, Moore A, …, Nalls MA, Mahmoud M, Sedlazeck FJ, Blauwendraat C, Gibbs JR, and Singleton AB: "**Genome-Wide Analysis of Structural Variants in Parkinson Disease**" *Annals of Neurology (2023);* *https://doi.org/10.1002/ana.26608*
10. Towns C, Richer M, Jasaityte S, Stafford E, Joubert J, Antar T, Carrasco-Martinez A, **Makarious MB**, Casey B, Vitale D, Levine K, Leonard HL, Wegel C, Solle J, Nalls MA, Blauwendraat C, Singleton AB, Tan MX, Iwaki H, Morris H on behalf of the Global Parkinsons Genetics Program: "**Global Parkinson's Genetics Program (GP2) Complex Network Protocol: Defining the Causes of Sporadic Parkinson's Disease**" medRxiv (2023); https://www.medrxiv.org/content/10.1101/2022.11.25.22282764v2; Accepted at npj Parkinson's disease
11. Lake J, Solsberg CW, Kim JJ, Acosta-Uribe J, **Makarious MB**, Li Z, Levine K, Heutink P, …, Singleton AB, Blauwendraat C, Nalls MA, Yokoyama JS, Leonard HL: "**Multi-ancestry meta-analysis and fine-mapping in Alzheimer's Disease**" *Molecular Psychiatry (2023);* *https://doi.org/10.1038/s41380-023-02089-w*
12. Koretsky M, Alvarado C, **Makarious MB**, Vitale D, Levine K, Sargent L, Blauwendraat C, Singleton AB, Nalls MA, Leonard HL: "**Genetic Risk Factor Clustering Within and Across Neurodegenerative Diseases**" *Brain (2023);* *https://doi.org/10.1093/brain/awad161*
13. Daida K, Funayama M, Billingsley KJ, Malik L, Miano-Burkhardt A, Leonard HL, **Makarious MB**, Iwaki H, Ding J, Gibbs RJ, Ishiguro M, Yoshino H, Ogaki K, Oyama G, Nishioka K, Nonaka R, Akamatsu W, Blauwendraat C, Hattori N: **"Long-Read Sequencing Resolves a Complex Structural Variant in *PRKN* Parkinson's Disease"** *Movement Disorders (2023);* https://doi.org/10.1002/mds.29610
14. Bandrés-Ciga S, Faghri F, Majounie E, Koretsky MJ, Kim JJ, Levine KS, Leonard HL, **Makarious MB**, …, Singleton AB, Nalls MA, Jeff J, Vitale D on behalf of the Global Parkinson's Genetics Program and the Center for Alzheimer's Disease and Related Dementias: **"NeuroBooster Array: A Genome-Wide Genotyping Platform to Study Neurological Disorders Across Diverse Populations"** *bioRxiv (2023);* https://doi.org/10.1101/2023.11.06.23298176
15. Saffie-Awad P, Elsayed I, Sanyaolu AO, Wild Crea P, Schumacher Schuh AF, …, **Makarious MB**, Mata IF, Bandres-Ciga S on behalf of the Global Parkinson's Genetics Program (GP2): **"Evaluating the Performance of Polygenic Risk Profiling Across Diverse Ancestry Populations in Parkinson's Disease"** *bioRxiv (2023);* https://doi.org/10.1101/2023.11.28.23299090
16. Kim JJ, Vitale D, Otani DV, Lian M, …, **Makarious MB**, Tan EK, Singleton AB, Blauwendraat C, Nalls MA, Foo JN, Mata I: **"Meta-Ancestry Genome-Wide Meta-Analysis in Parkinson's Disease"** *Nature Genetics (2023)*; https://doi.org/10.1038/s41588-023-01584-8
17. Alvarado CX, **Makarious MB**, Weller CA, Vitale D, Koretsky M, Bandres-Ciga S, Iwaki H, Levine K, Singleton AB, Faghri F, Nalls MA, Leonard HL: **"omicSynth: An Open Multi-omic Community Resource for Identifying Druggable Targets across Neurodegenerative Diseases"** *AJHG (2024)*; https://doi.org/10.1016/j.ajhg.2023.12.006

18. Pitz V, **Makarious MB**, Bandrés-Ciga S, Iwaki H, 23andMe Research Team, Singleton AB, Nalls MA, Heilbron K, Blauwendraat C: "**Associations of Rare Parkinson's Disease Variants in Millions of People**" *npj Parkinson's (2024);* https://doi.org/10.1038/s41531-023-00608-8
19. Ojo OO, Bandrés-Ciga S, **Makarious MB**, Wild Crea P, Hernandez DG, Houlden H, Rizig M, Singleton AB, Noyce AJ, Nalls MA, Blauwendraat C, Okubadejo NU on behalf of the Nigeria Parkinson's Disease Research Network and the Global Parkinson's Genetics Program (GP2): **"The non-coding *GBA1* rs3115534 Variant is Associated with REM Sleep Behavior Disorder in the Nigerian Population"** *Movement Disorders (2024)*; https://doi.org/10.1002/mds.29753
20. Alvarez Jerez A, Wild Crea P, Ramos DM, Gustavsson EK, Radefeldt M, **Makarious MB**, Ojo OO, Billingsley KJ, Malik L, Daida K, Bromberek S, Hu C, Schneider Z, ..., Singleton AB, Ward M, Okubadejo NU, Blauwendraat: **"African Ancestry Neurodegeneration Risk Variant Disrupts an Intronic Branchpoint in *GBA1*"** *medRxiv (2024)*; https://doi.org/10.1101/2024.02.20.24302827
21. Danek B, **Makarious MB**, Dadu A, Vitale D, Nalls MA, Sun J, Faghri F: **"Federated Learning for Multi-omics: A Performance Evaluation in Parkinson's Disease"** Cell Patterns (2024); https://doi.org/10.1016/j.patter.2024.100945

References

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74.

Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. “Tensorflow: A System for Large-Scale Machine Learning.” In *12th ${USENIX} Symposium on Operating Systems Design and Implementation ({OSDI}$ 16)*, 265–83.

Alvarado, Chelsea X., Mary B. Makarious, Cory A. Weller, Dan Vitale, Mathew J. Koretsky, Sara Bandres-Ciga, Hirotaka Iwaki, et al. 2024. “omicSynth: An Open Multi-Omic Community Resource for Identifying Druggable Targets across Neurodegenerative Diseases.” *American Journal of Human Genetics* 111 (1): 150–64.

Angermueller, Christof, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. 2016. “Deep Learning for Computational Biology.” *Molecular Systems Biology* 12 (7): 878.

Backman, Joshua D., Alexander H. Li, Anthony Marcketta, Dylan Sun, Joelle Mbatchou, Michael D. Kessler, Christian Benner, et al. 2021. “Exome Sequencing and Analysis of 454,787 UK Biobank Participants.” *Nature* 599 (7886): 628–34.

Bailey, Meagan, Lisa M. Shulman, Diane Ryan, Bichun Ouyang, Joshua M. Shulman, Aron S. Buchman, David A. Bennett, Lisa L. Barnes, and Deborah A. Hall. 2021. “Frequency of Parkinsonism and Parkinson Disease in African Americans in the Chicago Community.” *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 76 (7): 1340–45.

Balwani, Manisha, Laura Fuerstman, Ruth Kornreich, Lisa Edelmann, and Robert J. Desnick. 2010. “Type 1 Gaucher Disease: Significant Disease Manifestations in ‘Asymptomatic’ Homozygotes.” *Archives of Internal Medicine* 170 (16): 1463–69.

Bandres-Ciga, Sara, Monica Diez-Fairen, Jonggeol Jeff Kim, and Andrew B. Singleton. 2020. “Genetics of Parkinson’s Disease: An Introspection of Its Journey towards Precision Medicine.” *Neurobiology of Disease* 137 (April): 104782.

Bandres-Ciga, Sara, Faraz Faghri, Elisa Majounie, Mathew J. Koretsky, Jeffrey Kim, Kristin S. Levine, Hampton Leonard, et al. 2023. “NeuroBooster Array: A Genome-Wide Genotyping Platform to Study Neurological Disorders Across Diverse Populations.” *medRxiv : The Preprint Server for Health Sciences*, November. https://doi.org/10.1101/2023.11.06.23298176.

Bandres-Ciga, S., S. Saez-Atienzar, J. J. Kim, M. B. Makarious, F. Faghri, M. Diez-Fairen, H. Iwaki, et al. 2020. “Large-Scale Pathway Specific Polygenic Risk and Transcriptomic Community Network Analysis Identifies Novel Functional Pathways in Parkinson Disease.” *Acta Neuropathologica* 140 (3): 341–58.

Ben-Arie, N., D. Lancet, C. Taylor, M. Khen, N. Walker, D. H. Ledbetter, R. Carrozzo, K. Patel, D. Sheer, and H. Lehrach. 1994. “Olfactory Receptor Gene Cluster on Human Chromosome 17: Possible Duplication of an Ancestral Receptor Repertoire.” *Human Molecular Genetics* 3 (2): 229–35.

Billingsley, Kimberley J., Jinhui Ding, Pilar Alvarez Jerez, Anastasia Illarionova, Kristin Levine, Francis P. Grenn, Mary B. Makarious, et al. 2023. “Genome-Wide Analysis of Structural Variants in Parkinson Disease.” *Annals of Neurology*, January. https://doi.org/10.1002/ana.26608.

Blanckenberg, Janine, Soraya Bardien, Brigitte Glanzmann, Njideka U. Okubadejo, and Jonathan A. Carr. 2013. “The Prevalence and Genetics of Parkinson’s Disease in Sub-Saharan Africans.” *Journal of the Neurological Sciences* 335 (1-2): 22–25.

Blauwendraat, Cornelis, Faraz Faghri, Lasse Pihlstrom, Joshua T. Geiger, Alexis Elbaz, Suzanne Lesage, Jean-Christophe Corvol, et al. 2017. “NeuroChip, an Updated Version of the NeuroX Genotyping Platform to Rapidly Screen for Variants Associated with Neurological Diseases.” *Neurobiology of Aging* 57 (September): 247.e9–247.e13.

Blauwendraat, Cornelis, Karl Heilbron, Costanza L. Vallerga, Sara Bandres-Ciga, Rainer von Coelln, Lasse Pihlstrøm, Javier Simón-Sánchez, et al. 2019. “Parkinson’s Disease Age at Onset Genome-Wide Association Study: Defining Heritability, Genetic Loci, and α-Synuclein Mechanisms.” *Movement Disorders: Official Journal of the Movement Disorder Society* 34 (6): 866–75.

Blauwendraat, Cornelis, Hirotaka Iwaki, Mary B. Makarious, Sara Bandres-Ciga, Hampton L. Leonard, Francis P. Grenn, Julie Lake, et al. 2021. “Investigation of Autosomal Genetic Sex Differences in Parkinson’s Disease.” *Annals of Neurology*, April. https://doi.org/10.1002/ana.26090.

Blauwendraat, Cornelis, Demis A. Kia, Lasse Pihlstrøm, Ziv Gan-Or, Suzanne Lesage, J. Raphael Gibbs, Jinhui Ding, et al. 2018. “Insufficient Evidence for Pathogenicity of SNCA His50Gln (H50Q) in Parkinson’s Disease.” *Neurobiology of Aging* 64 (April): 159.e5–159.e8.

Blauwendraat, Cornelis, Mike A. Nalls, and Andrew B. Singleton. 2020. “The Genetic Architecture of Parkinson’s Disease.” *Lancet Neurology* 19 (2): 170–78.

Book, Adam, Ilaria Guella, Tara Candido, Alexis Brice, Nobutaka Hattori, Beomseok Jeon, Matthew J. Farrer, and SNCA Multiplication Investigators of the GEoPD Consortium. 2018. “A Meta-Analysis of α-Synuclein Multiplication in Familial Parkinsonism.” *Frontiers in Neurology* 9 (December): 1021.

Bray, Steven M., Jennifer G. Mulle, Anne F. Dodd, Ann E. Pulver, Stephen Wooding, and Stephen T. Warren. 2010. “Signatures of Founder Effects, Admixture, and Selection in the Ashkenazi Jewish Population.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (37): 16222–27.

Brice, Alexis. 2005. “Genetics of Parkinson’s Disease: LRRK2 on the Rise.” *Brain: A Journal of Neurology*.

Buervenich, Silvia, Andrea Carmine, Dagmar Galter, Haydeh N. Shahabi, Bo Johnels, Björn Holmberg, Jarl Ahlberg, et al. 2005. “A Rare Truncating Mutation in ADH1C (G78Stop) Shows Significant Association with Parkinson Disease in a Large International Sample.” *Archives of Neurology* 62 (1): 74–78.

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, et al. 2018. “The UK Biobank Resource with Deep Phenotyping and Genomic Data.” *Nature* 562 (7726): 203–9.

Cabeza-Arvelaiz, Yofre, and Robert H. Schiestl. 2012. “Transcriptome Analysis of a Rotenone Model of Parkinsonism Reveals Complex I-Tied and -Untied Toxicity Mechanisms Common to Neurodegenerative Diseases.” *PloS One* 7 (9): e44700.

Calogero, Alessandra M., Samanta Mazzetti, Gianni Pezzoli, and Graziella Cappelletti. 2019. “Neuronal Microtubules and Proteins Linked to Parkinson’s Disease: A Relevant Interaction?” *Biological Chemistry* 400 (9): 1099–1112.

Campabadal, Anna, Barbara Segura, Carme Junque, and Alex Iranzo. 2021. “Structural and Functional Magnetic Resonance Imaging in Isolated REM Sleep Behavior Disorder: A Systematic Review of Studies Using Neuroimaging Software.” *Sleep Medicine Reviews*. https://doi.org/10.1016/j.smrv.2021.101495.

Campbell, Meghan C., Peter S. Myers, Alexandra J. Weigand, Erin R. Foster, Nigel J. Cairns, Joshua J. Jackson, Christina N. Lessov-Schlaggar, and Joel S. Perlmutter. 2020. “Parkinson Disease Clinical Subtypes: Key Features & Clinical Milestones.” *Annals of Clinical and Translational Neurology* 7 (8): 1272–83.

Cartelli, Daniele, Alessandro Aliverti, Alberto Barbiroli, Carlo Santambrogio, Enzio M. Ragg, Francesca V. M. Casagrande, Francesca Cantele, et al. 2016. “α-Synuclein Is a Novel Microtubule Dynamase.” *Scientific Reports* 6 (September): 33289.

Chang, Christopher C., Carson C. Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. 2015. “Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets.” *GigaScience* 4 (February): 7.

Chang, Diana, Mike A. Nalls, Ingileif B. Hallgrímsdóttir, Julie Hunkapiller, Marcel van der Brug, Fang Cai, International Parkinson’s Disease Genomics Consortium, et al. 2017. “A Meta-Analysis of

Genome-Wide Association Studies Identifies 17 New Parkinson's Disease Risk Loci." *Nature Genetics* 49 (10): 1511–16.

Chartier-Harlin, Marie-Christine, Jennifer Kachergus, Christophe Roumier, Vincent Mouroux, Xavier Douay, Sarah Lincoln, Clotilde Levecque, et al. 2004. "Alpha-Synuclein Locus Duplication as a Cause of Familial Parkinson's Disease." *The Lancet* 364 (9440): 1167–69.

Cheng, Jun, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, et al. 2023. "Accurate Proteome-Wide Missense Variant Effect Prediction with AlphaMissense." *Science* 381 (6664): eadg7492.

Chen-Plotkin, Alice S. 2018. "Parkinson Disease: Blood Transcriptomics for Parkinson Disease?" *Nature Reviews. Neurology* 14 (1): 5–6.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. KDD '16. New York, NY, USA: ACM.

Choudhury, Ananyo, Shaun Aron, Laura R. Botigué, Dhriti Sengupta, Gerrit Botha, Taoufik Bensellak, Gordon Wells, et al. 2020. "High-Depth African Genomes Inform Human Migration and Health." *Nature* 586 (7831): 741–48.

Cilia, Roberto, Francesca Sironi, Albert Akpalu, Momodou Cham, Fred Stephen Sarfo, Tiziana Brambilla, Alba Bonetti, Marianna Amboni, Stefano Goldwurm, and Gianni Pezzoli. 2012. "Screening LRRK2 Gene Mutations in Patients with Parkinson's Disease in Ghana." *Journal of Neurology* 259 (3): 569–70.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. 2012. "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain w1118; Iso-2; Iso-3." *Fly* 6 (2): 80–92.

Clark, Lorraine N., Gilberto Levy, Ming-Xin Tang, Helen Mejia-Santana, Alejandra Ciappa, Benjamin Tycko, Lucien J. Cote, Elan D. Louis, Richard Mayeux, and Karen Marder. 2003. "The Saitohin 'Q7R' Polymorphism and Tau Haplotype in Multi-Ethnic Alzheimer Disease and Parkinson's Disease Cohorts." *Neuroscience Letters*. https://doi.org/10.1016/s0304-3940(03)00635-9.

Cookson, Mark R. 2012. "Parkinsonism due to Mutations in PINK1, Parkin, and DJ-1 and Oxidative Stress and Mitochondrial Pathways." *Cold Spring Harbor Perspectives in Medicine* 2 (9): a009415.

Daida, Kensuke, Manabu Funayama, Kimberley J. Billingsley, Laksh Malik, Abigail Miano-Burkhardt, Hampton L. Leonard, Mary B. Makarious, et al. 2023. "Long-Read Sequencing Resolves a Complex Structural Variant in PRKN Parkinson's Disease." *Movement Disorders: Official Journal of the Movement Disorder Society*, November. https://doi.org/10.1002/mds.29610.

Dardiotis, Efthimios, Vasileios Siokas, Eva Pantazi, Maria Dardioti, Dimitrios Rikos, Georgia Xiromerisiou, Aikaterini Markou, Dimitra Papadimitriou, Matthaios Speletas, and Georgios M. Hadjigeorgiou. 2017. "A Novel Mutation in TREM2 Gene Causing Nasu-Hakola Disease and Review of the Literature." *Neurobiology of Aging* 53 (May): 194.e13–194.e22.

Das, Sayantan, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E. Locke, Alan Kwong, Scott I. Vrieze, et al. 2016. "Next-Generation Genotype Imputation Service and Methods." *Nature Genetics* 48 (10): 1284–87.

Delaneau, Olivier, Jean-François Zagury, Matthew R. Robinson, Jonathan L. Marchini, and Emmanouil T. Dermitzakis. 2019. "Accurate, Scalable and Integrative Haplotype Estimation." *Nature Communications* 10 (1): 5436.

DePristo, Mark A., Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, et al. 2011. "A Framework for Variation Discovery and Genotyping Using next-Generation DNA Sequencing Data." *Nature Genetics* 43 (5): 491–98.

Derkach, Andriy, Jerry F. Lawless, and Lei Sun. 2013. "Robust and Powerful Tests for Rare Variants Using

Fisher's Method to Combine Evidence of Association from Two or More Complementary Tests." *Genetic Epidemiology* 37 (1): 110–21.
Di Fonzo, Alessio, Christan F. Rohé, Joaquim Ferreira, Hsin F. Chien, Laura Vacca, Fabrizio Stocchi, Leonor Guedes, et al. 2005. "A Frequent LRRK2 Gene Mutation Associated with Autosomal Dominant Parkinson's Disease." *The Lancet* 365 (9457): 412–15.
Dornbierer, D. A., M. Boxler, C. D. Voegel, B. Stucky, A. E. Steuer, T. M. Binz, M. R. Baumgartner, et al. 2019. "Nocturnal Gamma-Hydroxybutyrate Reduces Cortisol-Awakening Response and Morning Kynurenine Pathway Metabolites in Healthy Volunteers." *The International Journal of Neuropsychopharmacology / Official Scientific Journal of the Collegium Internationale Neuropsychopharmacologicum* 22 (10): 631–39.
Dorsey, E. Ray, Todd Sherer, Michael S. Okun, and Bastiaan R. Bloem. 2018. "The Emerging Evidence of the Parkinson Pandemic." *Journal of Parkinson's Disease* 8 (s1): S3–8.
Doty, R. L., P. Shaman, C. P. Kimmelman, and M. S. Dann. 1984. "University of Pennsylvania Smell Identification Test: A Rapid Quantitative Olfactory Function Test for the Clinic." *The Laryngoscope* 94 (2 Pt 1). https://doi.org/10.1288/00005537-198402000-00004.
Duncan, L., H. Shen, B. Gelaye, J. Meijsen, K. Ressler, M. Feldman, R. Peterson, and B. Domingue. 2019. "Analysis of Polygenic Risk Score Usage and Performance in Diverse Human Populations." *Nature Communications* 10 (1): 3328.
Elsayed, Inas, Alejandro Martinez-Carrasco, Mario Cornejo-Olivas, and Sara Bandres-Ciga. 2021. "Mapping the Diverse and Inclusive Future of Parkinson's Disease Genetics and Its Widespread Impact." *Genes* 12 (11). https://doi.org/10.3390/genes12111681.
El-Tallawy, Hamdy N., Wafaa M. Farghaly, Ghaydaa A. Shehata, Tarek A. Rageh, Nabil M. Abdel Hakeem, Mohamed Abd Al Hamed, and Reda Badry. 2013. "Prevalence of Parkinson's Disease and Other Types of Parkinsonism in Al Kharga District, Egypt." *Neuropsychiatric Disease and Treatment* 9 (November): 1821–26.
Eraslan, Gökcen, Žiga Avsec, Julien Gagneur, and Fabian J. Theis. 2019. "Deep Learning: New Computational Modelling Techniques for Genomics." *Nature Reviews. Genetics* 20 (7): 389–403.
Erikson, Galina A., Dale L. Bodian, Manuel Rueda, Bhuvan Molparia, Erick R. Scott, Ashley A. Scott-Van Zeeland, Sarah E. Topol, et al. 2016. "Whole-Genome Sequencing of a Healthy Aging Cohort." *Cell* 165 (4): 1002–11.
Eschbach, Judith, and Karin M. Danzer. 2014. "α-Synuclein in Parkinson's Disease: Pathogenic Function and Translation into Animal Models." *Neuro-Degenerative Diseases* 14 (1): 1–17.
"Extremely Randomized Trees." 2010. *Automation and Control Engineering*. https://doi.org/10.1201/9781439821091-a1.
Fan, Yu, Cheng-Yuan Mao, Ya-Li Dong, Si Shen, Qi-Meng Zhang, Da-Bao Yao, Fen Liu, et al. 2020. "ARSA Gene Variants and Parkinson's Disease." *Brain: A Journal of Neurology*.
Fayyad, Muneera, Safa Salim, Nour Majbour, Daniel Erskine, Erik Stoops, Brit Mollenhauer, and Omar M. A. El-Agnaf. 2019. "Parkinson's Disease Biomarkers Based on α-Synuclein." *Journal of Neurochemistry* 150 (5): 626–36.
Feng, Chien-Wei, Nan-Fu Chen, Chun-Sung Sung, Hsiao-Mei Kuo, San-Nan Yang, Chien-Liang Chen, Han-Chun Hung, Bing-Hung Chen, Zhi-Hong Wen, and Wu-Fu Chen. 2019. "Therapeutic Effect of Modulating TREM-1 via Anti-Inflammation and Autophagy in Parkinson's Disease." *Frontiers in Neuroscience* 13 (August): 769.
Fields, Carroll Rutherford, Nora Bengoa-Vergniory, and Richard Wade-Martins. 2019. "Targeting Alpha-Synuclein as a Therapy for Parkinson's Disease." *Frontiers in Molecular Neuroscience* 12 (December): 299.
Foo, Jia Nee, Elaine Guo Yan Chew, Sun Ju Chung, Rong Peng, Cornelis Blauwendraat, Mike A. Nalls, Kin Y. Mok, et al. 2020a. "Identification of Risk Loci for Parkinson Disease in Asians and Comparison of Risk

Between Asians and Europeans: A Genome-Wide Association Study." *JAMA Neurology* 77 (6): 746–54.

———. 2020b. "Identification of Risk Loci for Parkinson Disease in Asians and Comparison of Risk Between Asians and Europeans: A Genome-Wide Association Study." *JAMA Neurology* 77 (6): 746–54.

Fuchsberger, Christian, Gonçalo R. Abecasis, and David A. Hinds. 2015. "minimac2: Faster Genotype Imputation." *Bioinformatics* 31 (5): 782–84.

Funayama, Manabu, Kazuko Hasegawa, Etsuro Ohta, Noriko Kawashima, Masaru Komiyama, Hisayuki Kowa, Shoji Tsuji, and Fumiya Obata. 2005. "An LRRK2 Mutation as a Cause for the Parkinsonism in the Original PARK8 Family." *Annals of Neurology* 57 (6): 918–21.

Gaare, Johannes Jernqvist, Gonzalo Nido, Christian Dölle, Paweł Sztromwasser, Guido Alves, Ole-Bjørn Tysnes, Kristoffer Haugarvoll, and Charalampos Tzoulis. 2020. "Meta-Analysis of Whole-Exome Sequencing Data from Two Independent Cohorts Finds No Evidence for Rare Variant Enrichment in Parkinson Disease Associated Loci." *PloS One* 15 (10): e0239824.

Gan-Or, Z., N. Giladi, U. Rozovski, C. Shifrin, S. Rosner, T. Gurevich, A. Bar-Shira, and A. Orr-Urtreger. 2008. "Genotype-Phenotype Correlations between GBA Mutations and Parkinson Disease Risk and Onset." *Neurology* 70 (24): 2277–83.

García-Martín, Elena, Mónica Diez-Fairen, Pau Pastor, Javier Gómez-Tabales, Hortensia Alonso-Navarro, Ignacio Alvarez, María Cárcel, Miquel Aguilar, José A. G. Agúndez, and Félix Javier Jiménez-Jiménez. 2019. "Association between the Missense Alcohol Dehydrogenase rs1229984T Variant with the Risk for Parkinson's Disease in Women." *Journal of Neurology* 266 (2): 346–52.

GBD 2016 Parkinson's Disease Collaborators. 2018. "Global, Regional, and National Burden of Parkinson's Disease, 1990-2016: A Systematic Analysis for the Global Burden of Disease Study 2016." *Lancet Neurology* 17 (11): 939–53.

Gelber, Rebecca P., Lenore J. Launer, and Lon R. White. 2012. "The Honolulu-Asia Aging Study: Epidemiologic and Neuropathologic Research on Cognitive Impairment." *Current Alzheimer Research* 9 (6): 664–72.

"Genome-Wide Association Study of White Matter Hyperintensity Volume in Elderly Persons without Dementia." 2020. *NeuroImage: Clinical* 26 (January): 102209.

Gibbs, J. Raphael, Marcel P. van der Brug, Dena G. Hernandez, Bryan J. Traynor, Michael A. Nalls, Shiao-Lin Lai, Sampath Arepalli, et al. 2010. "Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain." *PLoS Genetics* 6 (5): e1000952.

Giladi, Nir, Roy N. Alcalay, Gary Cutter, Thomas Gasser, Tanya Gurevich, Günter U. Höglinger, Kenneth Marek, et al. 2023. "Safety and Efficacy of Venglustat in GBA1-Associated Parkinson's Disease: An International, Multicentre, Double-Blind, Randomised, Placebo-Controlled, Phase 2 Trial." *Lancet Neurology* 22 (8): 661–71.

Gillis, Jesse, Sara Ballouz, and Paul Pavlidis. 2014. "Bias Tradeoffs in the Creation and Analysis of Protein-Protein Interaction Networks." *Journal of Proteomics* 100 (April): 44–54.

Global Parkinson's Genetics Program. 2021. "GP2: The Global Parkinson's Genetics Program." *Movement Disorders: Official Journal of the Movement Disorder Society* 36 (4): 842–51.

Gloeckner, Christian Johannes, Norbert Kinkl, Annette Schumacher, Ralf J. Braun, Eric O'Neill, Thomas Meitinger, Walter Kolch, Holger Prokisch, and Marius Ueffing. 2006. "The Parkinson Disease Causing LRRK2 Mutation I2020T Is Associated with Increased Kinase Activity." *Human Molecular Genetics* 15 (2): 223–32.

Goedert, Michel, Maria Grazia Spillantini, Kelly Del Tredici, and Heiko Braak. 2013. "100 Years of Lewy Pathology." *Nature Reviews. Neurology* 9 (1): 13–24.

Green, Eric D., Chris Gunter, Leslie G. Biesecker, Valentina Di Francesco, Carla L. Easter, Elise A. Feingold, Adam L. Felsenfeld, et al. 2020. "Strategic Vision for Improving Human Health at The Forefront of

Genomics." *Nature* 586 (7831): 683–92.
Gwinn-Hardy, K., A. Singleton, P. O'Suilleabhain, M. Boss, D. Nicholl, A. Adam, J. Hussey, P. Critchley, J. Hardy, and M. Farrer. 2001. "Spinocerebellar Ataxia Type 3 Phenotypically Resembling Parkinson Disease in a Black Family." *Archives of Neurology* 58 (2). https://doi.org/10.1001/archneur.58.2.296.
Halestrap, Andrew P. 2013. "The SLC16 Gene Family - Structure, Role and Regulation in Health and Disease." *Molecular Aspects of Medicine* 34 (2-3): 337–49.
Halman, Andreas, Egor Dolzhenko, and Alicia Oshlack. 2022. "STRipy: A Graphical Application for Enhanced Genotyping of Pathogenic Short Tandem Repeats in Sequencing Data." *Human Mutation* 43 (7): 859–68.
Halsband, Claire, Antonia Zapf, Friederike Sixel-Döring, Claudia Trenkwalder, and Brit Mollenhauer. 2018. "The REM Sleep Behavior Disorder Screening Questionnaire Is Not Valid in De Novo Parkinson's Disease." *Movement Disorders Clinical Practice* 5 (2): 171–76.
Haynes, Winston A., Aurelie Tomczak, and Purvesh Khatri. 2018. "Gene Annotation Bias Impedes Biomedical Research." *Scientific Reports* 8 (1): 1362.
Heilbron, Karl, Alastair J. Noyce, Pierre Fontanillas, Babak Alipanahi, Mike A. Nalls, 23andMe Research Team, and Paul Cannon. 2019. "The Parkinson's Phenome-Traits Associated with Parkinson's Disease in a Broadly Phenotyped Cohort." *NPJ Parkinson's Disease* 5 (March): 4.
Heinzel, Sebastian, Daniela Berg, Thomas Gasser, Honglei Chen, Chun Yao, Ronald B. Postuma, and MDS Task Force on the Definition of Parkinson's Disease. 2019. "Update of the MDS Research Criteria for Prodromal Parkinson's Disease." *Movement Disorders: Official Journal of the Movement Disorder Society* 34 (10): 1464–70.
He, Xianghua, Lifang Zhang, Xiaoli Yao, Jing Hu, Lihua Yu, Hua Jia, Ran An, Zhuolin Liu, and Yanming Xu. 2013. "Association Studies of MMP-9 in Parkinson's Disease and Amyotrophic Lateral Sclerosis." *PloS One* 8 (9): e73777.
Hill-Burns, Erin M., Owen A. Ross, William T. Wissemann, Alexandra I. Soto-Ortolaza, Sepideh Zareparsi, Joanna Siuda, Timothy Lynch, et al. 2016. "Identification of Genetic Modifiers of Age-at-Onset for Familial Parkinson's Disease." *Human Molecular Genetics* 25 (17): 3849–62.
Hughes, A. J., S. E. Daniel, L. Kilford, and A. J. Lees. 1992. "Accuracy of Clinical Diagnosis of Idiopathic Parkinson's Disease: A Clinico-Pathological Study of 100 Cases." *Journal of Neurology, Neurosurgery, and Psychiatry* 55 (3): 181–84.
Hutchins, Elizabeth, David Craig, Ivo Violich, Eric Alsop, Bradford Casey, Samantha Hutten, Alyssa Reimer, et al. 2021. "Quality Control Metrics for Whole Blood Transcriptome Analysis in the Parkinson's Progression Markers Initiative (PPMI)." *medRxiv*, January, 2021.01.05.21249278.
Ibáñez, Kristina, César Boullosa, Rafael Tabarés-Seisdedos, Anaïs Baudot, and Alfonso Valencia. 2014. "Molecular Evidence for the Inverse Comorbidity between Central Nervous System Disorders and Cancers Detected by Transcriptomic Meta-Analyses." *PLoS Genetics* 10 (2). https://doi.org/10.1371/journal.pgen.1004173.
"Identification of 969 Protein Quantitative Trait Loci in an African American Population with Kidney Disease Attributed to Hypertension." 2022. *Kidney International* 102 (5): 1167–77.
Inzelberg, Rivka, Sharon Hassin-Baer, and Joseph Jankovic. 2014. "Genetic Movement Disorders in Patients of Jewish Ancestry." *JAMA Neurology* 71 (12): 1567–72.
Ishikawa, A., and S. Tsuji. 1996. "Clinical Analysis of 17 Patients in 12 Japanese Families with Autosomal-Recessive Type Juvenile Parkinsonism." *Neurology* 47 (1): 160–66.
Iwaki, Hirotaka, Cornelis Blauwendraat, Mary B. Makarious, Sara Bandrés-Ciga, Hampton L. Leonard, J. Raphael Gibbs, Dena G. Hernandez, et al. 2020. "Penetrance of Parkinson's Disease in LRRK2 p.G2019S Carriers Is Modified by a Polygenic Risk Score." *Movement Disorders: Official Journal of the Movement Disorder Society* 35 (5): 774–80.
Iwaki, Hirotaka, Hampton L. Leonard, Mary B. Makarious, Matt Bookman, Barry Landin, David Vismer,

Bradford Casey, et al. 2021. “Accelerating Medicines Partnership: Parkinson’s Disease. Genetic Resource.” *Movement Disorders: Official Journal of the Movement Disorder Society* 36 (8): 1795–1804.

Jain, Chirag, Arang Rhie, Nancy F. Hansen, Sergey Koren, and Adam M. Phillippy. 2022. “Long-Read Mapping to Repetitive Reference Sequences Using Winnowmap2.” *Nature Methods* 19 (6): 705–10.

Jansen, Iris E., J. Raphael Gibbs, Mike A. Nalls, T. Ryan Price, Steven Lubbe, Jeroen van Rooij, André G. Uitterlinden, et al. 2017. “Establishing the Role of Rare Coding Variants in Known Parkinson’s Disease Risk Loci.” *Neurobiology of Aging* 59 (November): 220.e11–220.e18.

Jia, Fangzhi, Avi Fellner, and Kishore Raj Kumar. 2022. “Monogenic Parkinson’s Disease: Genotype, Phenotype, Pathophysiology, and Genetic Testing.” *Genes* 13 (3). https://doi.org/10.3390/genes13030471.

Johansen, Krisztina Kunszt, Sverre Helge Torp, Matthew J. Farrer, Emil K. Gustavsson, and Jan O. Aasly. 2018. “A Case of Parkinson’s Disease with No Lewy Body Pathology due to a Homozygous Exon Deletion in.” *Case Reports in Neurological Medicine* 2018 (June): 6838965.

Jourquin, J., D. Duncan, Z. Shi, and B. Zhang. 2012. “GLAD4U: Deriving and Prioritizing Gene Lists from PubMed Literature.” *BMC Genomics* 13 Suppl 8 (Suppl 8). https://doi.org/10.1186/1471-2164-13-S8-S20.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89.

Kachuri, Linda, Angel C. Y. Mak, Donglei Hu, Celeste Eng, Scott Huntsman, Jennifer R. Elhawary, Namrata Gupta, et al. 2023. “Gene Expression in African Americans, Puerto Ricans and Mexican Americans Reveals Ancestry-Specific Patterns of Genetic Architecture.” *Nature Genetics* 55 (6): 952–63.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. “The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans.” *Nature* 581 (7809): 434–43.

Kasten, Meike, Corinna Hartmann, Jennie Hampf, Susen Schaake, Ana Westenberger, Eva-Juliane Vollstedt, Alexander Balck, et al. 2018. “Genotype-Phenotype Relations for the Parkinson’s Disease Genes Parkin, PINK1, DJ1: MDSGene Systematic Review.” *Movement Disorders: Official Journal of the Movement Disorder Society* 33 (5): 730–41.

Kasten, Meike, and Christine Klein. 2013. “The Many Faces of Alpha-Synuclein Mutations.” *Movement Disorders: Official Journal of the Movement Disorder Society* 28 (6): 697–701.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree.” In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 3146–54. Curran Associates, Inc.

Khedr, Eman M., Gharib Fawi, Mohammed Abd Allah Abbas, Talal A. Mohammed, Noha Abo El-Fetoh, Ghada Al Attar, and Ahmed F. Zaki. 2015. “Prevalence of Parkinsonism and Parkinson’s Disease in Qena governorate/Egypt: A Cross-Sectional Community-Based Survey.” *Neurological Research* 37 (7): 607–18.

Kim, Jonggeol Jeffrey, Sara Bandres-Ciga, Cornelis Blauwendraat, International Parkinson’s Disease Genomics Consortium, and Ziv Gan-Or. 2020. “No Genetic Evidence for Involvement of Alcohol Dehydrogenase Genes in Risk for Parkinson’s Disease.” *Neurobiology of Aging* 87 (March): 140.e19–140.e22.

Kim, Jonggeol Jeffrey, Dan Vitale, Diego Véliz Otani, Michelle Mulan Lian, Karl Heilbron, 23andMe Research Team, Hirotaka Iwaki, et al. 2024. “Multi-Ancestry Genome-Wide Association Meta-Analysis of Parkinson’s Disease.” *Nature Genetics* 56 (1): 27–36.

Kim, Jonggeol Jeffrey, Dan Vitale, Diego Véliz Otani, Michelle Lian, Karl Heilbron, Hirotaka Iwaki, Julie

Lake, et al. 2022. “Multi-Ancestry Genome-Wide Meta-Analysis in Parkinson’s Disease.” *bioRxiv*. https://doi.org/10.1101/2022.08.04.22278432.

Kim, Min Woo, Kyonghwan Choe, Jun Sung Park, Hyeon Jin Lee, Min Hwa Kang, Riaz Ahmad, and Myeong Ok Kim. 2022. “Pharmacological Inhibition of Spleen Tyrosine Kinase Suppressed Neuroinflammation and Cognitive Dysfunction in LPS-Induced Neurodegeneration Model.” *Cells* 11 (11). https://doi.org/10.3390/cells11111777.

King, Emily A., J. Wade Davis, and Jacob F. Degner. 2019. “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 Genetics* 15 (12): e1008489.

Kitada, T., S. Asakawa, N. Hattori, H. Matsumine, Y. Yamamura, S. Minoshima, M. Yokochi, Y. Mizuno, and N. Shimizu. 1998. “Mutations in the Parkin Gene Cause Autosomal Recessive Juvenile Parkinsonism.” *Nature* 392 (6676): 605–8.

Klein, Christine, and Ana Westenberger. 2012. “Genetics of Parkinson’s Disease.” *Cold Spring Harbor Perspectives in Medicine* 2 (1): a008888.

Koch, Sebastian, Björn-Hergen Laabs, Meike Kasten, Eva-Juliane Vollstedt, Jos Becktepe, Norbert Brüggemann, Andre Franke, et al. 2021. “Validity and Prognostic Value of a Polygenic Risk Score for Parkinson’s Disease.” *Genes* 12 (12). https://doi.org/10.3390/genes12121859.

Kolmogorov, Mikhail, Kimberley J. Billingsley, Mira Mastoras, Melissa Meredith, Jean Monlong, Ryan Lorig-Roach, Mobin Asri, et al. 2023. “Scalable Nanopore Sequencing of Human Genomes Provides a Comprehensive View of Haplotype-Resolved Variation and Methylation.” *bioRxiv : The Preprint Server for Biology*, January. https://doi.org/10.1101/2023.01.12.523790.

Konno, Takuya, Owen A. Ross, Andreas Puschmann, Dennis W. Dickson, and Zbigniew K. Wszolek. 2016. “Autosomal Dominant Parkinson’s Disease Caused by SNCA Duplications.” *Parkinsonism & Related Disorders* 22 Suppl 1 (January): S1–6.

Koretsky, Mathew J., Chelsea Alvarado, Mary B. Makarious, Dan Vitale, Kristin Levine, Anant Dadu, Sonja W. Scholz, et al. 2022. “Genetic Risk Factor Clustering within and across Neurodegenerative Diseases.” *bioRxiv*. https://doi.org/10.1101/2022.12.01.22282945.

Lake, Julie, Catherine S. Storm, Mary B. Makarious, and Sara Bandres-Ciga. 2021. “Genetic and Transcriptomic Biomarkers in Neurodegenerative Diseases: Current Situation and the Road Ahead.” *Cells* 10 (5). https://doi.org/10.3390/cells10051030.

Lake, Julie, Caroline Warly Solsberg, Jonggeol Jeffrey Kim, Juliana Acosta-Uribe, Mary B. Makarious, Zizheng Li, Kristin Levine, et al. 2023. “Multi-Ancestry Meta-Analysis and Fine-Mapping in Alzheimer’s Disease.” *Molecular Psychiatry* 28 (7): 3121–32.

Lee, Dong Ah, Ho-Joon Lee, Hyung Chan Kim, and Kang Min Park. 2021. “Application of Machine Learning Analysis Based on Diffusion Tensor Imaging to Identify REM Sleep Behavior Disorder.” *Sleep & Breathing = Schlaf & Atmung*, July. https://doi.org/10.1007/s11325-021-02434-9.

Lee, Jun Sung, Kazuaki Kanai, Mari Suzuki, Woojin S. Kim, Han Soo Yoo, Yuhong Fu, Dong-Kyu Kim, et al. 2019. “Arylsulfatase A, a Genetic Modifier of Parkinson’s Disease, Is an α-Synuclein Chaperone.” *Brain: A Journal of Neurology* 142 (9): 2845–59.

Lees, Andrew. 2017. “An Essay on the Shaking Palsy.” *Brain: A Journal of Neurology* 140 (3): 843–48.

Lee, Seunggeun, Mary J. Emond, Michael J. Bamshad, Kathleen C. Barnes, Mark J. Rieder, Deborah A. Nickerson, David C. Christiani, Mark M. Wurfel, and Xihong Lin. 2012. “Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies.” *The American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2012.06.007.

Lee, Seunggeun, Mary J. Emond, Michael J. Bamshad, Kathleen C. Barnes, Mark J. Rieder, Deborah A. Nickerson, NHLBI GO Exome Sequencing Project—ESP Lung Project Team, David C. Christiani, Mark M. Wurfel, and Xihong Lin. 2012. “Optimal Unified Approach for Rare-Variant Association Testing

with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies." *American Journal of Human Genetics* 91 (2): 224–37.
Lee, Seunggeun, Christian Fuchsberger, Sehee Kim, and Laura Scott. 2016. "An Efficient Resampling Method for Calibrating Single and Gene-Based Rare Variant Association Analysis in Case-Control Studies." *Biostatistics* 17 (1): 1–15.
Lee, Seunggeung, Gonçalo R. Abecasis, Michael Boehnke, and Xihong Lin. 2014. "Rare-Variant Association Analysis: Study Designs and Statistical Tests." *American Journal of Human Genetics* 95 (1): 5–23.
Lee, Seunggeun, Michael C. Wu, and Xihong Lin. 2012. "Optimal Tests for Rare Variant Effects in Sequencing Association Studies." *Biostatistics* 13 (4): 762–75.
Lehri-Boufala, Sonia, Mohand-Ouidir Ouidja, Véronique Barbier-Chassefière, Emilie Hénault, Rita Raisman-Vozari, Laure Garrigue-Antar, Dulce Papy-Garcia, and Christophe Morin. 2015. "New Roles of Glycosaminoglycans in α-Synuclein Aggregation in a Cellular Model of Parkinson Disease." *PloS One* 10 (1): e0116641.
Leonard, Hampton, Cornelis Blauwendraat, Lynne Krohn, Faraz Faghri, Hirotaka Iwaki, Glen Ferguson, Aaron G. Day-Williams, et al. 2020. "Genetic Variability and Potential Effects on Clinical Trial Outcomes: Perspectives in Parkinson's Disease." *Journal of Medical Genetics* 57 (5): 331–38.
Leonard, Hampton, Caroline Jonson, Kristin Levine, Julie Lake, Linnea Hertslet, Lietsel Jones, Dhairya Patel, et al. 2024. "Assessing the Lack of Diversity in Genetics Research across Neurodegenerative Diseases: A Systematic Review of the GWAS Catalog and Literature." *medRxiv*. https://doi.org/10.1101/2024.01.08.24301007.
Lesage, Suzanne, Valérie Drouet, Elisa Majounie, Vincent Deramecourt, Maxime Jacoupy, Aude Nicolas, Florence Cormier-Dequaire, et al. 2016. "Loss of VPS13C Function in Autosomal-Recessive Parkinsonism Causes Mitochondrial Dysfunction and Increases PINK1/Parkin-Dependent Mitophagy." *American Journal of Human Genetics* 98 (3): 500–513.
Liao, Jingling, Chun-Xiang Wu, Christopher Burlak, Sheng Zhang, Heather Sahm, Mu Wang, Zhong-Yin Zhang, et al. 2014. "Parkinson Disease-Associated Mutation R1441H in LRRK2 Prolongs the 'Active State' of Its GTPase Domain." *Proceedings of the National Academy of Sciences of the United States of America* 111 (11): 4055–60.
Liao, Yuxing, Jing Wang, Eric J. Jaehnig, Zhiao Shi, and Bing Zhang. 2019. "WebGestalt 2019: Gene Set Analysis Toolkit with Revamped UIs and APIs." *Nucleic Acids Research* 47 (W1): W199–205.
Libbrecht, Maxwell W., and William Stafford Noble. 2015. "Machine Learning Applications in Genetics and Genomics." *Nature Reviews. Genetics* 16 (6): 321–32.
Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." *Bioinformatics* 25 (14): 1754–60.
Liu, Yaowu, Sixing Chen, Zilin Li, Alanna C. Morrison, Eric Boerwinkle, and Xihong Lin. 2019. "ACAT: A Fast and Powerful P Value Combination Method for Rare-Variant Analysis in Sequencing Studies." *American Journal of Human Genetics* 104 (3): 410–21.
Li, Wen, Yuhong Fu, Glenda M. Halliday, and Carolyn M. Sue. 2021. "Genes Link Mitochondrial Dysfunction and Alpha-Synuclein Pathology in Sporadic Parkinson's Disease." *Frontiers in Cell and Developmental Biology* 9 (July): 612476.
Loesch, Douglas P., Andrea R. V. R. Horimoto, Karl Heilbron, Elif I. Sarihan, Miguel Inca-Martinez, Emily Mason, Mario Cornejo-Olivas, et al. 2021. "Characterizing the Genetic Architecture of Parkinson's Disease in Latinos." *Annals of Neurology* 90 (3): 353–65.
Lopez, Kevin, Samah J. Fodeh, Ahmed Allam, Cynthia A. Brandt, and Michael Krauthammer. 2020. "Reducing Annotation Burden Through Multimodal Learning." *Frontiers in Big Data* 3 (June): 19.
Lundberg, Scott M., Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. "From Local Explanations to Global

Understanding with Explainable AI for Trees." *Nature Machine Intelligence* 2 (1): 56–67.
Lundberg, Scott M., Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, et al. 2018. "Explainable Machine-Learning Predictions for the Prevention of Hypoxaemia during Surgery." *Nature Biomedical Engineering* 2 (10): 749–60.
Lv, Qiankun, Ziyu Wang, Zhen Zhong, and Wei Huang. 2020. "Role of Long Noncoding RNAs in Parkinson's Disease: Putative Biomarkers and Therapeutic Targets." *Parkinson's Disease* 2020 (June). https://doi.org/10.1155/2020/5374307.
Magalingam, K. B., A. Radhakrishnan, P. Ramdas, and N. Haleagrahara. 2015. "Quercetin Glycosides Induced Neuroprotection by Changes in the Gene Expression in a Cellular Model of Parkinson's Disease." *Journal of Molecular Neuroscience: MN* 55 (3). https://doi.org/10.1007/s12031-014-0400-x.
Mahungu, Amokelani C., David G. Anderson, Anastasia C. Rossouw, Riaan van Coller, Jonathan A. Carr, Owen A. Ross, and Soraya Bardien. 2020. "Screening of the Glucocerebrosidase (GBA) Gene in South Africans of African Ancestry with Parkinson's Disease." *Neurobiology of Aging* 88 (April): 156.e11–156.e14.
Makarious, Mary B., Monica Diez-Fairen, Lynne Krohn, Cornelis Blauwendraat, Sara Bandres-Ciga, Jinhui Ding, Lasse Pihlstrøm, Henry Houlden, Sonja W. Scholz, and Ziv Gan-Or. 2019. "ARSA Variants in α-Synucleinopathies." *Brain: A Journal of Neurology*.
Makarious, Mary B., Hampton L. Leonard, Dan Vitale, Hirotaka Iwaki, David Saffo, Lana Sargent, Anant Dadu, et al. 2021. "GenoML: Automated Machine Learning for Genomics." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2103.03221.
Martin, Alicia R., Christopher R. Gignoux, Raymond K. Walters, Genevieve L. Wojcik, Benjamin M. Neale, Simon Gravel, Mark J. Daly, Carlos D. Bustamante, and Eimear E. Kenny. 2020. "Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations." *American Journal of Human Genetics* 107 (4): 788–89.
Martin, Alicia R., Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. n.d. "Current Clinical Use of Polygenic Scores Will Risk Exacerbating Health Disparities." https://doi.org/10.1101/441261.
Martin, Shaun, Stefanie Smolders, Chris Van den Haute, Bavo Heeman, Sarah van Veen, David Crosiers, Igor Beletchi, et al. 2020. "Mutated ATP10B Increases Parkinson's Disease Risk by Compromising Lysosomal Glucosylceramide Export." *Acta Neuropathologica* 139 (6): 1001–24.
Mata, Ignacio F., Min Shi, Pinky Agarwal, Kathryn A. Chung, Karen L. Edwards, Stewart A. Factor, Douglas R. Galasko, et al. 2010. "SNCA Variant Associated with Parkinson Disease and Plasma Alpha-Synuclein Level." *Archives of Neurology* 67 (11): 1350–56.
McGuire, V., S. K. Van Den Eeden, C. M. Tanner, F. Kamel, D. M. Umbach, K. Marder, R. Mayeux, et al. 2011. "Association of DRD2 and DRD3 Polymorphisms with Parkinson's Disease in a Multiethnic Consortium." *Journal of the Neurological Sciences* 307 (1-2): 22.
McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." *Genome Research* 20 (9): 1297–1303.
McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. "The Ensembl Variant Effect Predictor." *Genome Biology* 17 (1): 122.
Mei, Jie, Christian Desrosiers, and Johannes Frasnelli. 2021. "Machine Learning for the Diagnosis of Parkinson's Disease: A Review of Literature." *Frontiers in Aging Neuroscience* 13 (May): 633752.
Mei, Jie, Shady Rahayel, Christian Desrosiers, Ronald B. Postuma, Jacques Montplaisir, Julie Carrier, Oury Monchi, Johannes Frasnelli, and Jean-François Gagnon. 2021. "Identification of REM Sleep Behavior Disorder by Structural Magnetic Resonance Imaging and Machine Learning." *bioRxiv*.

https://doi.org/10.1101/2021.09.18.21263779.
Mencacci, Niccolò E., Ioannis U. Isaias, Martin M. Reich, Christos Ganos, Vincent Plagnol, James M. Polke, Jose Bras, et al. 2014. “Parkinson’s Disease in GTP Cyclohydrolase 1 Mutation Carriers.” *Brain: A Journal of Neurology* 137 (Pt 9): 2480–92.
Migdalska-Richards, Anna, and Anthony H. V. Schapira. 2016. “The Relationship between Glucocerebrosidase Mutations and Parkinson Disease.” *Journal of Neurochemistry* 139 Suppl 1 (Suppl Suppl 1): 77–90.
Min, Seonwoo, Byunghan Lee, and Sungroh Yoon. 2017. “Deep Learning in Bioinformatics.” *Briefings in Bioinformatics* 18 (5): 851–69.
Mitt, Mario, Mart Kals, Kalle Pärn, Stacey B. Gabriel, Eric S. Lander, Aarno Palotie, Samuli Ripatti, et al. 2017. “Improved Imputation Accuracy of Rare and Low-Frequency Variants Using Population-Specific High-Coverage WGS-Based Imputation Reference Panel.” *European Journal of Human Genetics: EJHG* 25 (7): 869–76.
Mokretar, Katya, Daniel Pease, Jan-Willem Taanman, Aynur Soenmez, Ayesha Ejaz, Tammaryn Lashley, Helen Ling, et al. 2018. “Somatic Copy Number Gains of α-Synuclein (SNCA) in Parkinson’s Disease and Multiple System Atrophy Brains.” *Brain: A Journal of Neurology* 141 (8): 2419–31.
Moon, Intae, Jaclyn LoPiccolo, Sylvan C. Baca, Lynette M. Sholl, Kenneth L. Kehl, Michael J. Hassett, David Liu, Deborah Schrag, and Alexander Gusev. 2023. “Publisher Correction: Machine Learning for Genetics-Based Classification and Treatment Response Prediction in Cancer of Unknown Primary.” *Nature Medicine*, November. https://doi.org/10.1038/s41591-023-02693-x.
Morley, James F., Abigail Cohen, Laura Silveira-Moriyama, Andrew J. Lees, David R. Williams, Regina Katzenschlager, Christopher Hawkes, et al. 2018. “Optimizing Olfactory Testing for the Diagnosis of Parkinson’s Disease: Item Analysis of the University of Pennsylvania Smell Identification Test.” *Npj Parkinson’s Disease* 4 (1): 1–7.
Morris, Huw R., Maria Grazia Spillantini, Carolyn M. Sue, and Caroline H. Williams-Gray. 2024. “The Pathogenesis of Parkinson’s Disease.” *The Lancet* 403 (10423): 293–304.
Mosley, R. Lee, Jessica A. Hutter-Saunders, David K. Stone, and Howard E. Gendelman. 2012. “Inflammation and Adaptive Immunity in Parkinson’s Disease.” *Cold Spring Harbor Perspectives in Medicine* 2 (1): a009381.
Mu, Jesse, Kallol R. Chaudhuri, Concha Bielza, Jesus de Pedro-Cuesta, Pedro Larrañaga, and Pablo Martinez-Martin. 2017. “Parkinson’s Disease Subtypes Identified from Cluster Analysis of Motor and Non-Motor Symptoms.” *Frontiers in Aging Neuroscience* 9 (September): 301.
Nalls, Mike A., Cornelis Blauwendraat, Costanza L. Vallerga, Karl Heilbron, Sara Bandres-Ciga, Diana Chang, Manuela Tan, et al. 2019a. “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.
———. 2019b. “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.
Nalls, Mike A., Cory Y. McLean, Jacqueline Rick, Shirley Eberly, Samantha J. Hutten, Katrina Gwinn, Margaret Sutherland, et al. 2015. “Diagnosis of Parkinson’s Disease on the Basis of Clinical and Genetic Classification: A Population-Based Modelling Study.” *Lancet Neurology* 14 (10): 1002–9.
Nelson, Matthew R., Hannah Tipney, Jeffery L. Painter, Judong Shen, Paola Nicoletti, Yufeng Shen, Aris Floratos, et al. 2015. “The Support of Human Genetic Evidence for Approved Drug Indications.” *Nature Genetics* 47 (8): 856–60.
Ngo, Dai-Hung, and Thanh Sang Vo. 2019. “An Updated Review on Pharmaceutical Properties of Gamma-Aminobutyric Acid.” *Molecules* 24 (15). https://doi.org/10.3390/molecules24152678.
Noyce, Alastair J., Jonathan P. Bestwick, Laura Silveira-Moriyama, Christopher H. Hawkes, Charles H.

Knowles, John Hardy, Gavin Giovannoni, et al. 2014. “PREDICT-PD: Identifying Risk of Parkinson’s Disease in the Community: Methods and Baseline Results.” *Journal of Neurology, Neurosurgery, and Psychiatry* 85 (1): 31–37.

Ojo, Oluwadamilola Omolara, Sara Bandres-Ciga, Mary B. Makarious, Peter Wild Crea, Dena G. Hernandez, Henry Houlden, Mie Rizig, et al. 2024. “GBA1 rs3115534 Is Associated with REM Sleep Behavior Disorder in Parkinson’s Disease in Nigerians.” *Movement Disorders: Official Journal of the Movement Disorder Society*, February. https://doi.org/10.1002/mds.29753.

Ojo, Oluwadamilola O., Kolawole W. Wahab, Abiodun H. Bello, Sani A. Abubakar, Bertha C. Ekeh, Folajimi M. Otubogun, Emmanuel U. Iwuozo, et al. 2021. “A Cross-Sectional Comprehensive Assessment of the Profile and Burden of Non-Motor Symptoms in Relation to Motor Phenotype in the Nigeria Parkinson Disease Registry Cohort.” *Movement Disorders Clinical Practice* 8 (8): 1206–15.

Ojo, O. O., S. A. Abubakar, E. U. Iwuozo, E. O. Nwazor, O. S. Ekenze, T. H. Farombi, R. O. Akinyemi, et al. 2020. “The Nigeria Parkinson Disease Registry: Process, Profile, and Prospects of a Collaborative Project.” *Movement Disorders: Official Journal of the Movement Disorder Society* 35 (8). https://doi.org/10.1002/mds.28123.

Okubadejo, Njideka, Angela Britton, Cynthia Crews, Rufus Akinyemi, John Hardy, Andrew Singleton, and Jose Bras. 2008. “Analysis of Nigerians with Apparently Sporadic Parkinson Disease for Mutations in LRRK2, PRKN and ATXN3.” *PloS One* 3 (10): e3421.

Okubadejo, Njideka U., James H. Bower, Walter A. Rocca, and Demetrius M. Maraganore. 2006. “Parkinson’s Disease in Africa: A Systematic Review of Epidemiologic and Genetic Studies.” *Movement Disorders: Official Journal of the Movement Disorder Society* 21 (12): 2150–56.

Okubadejo, Njideka U., Olaitan Okunoye, Oluwadamilola O. Ojo, Babawale Arabambi, Rufus O. Akinyemi, Godwin O. Osaigbovo, Sani A. Abubakar, et al. 2022. “APOE E4 Is Associated with Impaired Self-Declared Cognition but Not Disease Risk or Age of Onset in Nigerians with Parkinson’s Disease.” *Npj Parkinson’s Disease* 8 (1): 1–6.

Okubadejo, Njideka U., Mie Rizig, Oluwadamilola O. Ojo, Hallgeir Jonvik, Olajumoke Oshinaike, Emmeline Brown, and Henry Houlden. 2018. “Leucine Rich Repeat Kinase 2 (LRRK2) GLY2019SER Mutation Is Absent in a Second Cohort of Nigerian Africans with Parkinson Disease.” *PloS One* 13 (12): e0207984.

Oliveira, Luis M. A., Thomas Gasser, Robert Edwards, Markus Zweckstetter, Ronald Melki, Leonidas Stefanis, Hilal A. Lashuel, et al. 2021. “Alpha-Synuclein Research: Defining Strategic Moves in the Battle against Parkinson’s Disease.” *Npj Parkinson’s Disease*. https://doi.org/10.1038/s41531-021-00203-9.

Ozelius, Laurie J., Geetha Senthil, Rachel Saunders-Pullman, Erin Ohmann, Amanda Deligtisch, Michele Tagliati, Ann L. Hunt, et al. 2006. “LRRK2 G2019S as a Cause of Parkinson’s Disease in Ashkenazi Jews.” *The New England Journal of Medicine* 354 (4): 424–25.

Paisán-Ruíz, Coro, Shushant Jain, E. Whitney Evans, William P. Gilks, Javier Simón, Marcel van der Brug, Adolfo López de Munain, et al. 2004. “Cloning of the Gene Containing Mutations That Cause PARK8-Linked Parkinson’s Disease.” *Neuron* 44 (4): 595–600.

Pal, Madhumita, Smita Parija, Ganapati Panda, Kuldeep Dhama, and Ranjan K. Mohapatra. 2022. “Risk Prediction of Cardiovascular Disease Using Machine Learning Classifiers.” *Open Medicine: A Peer-Reviewed, Independent, Open-Access Journal* 17 (1): 1100–1113.

Palmerini, Luca, Laura Rocchi, Sinziana Mazilu, Eran Gazit, Jeffrey M. Hausdorff, and Lorenzo Chiari. 2017. “Identification of Characteristic Motor Patterns Preceding Freezing of Gait in Parkinson’s Disease Using Wearable Sensors.” *Frontiers in Neurology* 8 (August): 394.

Parkinson, James. 1817. *An Essay on the Shaking Palsy*.

Park, J. K., V. Koprivica, D. Q. Andrews, V. Madike, N. Tayebi, D. L. Stone, and E. Sidransky. 2001. “Glucocerebrosidase Mutations among African-American Patients with Type 1 Gaucher Disease.”

American Journal of Medical Genetics 99 (2): 147–51.
Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1912.01703.
Paulsen, Jane S., Martha Nance, Ji-In Kim, Noelle E. Carlozzi, Peter K. Panegyres, Cheryl Erwin, Anita Goh, Elizabeth McCusker, and Janet K. Williams. 2013. "A Review of Quality of Life after Predictive Testing for and Earlier Identification of Neurodegenerative Diseases." *Progress in Neurobiology* 110 (November): 2–28.
Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and Others. 2011. "Scikit-Learn: Machine Learning in Python." *The Journal of Machine Learning Research* 12: 2825–30.
Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research: JMLR* 12 (85): 2825–30.
Pellegrini, Laura, Andrea Wetzel, Simone Grannó, George Heaton, and Kirsten Harvey. 2017. "Back to the Tubule: Microtubule Dynamics in Parkinson's Disease." *Cellular and Molecular Life Sciences: CMLS* 74 (3): 409–34.
Perandones, C., J. C. Giugni, D. S. Calvo, G. B. Raina, L. De Jorge Lopez, V. Volpini, C. P. Zabetian, et al. 2014. "Mosaicism of Alpha-Synuclein Gene Rearrangements: Report of Two Unrelated Cases of Early-Onset Parkinsonism." *Parkinsonism & Related Disorders* 20 (5): 558–61.
Perez-Rodriguez, Diego, Maria Kalyva, Melissa Leija-Salazar, Tammaryn Lashley, Maxime Tarabichi, Viorica Chelban, Steve Gentleman, et al. 2019. "Investigation of Somatic CNVs in Brains of Synucleinopathy Cases Using Targeted SNCA Analysis and Single Cell Sequencing." *Acta Neuropathologica Communications* 7 (1): 219.
Peters, S. P., R. E. Lee, and R. H. Glew. 1975. "A Microassay for Gaucher's Disease." *Clinica Chimica Acta; International Journal of Clinical Chemistry* 60 (3): 391–96.
Piccio, Laura, Cecilia Buonsanti, Marina Cella, Ilaria Tassi, Robert E. Schmidt, Chiara Fenoglio, John Rinker 2nd, et al. 2008. "Identification of Soluble TREM-2 in the Cerebrospinal Fluid and Its Association with Multiple Sclerosis and CNS Inflammation." *Brain: A Journal of Neurology* 131 (Pt 11): 3081–91.
Picillo, M., M. T. Pellecchia, R. Erro, M. Amboni, C. Vitale, A. Iavarone, M. Moccia, R. Allocca, G. Orefice, and P. Barone. 2014. "The Use of University of Pennsylvania Smell Identification Test in the Diagnosis of Parkinson's Disease in Italy." *Neurological Sciences: Official Journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology* 35 (3). https://doi.org/10.1007/s10072-013-1522-6.
Pickrell, Alicia M., and Richard J. Youle. 2015. "The Roles of PINK1, Parkin, and Mitochondrial Fidelity in Parkinson's Disease." *Neuron* 85 (2): 257–73.
Pihlstrøm, Lasse, Cornelis Blauwendraat, Chiara Cappelletti, Victoria Berge-Seidl, Margrete Langmyhr, Sandra Pilar Henriksen, Wilma D. J. van de Berg, et al. 2018. "A Comprehensive Analysis of SNCA-Related Genetic Risk in Sporadic Parkinson Disease." *Annals of Neurology* 84 (1): 117–29.
Pitz, Vanessa, Mary B. Makarious, Sara Bandres-Ciga, Hirotaka Iwaki, 23andMe Research Team, Andrew B. Singleton, Mike Nalls, Karl Heilbron, and Cornelis Blauwendraat. 2024. "Analysis of Rare Parkinson's Disease Variants in Millions of People." *NPJ Parkinson's Disease* 10 (1): 11.
Polymeropoulos, M. H., C. Lavedan, E. Leroy, S. E. Ide, A. Dehejia, A. Dutra, B. Pike, et al. 1997. "Mutation in the Alpha-Synuclein Gene Identified in Families with Parkinson's Disease." *Science* 276 (5321): 2045–47.
Polymeropoulos, Mihael H., Joseph J. Higgins, Lawrence I. Golbe, William G. Johnson, Susan E. Ide, Giuseppe Di Iorio, Giuseppe Sanges, et al. 1996. "Mapping of a Gene for Parkinson's Disease to

Chromosome 4q21-q23." *Science*. https://doi.org/10.1126/science.274.5290.1197.
Polymeropoulos, Mihael H., Christian Lavedan, Elisabeth Leroy, Susan E. Ide, Anindya Dehejia, Amalia Dutra, Brian Pike, et al. 1997. "Mutation in the α-Synuclein Gene Identified in Families with Parkinson's Disease." *Science*. https://doi.org/10.1126/science.276.5321.2045.
Poplin, Ryan, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, et al. 2018. "Scaling Accurate Genetic Variant Discovery to Tens of Thousands of Samples." *bioRxiv*. https://doi.org/10.1101/201178.
Prashanth, R., Sumantra Dutta Roy, Pravat K. Mandal, and Shantanu Ghosh. 2016. "High-Accuracy Detection of Early Parkinson's Disease through Multimodal Features and Machine Learning." *International Journal of Medical Informatics* 90 (June): 13–21.
Price, Alkes L., Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. 2006. "Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies." *Nature Genetics* 38 (8): 904–9.
Pu, Jia-Li, Zhi-Hao Lin, Ran Zheng, Yi-Qun Yan, Nai-Jia Xue, Xin-Zhen Yin, and Bao-Rong Zhang. 2022. "Association Analysis of SYT11, FGF20, GCH1 Rare Variants in Parkinson's Disease." *CNS Neuroscience & Therapeutics* 28 (1): 175–77.
Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *American Journal of Human Genetics* 81 (3): 559–75.
Rahayel, Shady, Ronald B. Postuma, Jacques Montplaisir, Daphné Génier Marchand, Frédérique Escudier, Malo Gaubert, Pierre-Alexandre Bourgouin, et al. 2018. "Cortical and Subcortical Gray Matter Bases of Cognitive Deficits in REM Sleep Behavior Disorder." *Neurology* 90 (20): e1759–70.
Raj, T., K. Rothamel, S. Mostafavi, C. Ye, M. N. Lee, J. M. Replogle, T. Feng, et al. 2014. "Polarization of the Effects of Autoimmune and Neurodegenerative Risk Alleles in Leukocytes." *Science* 344 (6183). https://doi.org/10.1126/science.1249547.
Real, Raquel, Anni Moore, Cornelis Blauwendraat, Huw R. Morris, Sara Bandres-Ciga, and International Parkinson's Disease Genomics Consortium (IPDGC). 2020. "ATP10B and the Risk for Parkinson's Disease." *Acta Neuropathologica*.
Reed, Xylena, Sara Bandrés-Ciga, Cornelis Blauwendraat, and Mark R. Cookson. 2019. "The Role of Monogenic Genes in Idiopathic Parkinson's Disease." *Neurobiology of Disease* 124 (April): 230–39.
Regier, Allison A., Yossi Farjoun, David E. Larson, Olga Krasheninina, Hyun Min Kang, Daniel P. Howrigan, Bo-Juen Chen, et al. 2018. "Functional Equivalence of Genome Sequencing Analysis Pipelines Enables Harmonized Variant Calling across Human Genetics Projects." *Nature Communications* 9 (1): 4038.
Regier, Allison A., Yossi Farjoun, David Larson, Olga Krasheninina, Hyun Min Kang, Daniel P. Howrigan, Bo-Juen Chen, et al. n.d. "Functional Equivalence of Genome Sequencing Analysis Pipelines Enables Harmonized Variant Calling across Human Genetics Projects." https://doi.org/10.1101/269316.
Ren, Yong, Jinghui Zhao, and Jian Feng. 2003. "Parkin Binds to Alpha/beta Tubulin and Increases Their Ubiquitination and Degradation." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 23 (8): 3316–24.
Riley, Ekemini A. U., Ekemini AU Riley, and Randy Schekman. 2021. "Open Science Takes on Parkinson's Disease." *eLife*. https://doi.org/10.7554/elife.66546.
Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. "Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47.
Rizzo, Giovanni, Massimiliano Copetti, Simona Arcuti, Davide Martino, Andrea Fontana, and Giancarlo Logroscino. 2016. "Accuracy of Clinical Diagnosis of Parkinson Disease: A Systematic Review and Meta-Analysis." *Neurology* 86 (6): 566–76.

Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. 2011. “Integrative Genomics Viewer.” *Nature Biotechnology* 29 (1): 24–26.

Ross, Owen A., Greggory J. Wilhoite, Justin A. Bacon, Alexandra Soto-Ortolaza, Jennifer Kachergus, Stephanie A. Cobb, Andreas Puschmann, et al. 2010. “LRRK2 Variation and Parkinson’s Disease in African Americans.” *Movement Disorders: Official Journal of the Movement Disorder Society* 25 (12): 1973–76.

Rudakou, Uladzislau, Jennifer A. Ruskey, Lynne Krohn, Sandra B. Laurent, Dan Spiegelman, Lior Greenbaum, Gilad Yahalom, et al. 2020. “Analysis of Common and Rare Variants in Late-Onset Parkinson Disease.” *Neurology. Genetics* 6 (1): 385.

Rudakou, Uladzislau, Eric Yu, Lynne Krohn, Jennifer A. Ruskey, Farnaz Asayesh, Yves Dauvilliers, Dan Spiegelman, et al. 2021. “Targeted Sequencing of Parkinson’s Disease Loci Genes Highlights SYT11, FGF20 and Other Associations.” *Brain: A Journal of Neurology* 144 (2): 462–72.

Ruopp, Marcus D., Neil J. Perkins, Brian W. Whitcomb, and Enrique F. Schisterman. 2008. “Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower Limit of Detection.” *Biometrical Journal. Biometrische Zeitschrift* 50 (3): 419–30.

Saffie-Awad, Paula, Inas Elsayed, Arinola O. Sanyaolu, Peter Wild Crea, Artur F. Schumacher Schuh, Kristin S. Levine, Dan Vitale, et al. 2023. “Evaluating the Performance of Polygenic Risk Profiling across Diverse Ancestry Populations in Parkinson’s Disease.” *medRxiv : The Preprint Server for Health Sciences*, November. https://doi.org/10.1101/2023.11.28.23299090.

Safiri, Saeid, Maryam Noori, Seyed Aria Nejadghaderi, Seyed Ehsan Mousavi, Mark J. M. Sullman, Mostafa Araj-Khodaei, Kuljit Singh, Ali-Asghar Kolahi, and Kurosh Gharagozli. 2023. “The Burden of Parkinson’s Disease in the Middle East and North Africa Region, 1990-2019: Results from the Global Burden of Disease Study 2019.” *BMC Public Health* 23 (1): 107.

Sauerbier, Anna, Peter Jenner, Antoniya Todorova, and K. Ray Chaudhuri. 2016. “Non Motor Subtypes and Parkinson’s Disease.” *Parkinsonism & Related Disorders* 22 Suppl 1 (January): S41–46.

Savica, R., J. M. Carlin, B. R. Grossardt, J. H. Bower, J. E. Ahlskog, D. M. Maraganore, A. E. Bharucha, and W. A. Rocca. 2009. “Medical Records Documentation of Constipation Preceding Parkinson Disease: A Case-Control Study.” *Neurology* 73 (21): 1752–58.

Schapira, Anthony H. V. 2015. “Glucocerebrosidase and Parkinson Disease: Recent Advances.” *Molecular and Cellular Neurosciences* 66 (Pt A): 37–42.

Schlebusch, Carina M., and Mattias Jakobsson. 2018. “Tales of Human Migration, Admixture, and Selection in Africa.” *Annual Review of Genomics and Human Genetics* 19 (August): 405–28.

Schmaderer, Theresa M., Clodagh Towns, Simona Jasaityte, Manuela M. X. Tan, Miriam Pollard, Megan Hodgson, Lesley Wu, et al. 2023. “Parkinson’s Families Project: A UK-Wide Study of Early Onset and Familial Parkinson’s Disease.” *medRxiv*. https://doi.org/10.1101/2023.12.05.23299397.

Scholz, Sonja W., and Jose Bras. 2015. “Genetics Underlying Atypical Parkinsonism and Related Neurodegenerative Disorders.” *International Journal of Molecular Sciences* 16 (10): 24629–55.

Schork, Nicholas J., Sarah S. Murray, Kelly A. Frazer, and Eric J. Topol. 2009. “Common vs. Rare Allele Hypotheses for Complex Diseases.” *Current Opinion in Genetics & Development* 19 (3): 212–19.

Schumacher-Schuh, Artur Francisco, Andrei Bieger, Olaitan Okunoye, Kin Ying Mok, Shen-Yang Lim, Soraya Bardien, Azlina Ahmad-Annuar, et al. 2022. “Underrepresented Populations in Parkinson’s Genetics Research: Current Landscape and Future Directions.” *Movement Disorders: Official Journal of the Movement Disorder Society* 37 (8): 1593–1604.

Sen, Saurabh, and Andrew B. West. 2009. “The Therapeutic Potential of LRRK2 and Alpha-Synuclein in Parkinson’s Disease.” *Antioxidants & Redox Signaling* 11 (9): 2167–87.

Sidransky, E., M. A. Nalls, J. O. Aasly, J. Aharon-Peretz, G. Annesi, E. R. Barbosa, A. Bar-Shira, et al. 2009. “Multicenter Analysis of Glucocerebrosidase Mutations in Parkinson’s Disease.” *The New England Journal of Medicine* 361 (17): 1651–61.

Silva de Lima, Ana Lígia, Luc J. W. Evers, Tim Hahn, Lauren Bataille, Jamie L. Hamilton, Max A. Little, Yasuyuki Okuma, Bastiaan R. Bloem, and Marjan J. Faber. 2017. “Freezing of Gait and Fall Detection in Parkinson’s Disease Using Wearable Sensors: A Systematic Review.” *Journal of Neurology* 264 (8): 1642–54.

Simón-Sánchez, Javier, Claudia Schulte, Jose M. Bras, Manu Sharma, J. Raphael Gibbs, Daniela Berg, Coro Paisan-Ruiz, et al. 2009. “Genome-Wide Association Study Reveals Genetic Risk Underlying Parkinson’s Disease.” *Nature Genetics* 41 (12): 1308–12.

“Single-Cell Transcriptomics of Parkinson’s Disease Human In Vitro Models Reveals Dopamine Neuron-Specific Stress Responses.” 2020. *Cell Reports* 33 (2): 108263.

Singleton, A. B., M. Farrer, J. Johnson, A. Singleton, S. Hague, J. Kachergus, M. Hulihan, et al. 2003. “Alpha-Synuclein Locus Triplication Causes Parkinson’s Disease.” *Science* 302 (5646): 841.

Singleton, Andrew, and John Hardy. 2011. “A Generalizable Hypothesis for the Genetic Architecture of Disease: Pleomorphic Risk Loci.” *Human Molecular Genetics* 20 (R2): R158–62.

Siva, Nayanah. 2008. “1000 Genomes Project.” *Nature Biotechnology* 26 (3): 256.

Smolka, Moritz, Luis F. Paulin, Christopher M. Grochowski, Medhat Mahmoud, Sairam Behera, Mira Gandhi, Karl Hong, et al. 2022. “Comprehensive Structural Variant Detection: From Mosaic to Population-Level.” *bioRxiv*. https://doi.org/10.1101/2022.04.04.487055.

Soldner, Frank, Yonatan Stelzer, Chikdu S. Shivalila, Brian J. Abraham, Jeanne C. Latourelle, M. Inmaculada Barrasa, Johanna Goldmann, Richard H. Myers, Richard A. Young, and Rudolf Jaenisch. 2016. “Parkinson-Associated Risk Variant in Distal Enhancer of α-Synuclein Modulates Target Gene Expression.” *Nature* 533 (7601): 95–99.

Stephens, Zachary D., Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, and Gene E. Robinson. 2015. “Big Data: Astronomical or Genomical?” *PLoS Biology* 13 (7): e1002195.

Stumvoll, M., A. Fritsche, A. Madaus, N. Stefan, M. Weisser, F. Machicao, and H. Häring. 2001. “Functional Significance of the UCSNP-43 Polymorphism in the CAPN10 Gene for Proinsulin Processing and Insulin Secretion in Nondiabetic Germans.” *Diabetes* 50 (9): 2161–63.

Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, et al. 2015. “UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age.” *PLoS Medicine* 12 (3): e1001779.

Takahashi, H., E. Ohama, S. Suzuki, Y. Horikawa, A. Ishikawa, T. Morita, S. Tsuji, and F. Ikuta. 1994. “Familial Juvenile Parkinsonism: Clinical and Pathologic Study in a Family.” *Neurology* 44 (3 Pt 1): 437–41.

Takamura, Shogo, Aya Ikeda, Kenya Nishioka, Hirokazu Furuya, Mari Tashiro, Takashi Matsushima, Yuanzhe Li, et al. 2016. “Schizophrenia as a Prodromal Symptom in a Patient Harboring SNCA Duplication.” *Parkinsonism & Related Disorders* 25 (April): 108–9.

Taliun, Daniel, Daniel N. Harris, Michael D. Kessler, Jedidiah Carlson, Zachary A. Szpiech, Raul Torres, Sarah A. Gagliano Taliun, et al. 2021. “Sequencing of 53,831 Diverse Genomes from the NHLBI TOPMed Program.” *Nature* 590 (7845): 290–99.

Tan, Manuela M. X., Michael A. Lawton, Edwin Jabbari, Regina H. Reynolds, Hirotaka Iwaki, Cornelis Blauwendraat, Sofia Kanavou, et al. 2021. “Genome-Wide Association Studies of Cognitive and Motor Progression in Parkinson’s Disease.” *Movement Disorders: Official Journal of the Movement Disorder Society* 36 (2): 424–33.

Tayebi, N., J. Park, V. Madike, and E. Sidransky. 2000. “Gene Rearranagement on 1q21 Introducing a Duplication of the Glucocerebrosidase Pseudogene and a Metaxin Fusion Gene.” *Human Genetics* 107 (4): 400–403.

Tesson, Christelle, Ebba Lohmann, David Devos, Hélène Bertrand, Suzanne Lesage, and Alexis Brice. 2020. “Segregation of ATP10B Variants in Families with Autosomal Recessive Parkinsonism.” *Acta*

Neuropathologica.
The Lancet. 2024. “What next in Parkinson’s Disease?” *The Lancet* 403 (10423): 219.
Toffoli, Marco, Xiao Chen, Fritz J. Sedlazeck, Chiao-Yin Lee, Stephen Mullin, Abigail Higgins, Sofia Koletsi, et al. 2022. “Comprehensive Short and Long Read Sequencing Analysis for the Gaucher and Parkinson’s Disease-Associated GBA Gene.” *Communications Biology* 5 (1): 670.
Toffoli, Marco, Harneek Chohan, Stephen Mullin, Aaron Jesuthasan, Selen Yalkic, Sofia Koletsi, Elisa Menozzi, et al. 2023. “Phenotypic Effect of GBA1 Variants in Individuals with and without Parkinson’s Disease: The RAPSODI Study.” *Neurobiology of Disease* 188 (November): 106343.
Toffoli, Marco, Laura Smith, and Anthony H. V. Schapira. 2020. “The Biochemical Basis of Interactions between Glucocerebrosidase and Alpha-Synuclein in GBA1 Mutation Carriers.” *Journal of Neurochemistry* 154 (1): 11–24.
Towns, Clodagh, Madeleine Richer, Simona Jasaityte, Eleanor J. Stafford, Julie Joubert, Tarek Antar, Alejandro Martinez-Carrasco, et al. 2023. “Defining the Causes of Sporadic Parkinson’s Disease in the Global Parkinson's Genetics Program (GP2).” *NPJ Parkinson’s Disease* 9 (1): 131.
Traag, V. A., R. Aldecoa, and J-C Delvenne. 2015. “Detecting Communities Using Asymptotical Surprise.” *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* 92 (2): 022816.
Trabzuni, Daniah, United Kingdom Brain Expression Consortium (UKBEC), and Peter C. Thomson. 2014. “Analysis of Gene Expression Data Using a Linear Mixed Model/finite Mixture Model Approach: Application to Regional Differences in the Human Brain.” *Bioinformatics* 30 (11): 1555–61.
Tysnes, Ole-Bjørn, and Anette Storstein. 2017. “Epidemiology of Parkinson’s Disease.” *Journal of Neural Transmission* 124 (8): 901–5.
Uehara, Yuya, Shin-Ichi Ueno, Haruka Amano-Takeshige, Shuji Suzuki, Yoko Imamichi, Motoki Fujimaki, Noriyasu Ota, et al. 2021. “Non-Invasive Diagnostic Tool for Parkinson’s Disease by Sebum RNA Profile with Machine Learning.” *Scientific Reports*. https://doi.org/10.1038/s41598-021-98423-9.
Valente, Enza Maria, Patrick M. Abou-Sleiman, Viviana Caputo, Miratul M. K. Muqit, Kirsten Harvey, Suzana Gispert, Zeeshan Ali, et al. 2004. “Hereditary Early-Onset Parkinson’s Disease Caused by Mutations in PINK1.” *Science* 304 (5674): 1158–60.
Van der Auwera, Geraldine A., Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo Del Angel, Ami Levy-Moonshine, Tadeusz Jordan, et al. 2013. “From FastQ Data to High Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline.” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 43: 11.10.1–11.10.33.
Van der Auwera, Geraldine A., and Brian D. O’Connor. 2020. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O’Reilly Media.
Vihinen, Mauno. 2023. “Systematic Errors in Annotations of Truncations, Loss-of-Function and Synonymous Variants.” *Frontiers in Genetics* 14 (January): 1015017.
Wang, Kai, Mingyao Li, and Hakon Hakonarson. 2010. “ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data.” *Nucleic Acids Research* 38 (16): e164.
Wang, K., M. Li, and H. Hakonarson. 2010. “ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkq603.
Wang, Rui, Hongyang Sun, Guanghui Wang, and Haigang Ren. 2020. “Imbalance of Lysine Acetylation Contributes to the Pathogenesis of Parkinson’s Disease.” *International Journal of Molecular Sciences* 21 (19). https://doi.org/10.3390/ijms21197182.
Welsh, Natalie J., Christina A. Gewinner, Kavita Mistry, Mumta Koglin, Juniebel Cooke, Matthew Butler, Ben Powney, Malcolm Roberts, James M. Staddon, and Anthony H. V. Schapira. 2020. “Functional Assessment of Glucocerebrosidase Modulator Efficacy in Primary Patient-Derived Macrophages Is Essential for Drug Development and Patient Stratification.” *Haematologica* 105 (5): e206–9.
Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer.
Willer, Cristen J., Yun Li, and Gonçalo R. Abecasis. 2010. “METAL: Fast and Efficient Meta-Analysis of

Genomewide Association Scans." *Bioinformatics* 26 (17): 2190–91.
Wishart, David S., Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. 2008. "DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets." *Nucleic Acids Research* 36 (Database issue): D901–6.
Wojewska, Dominika Natalia, and Arjan Kortholt. 2021. "LRRK2 Targeting Strategies as Potential Treatment of Parkinson's Disease." *Biomolecules* 11 (8). https://doi.org/10.3390/biom11081101.
Wu, Michael C., Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. 2011. "Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test." *American Journal of Human Genetics* 89 (1): 82–93.
Xie, Tao, Chuanhong Liao, Danielle Lee, Huiyan Yu, Mahesh Padmanaban, Wenjun Kang, Julie Johnson, et al. 2021. "Disparities in Diagnosis, Treatment and Survival between Black and White Parkinson Patients." *Parkinsonism & Related Disorders* 87 (June): 7–12.
Yang, Zongcheng, Xiuming Liang, Yue Fu, Yingjiao Liu, Lixin Zheng, Fen Liu, Tongyu Li, Xiaolin Yin, Xu Qiao, and Xin Xu. 2019. "Identification of AUNIP as a Candidate Diagnostic and Prognostic Biomarker for Oral Squamous Cell Carcinoma." *EBioMedicine* 47 (September): 44–57.
Yonova-Doing, Ekaterina, Masharip Atadzhanov, Marialuisa Quadri, Paul Kelly, Nyambura Shawa, Sheila T. S. Musonda, Erik J. Simons, Guido J. Breedveld, Ben A. Oostra, and Vincenzo Bonifati. 2012. "Analysis of LRRK2, SNCA, Parkin, PINK1, and DJ-1 in Zambian Patients with Parkinson's Disease." *Parkinsonism & Related Disorders* 18 (5): 567–71.
Zeggini, Eleftheria, and Andrew Morris. 2015. *Assessing Rare Variation in Complex Traits: Design and Analysis of Genetic Studies*. Springer.
Zhang, Jiayu, Wenchao Zhou, Hongmei Yu, Tong Wang, Xiaqiong Wang, Long Liu, and Yalu Wen. 2023. "Prediction of Parkinson's Disease Using Machine Learning Methods." *Biomolecules* 13 (12). https://doi.org/10.3390/biom13121761.
Zhang, X., C. Zhu, G. Beecham, B. N. Vardarajan, Y. Ma, D. Lancour, J. J. Farrell, et al. 2019. "A Rare Missense Variant of CASP7 Is Associated with Familial Late-Onset Alzheimer's Disease." *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 15 (3). https://doi.org/10.1016/j.jalz.2018.10.005.
Zimprich, Alexander, Saskia Biskup, Petra Leitner, Peter Lichtner, Matthew Farrer, Sarah Lincoln, Jennifer Kachergus, et al. 2004. "Mutations in LRRK2 Cause Autosomal-Dominant Parkinsonism with Pleomorphic Pathology." *Neuron* 44 (4): 601–7.
Zirra, Alexandra, Shilpa C. Rao, Jonathan Bestwick, Rajasumi Rajalingam, Connie Marras, Cornelis Blauwendraat, Ignacio F. Mata, and Alastair J. Noyce. 2023. "Gender Differences in the Prevalence of Parkinson's Disease." *Movement Disorders Clinical Practice* 10 (1): 86–93.