Identification of variants, genes and pathways in synucleinopathies using bioinformatics and machine learning

Eric Yu

Department of Human Genetics

Faculty of Medicine and Health Sciences

McGill University, Montreal, Quebec, Canada

August 2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

© Eric Yu, 2023

Abstract

Synucleinopathies are a group of neurodegenerative diseases characterized by the presence of alpha-synuclein in the brain of the patient. Synucleinopathies are primarily composed of Parkinson's disease (PD), dementia with Lewy body (DLB) and multiple system atrophies (MSA). Currently, there is no cure for any of the above-mentioned disorders. Since degeneration occurs before disease diagnosis, drugs and therapeutics slowing down or stopping disease progression are of crucial importance. Early diagnosis is important for patients to administer treatment before severe degeneration occurs. For example, REM-sleep behavior disorder (RBD) is considered one of the best predictors for synucleinopathies as more than 80% of patients phenoconvert to PD, DLB or MSA.

Genetic studies have been conducted to characterize the genetic landscape of synucleinopathies. Familial PD revealed monogenic PD genes such as *LRRK2*, *PRKN*, *PINK1*, and *DJ-1* and case-control studies identified PD risk factors such as *GBA1*. Genome-wide association studies (GWAS) nominated genes such as *GBA1*, *SNCA* and *TMEM175* that were central to PD, DLB and RBD. Many other genes were found to be distinct for a certain disorder such as *LRRK2* in PD and *APOE* in DLB. Although GWAS nominated numerous novel loci, most of these loci have unknown causal genes due to a lack of additional biological evidence.

In this thesis, I used bioinformatics and machine learning to characterize patients and identify novel genetic targets. In Chapter 2, I investigated the association of heterozygous *PRKN* single nucleotide variants (SNVs) and copy number variations (CNVs) with PD. While *PRKN* is an autosomal recessive PD gene, the role of heterozygous *PRKN* variants is controversial. Using targeted next-generation sequencing, we sequenced the coding and untranslated region of *PRKN*

in 2,809 PD patients and 3,629 healthy controls and performed CNV calling. After including both SNVs and CNVs, I examined the association between PD and heterozygous *PRKN* variants. The results from this study suggest that the frequency of heterozygous *PRKN* variants is similar between PD patients and control. These findings have applications in future clinical trials and precision medicine. For example, heterozygous carriers of *PRKN* variants should not be included in *PRKN* clinical trials but they may be included in other PD trials.

In Chapters 3 and 4, I performed fine-mapping of the *HLA* locus in PD, DLB and RBD. Although neuroinflammation is involved in disease pathogenesis, the role of *HLA* is unclear. *HLA* was nominated by the PD GWAS with a protective effect. However, due to the complex linkage disequilibrium (LD) of the region, it is unclear which genes and alleles are causal. Using *HLA* imputation methods, I imputed *HLA* alleles and amino acids with high accuracy from genotyping data. In PD, our results suggest that the protective association is driven by three *HLA-DRB1* gene residues in high LD: 11V, 13H and 33H. Meanwhile, *HLA-DRB1* 11:01 was associated with risk for RBD. No association was found in LBD. *HLA* genes could be potential therapeutic targets.

In Chapter 5, I conducted a multi-omic study on the PD GWAS to prioritize genes for future functional studies. To create a machine learning model specific to PD, I selected datasets from brain tissues and dopaminergic neurons and trained a multi-omic model from genetic, transcriptomic, epigenetic and distance measures. Then, I performed post hoc analyses such as pathway enrichment analysis, rare variant burden tests and structural analysis on the top genes from each locus. *IP6K2*, *ITPKB*, *PPIP5K2*, *INPP5F*, *SPNS1*, and *MLX* are the top genes in their respective locus. The inositol phosphate biosynthetic process was suggested to be involved in PD.

This thesis identified variants, genes and pathways in synucleinopathies using bioinformatics and machine learning. These results can be applied to clinical trials, drug discovery, precision medicine and functional studies.

Résumé

Les synucléinopathies sont un groupe de maladies neurodégénératives caractérisées par la présence d'alpha-synucléine dans le cerveau du patient. Les synucléinopathies sont principalement composées de la maladie de Parkinson (PD), de la démence à corps de Lewy (DCL) et des atrophies multisystémiques (AMS). Actuellement, il n'existe aucun remède pour l'un des troubles mentionnés ci-dessus. Étant donné que la dégénérescence se produit avant le diagnostic de la maladie, les médicaments et les thérapies ralentissant ou arrêtant la progression de la maladie sont d'une importance cruciale. Un diagnostic précoce est important pour que les patients puissent administrer un traitement avant qu'une dégénérescence sévère ne se produise. Par exemple, le trouble du comportement en sommeil paradoxal (RBD) est considéré comme l'un des meilleurs prédicteurs des synucléinopathies, car plus de 80% des patients phénoconvertissent en PD, DCL ou AMS.

Des études génétiques ont été menées pour caractériser le paysage génétique des synucléinopathies. La PD familiale a révélé des gènes de PD monogéniques tels que *LRRK2, PRKN, PINK1* et *DJ-1*, et des études cas-témoins ont identifié des facteurs de risque de PD tels que *GBA1*. Les études d'association pangénomique (GWAS) ont désigné des gènes tels que *GBA1, SNCA* et *TMEM175* qui étaient centraux pour la PD, la DCL et le RBD. De nombreux autres gènes se sont avérés distincts pour un certain trouble, comme *LRRK2* dans la PD et *APOE* dans la DCL. Bien que les GWAS aient désigné de nombreux nouveaux loci, la plupart de ces loci ont des gènes causaux inconnus en raison d'un manque de preuves biologiques supplémentaires.

Dans cette thèse, j'ai utilisé la bioinformatique et l'apprentissage automatique pour caractériser les patients et identifier de nouvelles cibles génétiques. Dans le chapitre 2, j'ai étudié l'association des

variants mononucléotidiques hétérozygotes (SNV) et des variations du nombre de copies (CNV) de *PRKN* avec la maladie de Parkinson (PD). Bien que *PRKN* soit un gène autosomique récessif de la PD, le rôle des variants hétérozygotes de *PRKN* est controversé. En utilisant le séquençage de nouvelle génération ciblé, nous avons séquencé la région codante et non traduite de *PRKN* chez 2 809 patients atteints de PD et 3 629 témoins sains et avons effectué l'appel des CNV. Après avoir inclus les SNV et les CNV, j'ai examiné l'association entre la PD et les variants hétérozygotes de *PRKN*. Les résultats de cette étude suggèrent que la fréquence des variants hétérozygotes de *PRKN* est similaire entre les patients atteints de PD et les témoins. Ces résultats ont des applications dans les futurs essais cliniques et la médecine de précision. Par exemple, les porteurs hétérozygotes de variants *PRKN* ne devraient pas être inclus dans les essais cliniques *PRKN*, mais ils pourraient être inclus dans d'autres essais de PD.

Dans les chapitres 3-4, j'ai effectué un mappage fin du locus *HLA* dans la PD, la DCL et le RBD. Bien que la neuroinflammation soit impliquée dans la pathogenèse de la maladie, le rôle du *HLA* n'est pas clair. Le *HLA* a été désigné par les GWAS de la PD avec un effet protecteur. Cependant, en raison de la liaison déséquilibrée (LD) complexe de la région, il n'est pas clair quels gènes et allèles sont causaux. En utilisant des méthodes d'imputation *HLA*, j'ai imputé les allèles et les acides aminés *HLA* avec une grande précision à partir des données de génotypage. Dans la PD, nos résultats suggèrent que l'association protectrice est entraînée par trois résidus du gène *HLA-DRB1* en LD élevé : 11V, 13H et 33H. Pendant ce temps, le *HLA-DRB1* 11:01 était associé au risque pour le RBD. Aucune association n'a été trouvée dans la DCL. Les gènes *HLA* pourraient être des cibles thérapeutiques potentielles.

Dans le chapitre 5, j'ai mené une étude multi-omique sur les GWAS de la PD pour hiérarchiser les gènes pour de futures études fonctionnelles. Pour créer un modèle d'apprentissage automatique

spécifique à la PD, j'ai sélectionné des ensembles de données provenant de tissus cérébraux et de neurones dopaminergiques et j'ai entraîné un modèle multi-omique à partir de mesures génétiques, transcriptomiques, épigénétiques et de distance. Ensuite, j'ai effectué des analyses post hoc telles que l'analyse d'enrichissement des voies, les tests de charge de variants rares et l'analyse structurale sur les gènes principaux de chaque locus. *IP6K2*, *ITPKB*, *PPIP5K2*, *INPP5F*, *SPNS1* et *MLX* sont les gènes principaux de leur locus respectif. Le processus de biosynthèse du phosphate d'inositol a été suggéré d'être impliqué dans la PD. Cette thèse a identifié des variants, des gènes et des voies dans les synucléinopathies en utilisant la bioinformatique et l'apprentissage automatique. Ces résultats peuvent être appliqués aux essais cliniques, à la découverte de médicaments, à la médecine de précision et aux études fonctionnelles.

Table of Content

ABSTRACT **2**

RÉSUMÉ **5**

TABLE OF CONTENT **8**

LIST OF ABBREVIATIONS **11**

LIST OF FIGURES **13**

LIST OF TABLES **14**

ACKNOWLEDGMENTS **15**

CONTRIBUTION TO ORIGINAL KNOWLEDGE **16**

FORMAT OF THE THESIS **17**

CONTRIBUTION OF AUTHORS **18**

CHAPTER 1: INTRODUCTION **20**

PARKINSON'S DISEASE 20

Epidemiology *20*

Clinical symptoms *20*

Neuropathology of Parkinson' disease *22*

Environmental risk factors of Parkinson's disease *23*

Treatment for Parkinson's disease *23*

Prodromal Parkinson's disease *24*

Isolated REM sleep behavior disorder *24*

Dementia with Lewy Bodies *25*

Familial genetic risk factors of Parkinson's disease *25*

SNCA 25

PRKN 26

LRRK2 28

Genome-wide association study (GWAS) of Parkinson's disease *29*

Genes nominated from Parkinson's disease genome-wide association studies *30*

GBA1 30

VPS13C 31

MAPT 32

TMEM175 33

GCH1 33

HLA *34*

Genetics of REM sleep behavior disorder and dementia with Lewy bodies *35*

Biological pathways involved in synucleinopathies *36*

Autophagy lysosomal pathway 36

Mitophagy 37

Neuroinflammation 38

Clinical interventions in Parkinson's disease *39*

HYPOTHESIS 40

OBJECTIVES 40

CHAPTER 2: ANALYSIS OF HETEROZYGOUS *PRKN* VARIANTS AND COPY NUMBER VARIATIONS IN PARKINSON'S DISEASE 43

ABSTRACT 44
INTRODUCTION 45
METHODS 46
Study Population *46*
Genetic analysis *47*
PRKN sequencing 47
Detection and validation of copy number variations 48
Quality Control of MIPs for CNV detection 49
Statistical Analysis *49*
RESULTS 52
Identification of PRKN SNVs and CNVs *52*
Heterozygous PRKN SNVs and CNVs are not associated with Parkinson's disease *53*
Heterozygous PRKN SNVs and CNVs are not associated with AAO of Parkinson's disease *55*
Identification of PRKN-associated parkinsonism patients *57*
DISCUSSION 58
REFERENCES 62

CHAPTER 3: FINE MAPPING OF THE *HLA* LOCUS IN PARKINSON'S DISEASE IN EUROPEANS 70

ABSTRACT 71
INTRODUCTION 71
RESULTS 72
Meta-analysis of HLA types in Parkinson's disease suggests a single association *72*
HLA haplotype analysis *74*
Meta-analysis of the association of HLA amino acid changes with Parkinson's disease *75*
*Conditional analyses confirm that DRB1*04 amino acid variants likely drive the association of the HLA locus with PD* *78*
DISCUSSION 79
METHODS 83
Study population *83*
Pre-imputation genotype quality control *84*
UK Biobank quality control *84*
Imputation *85*
Association analysis of common variants on chromosome 6 ..... *85*
HLA locus analysis *86*
HLA imputation 86
Statistical analysis 87
Code availability *88*
Data availability *88*
References *89*

CHAPTER 4: *HLA* IN ISOLATED REM SLEEP BEHAVIOR DISORDER AND LEWY BODY DEMENTIA. 95

ABSTRACT 97
Background and Objectives *97*
Methods *97*
Results *97*
Discussion *97*
INTRODUCTION 98

METHODS....99
Study population....*99*
Quality control....*100*
HLA imputation....*100*
Power calculations....*100*
Statistical analysis....*101*
Data availability....*101*
Code availability....*101*
RESULTS....101
DISCUSSION....103
REFERENCE....106

CHAPTER 5: IDENTIFICATION OF NOVEL VARIANTS, GENES AND PATHWAYS POTENTIALLY LINKED TO PARKINSON'S DISEASE USING MACHINE LEARNING. 110

ABSTRACT....111
INTRODUCTION....111
RESULTS....113
Machine learning model nominates PD-associated genes in each PD locus....*113*
Gene expression in specific PD-associated dopaminergic neuron subtypes is an important feature predicting PD-relevant genes....*116*
Differential gene expression of genes from the inositol phosphate biosynthetic pathway and MLX1 in PD....*118*
Structural analysis of SPNS1 and MLX....*118*
Gene enrichment nominates the inositol phosphate biosynthetic pathway as a novel pathway involved in PD....*121*
Pathway specific polygenic risk score of the inositol phosphate biosynthetic pathway is associated with PD 123
Association of rare variants with nominated PD genes....*125*
DISCUSSION....125
METHODS....129
General structure of the study....*129*
Definition of loci and genes within each locus....*129*
Feature preprocessing....*130*
Neighborhood scores....*131*
Machine learning model to prioritize genes....*131*
Functional enrichment analysis....*131*
Single-cell and bulk RNAseq analyses....*132*
Pathway polygenic risk score analyses....*132*
Rare variant burden analyses....*133*
Structural analysis....*133*
Data availability....*134*
Code availability....*134*
REFERENCES....134

CHAPTER 6: GENERAL DISCUSSION 142

PARKINSON'S DISEASE GENES NOMINATED BY MACHINE LEARNING....146

CHAPTER 7: CONCLUSIONS AND FUTURE DIRECTIONS.... 152

REFERENCE 154

APPENDICES.... 172

List of Abbreviations

Alzheimer's disease (AD)

Autophagy lysosomal pathway (ALP)

Copy number variation (CNV)

Dementia with Lewy body (DLB)

Early-onset Parkinson's Disease (EOPD)

Expression quantitative trait loci (eQTL)

Gaucher's disease (GD)

Genome-wide association study (GWAS)

Glucocerebrosidase (GCase)

Human leukocyte antigens (HLA)

Induced pluripotent stem cells (iPSC)

Isolated REM sleep behavior disorder (iRBD)

Linkage disequilibrium (LD)

Lysosomal Integral Membrane Protein-2 (LIMP-2)

Multiplex ligation-dependent probe amplification (MLPA)

Multiple systems atrophy (MSA)

Next-generation sequencing (NGS)

Parkinson’s disease (PD)

Parkinson’s disease with dementia (PDD)

Reactive oxygen species (ROS)

Restless leg syndrome (RLS)

Shared epitope (SE)

Single-nucleotide polymorphism (SNP)

Single nucleotide variant (SNV)

Untranslated region (UTR)

Video polysomnography (vPSG)

List of Figures

Chapter 1

- Figure 1. Clinical symptoms and time course of Parkinson's disease progression (22)

Chapter 2

- Figure 1. Flow chart of different analysis phases. (51)

Chapter 3

- Figure 1. Validation of previously associated top *HLA* locus SNP (rs112485576) in our cohort. (73)
- Figure 2. Association of the *HLA-DRB1* alleles and location of associated amino acids. (77)

Chapter 5

- Figure 1. Workflow summary. (113)
- Figure 2. Probability score of the Parkinson's disease GWAS candidate genes (114)
- Figure 3. Feature importance for the Parkinson's disease GWAS gene prioritization model (115)
- Figure 4. Waterfall plots for Parkinson's disease GWAS candidate genes. (117)
- Figure 5. Structural analysis of *SPNS1* p.L512M and *MLX* p.Q223R. (120)
- Figure 6. Volcano plots of Gene Ontology biological processes and cellular components. (122)

List of Tables

Chapter 2

- Table 1: Study populations included in the analysis after quality control and exclusion of biallelic pathogenic/likely pathogenic *PRKN* SNV and CNV carriers. (47)
- Table 2: Rare *PRKN* heterozygous SNV and CNV analysis for risk of Parkinson's disease using SKAT-O adjusted for age, sex, ethnicity and *GBA* and *LRRK2* status. (54)
- Table 3: Rare *PRKN* heterozygous SNV and CNV analysis for Age at Onset of Parkinson's disease adjusted age, sex, ethnicity and *GBA* and *LRRK2* status. (56)
- Table 4: Parkinson's disease patients with biallelic pathogenic and likely pathogenic *PRKN* SNVs and CNVs. (57)

Chapter 3

- Table 1: Meta-analyses of *HLA* alleles association. (76)
- Table 2: Meta-analyses of *HLA* haplotype association. (79)
- Table 3: Meta-analyses of *HLA* amino acid changes association. (79)

Chapter 4

- Table 1: Study population after quality control. (99)
- Table 2: *HLA* association in isolated REM sleep behavior disorder. (102)

Chapter 5

- Table 1: Meta-analyses of pathway-specific polygenic risk scores. (123)
- Table 2: Meta-analysis of rare variant analysis of putative causal genes. (124)

Acknowledgments

I would like to first thank my supervisor, Professor Ziv Gan-Or, for all the guidance and mentorship throughout this journey. He believed in me from the start and always pushed me to be better. To my friends Mehrdad, Konstantin, Lynne, Vlad, Jenn, Khair, Dan, Lang, Emma and fellow NAP-Med lab members, thank you for the help and the great time we spent working together. I would also like to thank Dr. Alain Dagher and Dr. Simon Gravel for their feedback and contribution on the supervisory committee. Thank you to Ross and Rimi for helping all graduate students from the department. Thank you to Anjie, Zac, Lilit and other HGEN students who supported me and the fun times we had. To my family, thank you for help and support throughout everything. To my sweet and beloved girlfriend Audrey, thank you for giving me the strength to complete my studies.

Contribution to original knowledge

This thesis contains work of original knowledge. Chapter 2 showed the lack of evidence between heterozygous variants in *PRKN* and risk for PD and age at disease onset. Chapter 3 nominated *HLA* alleles, haplotypes and amino acid associated with PD. In Chapter 4, the *HLA* locus was examined in RBD and LBD and showed association between *HLA* alleles and RBD. In Chapter 5, I performed machine learning by building a model to examine the most likely causal genes in the PD GWAS. This model included multi-omics data such as transcriptomics, epigenetics, single-cell RNA sequencing from dopaminergic neurons.

Format of the thesis

The thesis uses a manuscript-based format. It contains two published manuscripts (Chapter 2-3) and two in preparation for submission (Chapter 4-5).

Chapter 2. Analysis of heterozygous *PRKN* variants and copy number variations in Parkinson's disease. *Movement Disorders*. 2021 Jan;36(1):178-187.

Chapter 3. Fine mapping of the *HLA* locus in Parkinson's disease in Europeans. *npj Parkinsons Dis*. 2021 Sep 21;7(1):84

Chapter 4. *HLA* in isolated REM sleep behavior disorder and Lewy body dementia. Submitting to *Annals of Clinical and Translational Neurology*.

Chapter 5. Identification of novel variants, genes and pathways potentially linked to Parkinson's disease using machine learning. Submitting to *Nature Communications*.

Supplementary files:

https://drive.google.com/drive/folders/1l3K4UBFI1TiUcwNlQM_2HHENFGojXiIT?usp=share_link

Contribution of authors

Chapter 2

- EY and ZGO conceptualized and designed the study.
- UR, LK, KM, JAR, FA, MAE, DS, MS, SF, CHW, LG, AJE, YD, ND, GAR, SH-B, EAF, RNA and ZGO had major role in the acquisition of the data.
- EY analyzed the data and drafted the manuscript.
- UR, LK, KM, JAR, FA, MAE, DS, MS, SF, CHW, LG, AJE, YD, ND, GAR, SH-B, EAF, RNA and ZGO revised the manuscript.

Chapter 3

- EY, EM and ZGO conceptualized and designed the study.
- AA, MSA, LK, JAR, FA, DS, MT, MKV and LP had major role in the acquisition of the data.
- EY analyzed the data and drafted the manuscript.
- AA, MSA, LK, MAE, PS, KS, YLS, AAKS, JAR, FA, DS, MT, MKV, MS, CB, LP, EM and ZGO revised the manuscript.

Chapter 4

- EY and ZGO conceptualized and designed the study.
- JAR, FS, DS, IA, MTMH, JYM, JFG, AD, YD, GLG, MV, FJ, AB, BH, AS, AI, AH, KS, PD, DK, WO, AJ, GP, EA, MF, MP, BM, CT, FSD, VCDC, LFS, FD, MV, BA, BFB. GAR, RBP, SWS and ZGO had major role in the acquisition of the data.

- EY, LK, KS and RC analyzed the data.
- EY drafted the manuscript.
- LK, JAR, FS, DS, ZS, RC, IA, MTMH, JYM, JFG, AD, YD, GLG, MV, FJ, AB, BH, AS, AI, AH, KS, PD, DK, WO, AJ, GP, EA, MF, MP, BM, CT, FSD, VCDC, LFS, FD, MV, BA, BFB. GAR, RBP, SWS and ZGO revised the manuscript.

Chapter 5

- EY and ZGO conceptualized and designed the study.
- EY, LL, KS and JFT had major role in the acquisition of the data.
- EY, LL, KS and JFT analyzed the data.
- EY drafted the manuscript.
- RL, RAT, LL, KS, SR, JFT, EAF and ZGO revised the manuscript.

Chapter 1: Introduction

Neurodegeneration is characterized by the progressive loss of neuronal structure and cell death. A common pathological hallmark of neurodegenerative disorders is the accumulation of protein aggregates. For example, synucleinopathies are a group of diseases associated with the inclusion bodies of the protein alpha-synuclein.[1] These neuronal inclusions can be found in the form of Lewy bodies or Lewy neurites (e.g., axon or dendrite). Synucleinopathies primarily consist of three diseases: Parkinson's disease (PD), dementia with Lewy bodies (DLB) and multiple systems atrophy (MSA), which has non-neuronal inclusions.[1] In this thesis, PD and DLB will be the focus, as MSA is a rare condition in comparison.

Parkinson's disease

Epidemiology

In 1817, James Parkinson first described PD in his work "An Essay on the Shaking Palsy."[2] PD is a slowly progressing neurodegenerative disorder that affected more than six million people worldwide in 2016.[3,4] PD is commonly diagnosed between the ages of 55 and 65.[5] However, early-onset PD (EOPD), with onset before the age of 50, may also occur in familial PD.[6] Since 1990, the age-standardized prevalence rate of PD has rapidly increased by 21.7%.[4] With the current rate of change in prevalence, the number of PD patients worldwide is likely to double between 2005 and 2030.[3]

Clinical symptoms

PD can be diagnosed using the UK PD Society Brain Bank Diagnostic Criteria and the Movement Disorder Society PD Criteria.[7,8] As shown in Figure 1, it includes the diagnosis of parkinsonism based on the manifestation of bradykinesia and two other essential motor symptoms: rigidity and

tremor at rest. Bradykinesia is characterized by a decline in the speed and amplitude of movements, especially in the limbs of patients. Rigidity refers to muscular rigidity of the neck and limbs of patients. Rest tremor is defined as a 4-6Hz tremor of a limb at rest. However, parkinsonism symptoms can also be caused by drugs, vascular conditions, brain damage, and viral infections, so these factors must not be present in PD patients.[7] Some rare conditions such as progressive supranuclear palsy and frontotemporal dementia also have atypical parkinsonism symptoms.[9]

PD is not solely characterized by motor symptoms. Patients can present with a range of non-motor symptoms, some of which can occur during the prodromal phase even before the onset of motor symptoms.[10] These non-motor symptoms can greatly impact the quality of life of individuals with PD. Some common non-motor symptoms of PD include hyposmia (reduced ability to smell), constipation, isolated REM sleep behavior disorder (iRBD) and depression.[10] iRBD describes patients with RBD during the prodromal phase without parkinsonism. Meanwhile, RBD can also occur after PD disease onset.

After diagnosis, the disease prognosis of patients will vary considerably over the next 10 years, on average.[11] In the early stage, most patients have milder motor symptoms and respond well to treatment.[12] Patients at later stages may be affected by postural instability, cognitive decline and dementia.[12] A previous systematic review showed that around 30% of PD patients develop PD with dementia.[13] It can occur in 75% of PD patients who survived more than 10 years.[13] Other patients may develop psychosis, sleep-wake cycle dysregulation and autonomic failure.[11]

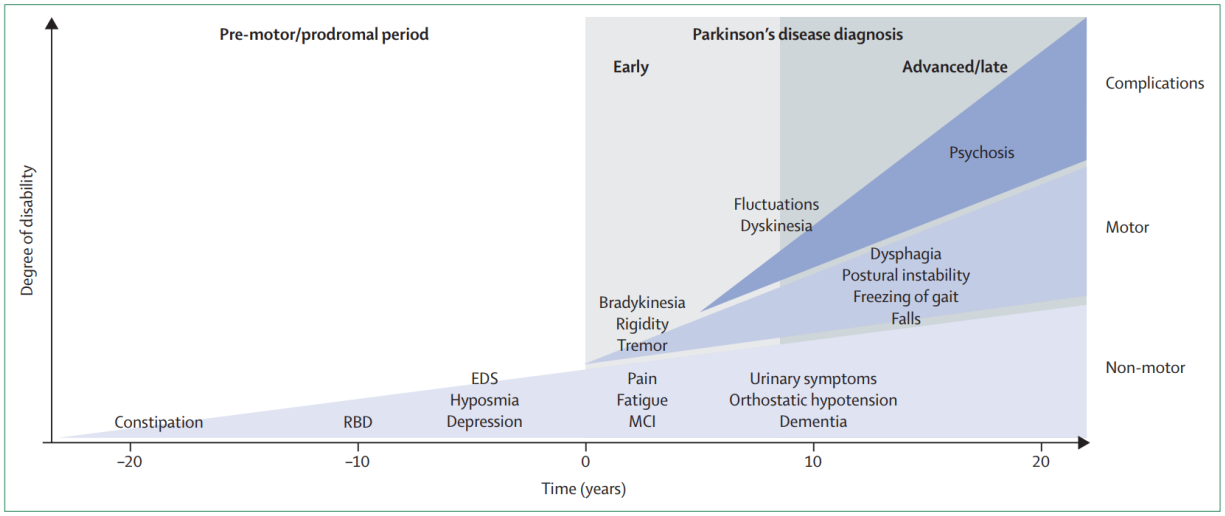


Figure 1. Clinical symptoms and time course of Parkinson's disease progression

EDS=excessive daytime sleepiness. MCI=mild cognitive impairment. RBD=REM sleep behaviour disorder. Adapted from Kalia, *et al.*, 2015.

Neuropathology of Parkinson' disease

PD is caused by the progressive loss of dopaminergic neurons in the substantia nigra pars compacta.[14] However, PD pathology is not restricted to the substantia nigra but may spreads across various brain areas and neuronal cell types.[15] A previous study suggested a stage-based progression called the Braak hypothesis, using post-mortem brain tissue from PD cases with Lewy bodies and Lewy neurites.[16] In stage 1, pathology starts in the olfactory bulb and the dorsal motor nucleus of the vagus nerve, and then it travels up the brainstem to the dorsal pons in stage 2. Next, in stage 3, it reaches the midbrain substantia nigra. Afterwards, the disease progresses to the thalamus and then the prefrontal cortex. At the last stage, stage 6, it affects the neocortex, an area involved in many cognitive processes.

However, this system does not describe the complexity of PD. For example, many subjects in Braak's study with Braak stage 4 or higher did not have PD.[17] Also, Lewy pathology is not an

indicator of PD, as many PD patients lack Lewy bodies, such as around half of the patients carrying *LRRK2* mutations.[18]

Environmental risk factors of Parkinson's disease

While the involvement of Lewy bodies is not well understood, many environmental risk factors are associated with PD. For instance, aging is the strongest risk factor for PD.[19] Previous studies have shown an increasing disease prevalence in the older population.[20] Although the underlying mechanism is not well understood, aging is associated with numerous physiological changes that can lead to impaired cellular function, such as lysosomal dysfunction, mitochondrial dysfunction and oxidative stress.[21] This association suggests that aging predisposes individuals to PD through decades of accumulation of environmental exposure or genetic mechanisms. PD also affects males and females at a ratio of 3:2.[22] Estrogen has been suggested to convey a protective effect for PD, as there is an increased risk of parkinsonism in women who have undergone ovary removal surgery.[23]

In a meta-analysis, 11 environmental risk factors for PD were discovered.[24] Pesticide exposure, head injury, rural living, beta-blocker usage, agricultural occupation, dairy and well-water drinking were shown to increase the risk. Smoking, coffee, non-steroidal anti-inflammatory drug use, calcium channel blocker use, and alcohol consumption were associated with a decreased risk of PD.

Treatment for Parkinson's disease

Current therapies are only available to treat PD symptoms. Levodopa, a dopamine precursor, is commonly prescribed for motor symptoms.[25] Dopamine agonists and monoamine oxidase type B

inhibitors have also been shown to be beneficial.[26,27] However, drug-induced adverse reactions can be debilitating and greatly affect the quality of life of patients. For example, long-term use of levodopa is associated with dyskinesia and motor fluctuations.[28] Dopamine agonists can cause impulse control disorders and hallucinations.[29] Despite advances in the drug industry since the discovery of PD, none of the drugs available are neuroprotective or disease-modifying.

Prodromal Parkinson's disease

Since the loss of dopaminergic neurons may initiate decades before the onset of motor symptoms, early diagnosis to define the prodromal stage of PD is necessary.[10] IRBD is currently one of the strongest predictors of synucleinopathies.[10] Previous studies have shown that more than 80% of iRBD patients will convert mainly to PD and DLB after 10 years from disease onset.[30]

Isolated REM sleep behavior disorder

During REM sleep where dreaming occurs, muscle paralysis or muscle atonia prevents us from enacting our dreams. IRBD is a parasomnia characterized by the loss of muscle atonia.[31] One of the characteristic symptoms of iRBD is the enactment of dreams using movement and vocalization during REM sleep.[31] These movements may involve complex behaviors such as crying, kicking, and punching.

Around 30% of PD, 50-80% of DLB, and 80-94% of MSA patients develop RBD symptoms after disease onset.[32] However, even though 80% of iRBD patients phenoconvert, iRBD patients do not have parkinsonism symptoms at disease onset.[31]

IRBD is diagnosed using video polysomnography (vPSG) to confirm that the motor events occur during REM sleep.[31] However, vPSG is expensive, time-consuming, and not available at all clinical centers, making it difficult to estimate the prevalence of iRBD from epidemiological

studies. Other screening methods may misdiagnose other sleep conditions as RBD, such as sleepwalking and non-REM sleep parasomnia. Currently, the prevalence of iRBD is estimated to be around 1% in individuals above 60 years old.[33–35]

Clonazepam, a medication for seizures, and melatonin can be prescribed to improve RBD symptoms.[31] However, no treatment can prevent phenoconversion or slow down disease progression.

Dementia with Lewy Bodies

DLB is a synucleinopathy characterized by parkinsonism symptoms and cognitive fluctuations, such as attention deficit and executive dysfunction.[26] In contrast to Alzheimer's disease (AD), visual hallucinations are more likely to manifest in DLB.[36] The relationship between Parkinson's disease with dementia (PDD) and DLB has been a subject of controversy. DLB typically has milder motor symptoms that may worsen as the disease progresses.[37] PDD is diagnosed when prominent motor symptoms are present, with cognitive symptoms developing in later stages. To distinguish between PDD and DLB, DLB is diagnosed when dementia occurs at least one year before parkinsonism.[38,39]

Despite careful assessment, the accurate diagnosis of dementia due to AD, DLB, or PD has proven to be challenging due to the lack of strong biomarkers.

Familial genetic risk factors of Parkinson's disease

SNCA

The first genetic evidence for PD was found in 1997 by Polymeropoulos *et al* through the discovery of the *SNCA* p.A53T mutation.[40] This mutation was identified in a large Italian family with

autosomal dominant PD. Subsequently, many other pathogenic mutations in the *SNCA* gene, such as p.A30P, E46K, and G51D, were discovered.[41]

SNCA encodes alpha-synuclein, which is a major component of Lewy bodies, a pathological feature of PD. For instance, *SNCA* triplication has been found to cause PD in a large family from the Midwestern US.[42] Additional studies have also discovered *SNCA* duplication as a genetic alteration associated with PD.[43] *SNCA* copy number variations (CNV) have been linked to earlier age at onset, rapid disease progression, severe cognitive fluctuations, and Lewy body pathology compared to idiopathic PD. In cases of duplications, the clinical and pathological presentations tend to be milder than triplication with reduced penetrance.[44]

While the biological function of *SNCA* remains unclear, overexpression of alpha-synuclein has been associated with protein aggregation and neuronal dysfunction.[43] Alpha-synuclein has been implicated in various cellular pathways, including membrane interactions, protein degradation, synaptic vesicle function, dopamine release and transport, mitochondrial dysfunction, and the autophagy-lysosomal pathway.[45]

According to the current disease model, alpha-synuclein monomers assemble into oligomers (e.g., tetramers), which then form fibrils.[46] Alpha-synuclein fibrils are the predominant form found in Lewy bodies and Lewy neurites.[47] However, the presence of incidental Lewy pathology (Lewy pathology in healthy individuals) and the lack of success in synuclein-targeted therapy challenge this model.[48]

PRKN

PRKN (also known as *PARK2*) is an autosomal recessive gene that is associated with early-onset Parkinson's disease (EOPD).[49] It is the most common cause of familial PD, with biallelic *PRKN*

variants estimated to have a prevalence of 4.3% in sporadic EOPD cases and 15.5% in familial PD cases.[50,51] While homozygous and compound heterozygous variants have been found to cause PD, the role of heterozygous variants has been controversial.[52–61] However, many studies did not examine copy number variations (CNVs) in the *PRKN* gene. *PRKN* is located in a common chromosomal fragile site, making it prone to deletions and duplications.[62] This region is often affected by chromosomal breaks due to collisions between replication and transcription complexes in large genes like *PRKN*, which spans 1.4Mb and produces a 4kb mature mRNA.[63]

Parkin, the protein encoded by *PRKN*, is an E3 ubiquitin ligase involved in mitophagy.[49] It interacts with *PINK1* to activate the ubiquitin-proteasome system and promote the degradation of dysfunctional mitochondria through autophagy.[64] *PINK1* is another autosomal recessive cause of EOPD.[65] Accumulation of reactive oxygen species (ROS) in mitochondria can induce mitophagy as a protective mechanism against ROS-mediated cellular damage.[66] *PRKN* has also been suggested to play a role in the innate immune system.[67]

Neuropathological studies of patients with biallelic *PRKN* variants have revealed distinct characteristics compared to idiopathic PD.[68] For example, *PRKN*-PD patients lack Lewy bodies pathology.[69] Neuropathology is limited to the substantia nigra and locus coeruleus.[69] These patients also tend to have milder PD symptoms and a slow disease progression.[70]

Despite the lack of current clinical trials specifically targeting *PRKN*, it is important to distinguish *PRKN*-PD patients from other PD patients due to differences in pathology and disease prognosis. Many PD patients who are carriers of *PRKN* heterozygous variants may carry cryptic variants that were not detected by next-generation sequencing (NGS) methods.[71] Currently, multiplex ligation-dependent probe amplification (MLPA) is considered the gold standard for

detecting CNVs.[72] Additional efforts are required to detect both single nucleotide variants (SNVs) and CNVs in clinical trials to minimize the risk of false negatives.

LRRK2

LRRK2 (leucine-rich repeat protein kinase-2) is another gene that has been discovered through familial studies and is associated with PD. It was first identified in 2004, and the *LRRK2* p.R1441G variant was found to be the causal variant for autosomal dominant PD in Spain.[73] In case-control studies, the *LRRK2* p.G2019S variant has been identified as one of the strongest risk variants for PD in Europeans.[74] This variant has been reported in approximately 4% of familial PD cases and 1% of sporadic cases.[75] In certain populations, such as North African Berbers and Ashkenazi Jews, the prevalence of the p.G2019S variant can be as high as 40% and 26%, respectively.[76,77] The penetrance of this variant is estimated to be between 25-42% at age 80 in Europeans.[78]

In addition to p.G2019S, many other pathogenic variants of *LRRK2* have been discovered, such as p.N1437H, p.R1441G/C/H/S, and p.Y1699C.[79,80] These variants have been found to increase the kinase activity of *LRRK2*. However, the specific kinase substrates of *LRRK2* in PD are still not fully understood. Previous studies have also identified a common haplotype (*LRRK2* p.N551K-p.R1398H-p.K1423K) that is associated with decreased kinase activity and reduced risk for PD.[81]

LRRK2 has been implicated in various pathways, including inflammatory pathways, lysosomal function, and autophagy.[82,83] Many of these mechanisms are interconnected with the pathophysiology of PD.

The clinical presentation of *LRRK2*-associated PD resembles sporadic PD, but with milder symptoms and slower disease progression. Hyposmia and RBD are also less frequently observed

in *LRRK2*-associated PD. However, individuals with *LRRK2* mutations tend to have an earlier age at onset and are more prone to postural instability and gait difficulties.

The neuropathology of *LRRK2*-PD differs considerably from sporadic PD. Only 62% of patients was reported with Lewy body pathology and 71% with tau pathology.[84]

Genome-wide association study (GWAS) of Parkinson's disease

Performing GWAS has been instrumental in examining the genetic landscape of PD. GWAS involves testing the association of single nucleotide polymorphisms (SNPs) across the genome.[85] Given the large number of SNPs tested, multiple correction testing is essential, and meta-analysis of large population studies increases the statistical power to detect associations.[85]

The largest PD GWAS to date, which included 37,688 cases, 18,618 proxy cases (healthy individuals with relatives with PD), and 1.4 million controls, identified 90 independent risk variants across 78 genomic risk loci.[74] Notably, the PD GWAS identified genome-wide significant variants within the *SNCA* and *LRRK2* loci.

In the *SNCA* locus, the most statistically significant signal in the PD GWAS was found in SNPs located in the 3' untranslated region (UTR) of the gene. This highlights the importance of regulatory regions in the genetic susceptibility to PD.

Regarding the *LRRK2* locus, the main signal identified in the PD GWAS was the p.G2019S variant. However, another independent SNP (rs76904798) within the *LRRK2* locus was also identified as associated with PD risk.[74]

Despite the valuable insights provided by GWAS in nominating novel variants and loci associated with polygenic traits, there are several challenges that remain in the field. One of the challenges is that GWAS typically explain only a small fraction of the heritability of polygenic

traits.[86] This is because GWAS primarily capture the effects of common variants, while rare variants, CNVs, and interaction effects can be missed.[87] Additionally, the majority of the nominated variants are located in non-coding regions of the genome, making it unclear how these variants functionally contribute to the traits or diseases of interest and which genes they affect.

Another challenge arises from the correlation between variants due to linkage disequilibrium (LD). While LD can be leveraged to identify novel loci associated with a trait, it becomes more difficult to discern the causal variants and genes within these loci.[87] Fine-mapping methods have been developed to tackle this challenge by using GWAS summary statistics and LD structure to nominate causal variants.[88,89]

In this thesis, I will further discuss these GWAS genes and their implications in the context of future chapters. Understanding the genetic factors and risk loci associated with PD can provide insights into disease mechanisms, potential therapeutic targets, and personalized approaches to treatment. Addressing GWAS limitation and incorporating more comprehensive and context-specific functional data will be crucial for further unraveling the complex genetic architecture underlying polygenic traits and diseases.

Genes nominated from Parkinson's disease genome-wide association studies

GBA1

GBA1 (previously known as *GBA*) is a gene that is associated with the autosomal recessive form of Gaucher's disease (GD), which is a lysosomal storage disorder.[90] Variants in the *GBA1* gene have been found to be prevalent in different populations, ranging from 5% in Asians to 20% in the Ashkenazi Jewish population. Hundreds of *GBA1* variants have been identified, including p.N370S, p.E326K, and p.L444P.[91]

The penetrance of *GBA1* variants in PD is relatively low, estimated to be between 10-30%.[92] *GBA1* variants are classified based on their association with the severity of Gaucher's disease, where severe variants are linked to the neuronopathic form of GD, and milder variants are associated with the non-neuronopathic form. In PD, carriers of severe *GBA1* variants typically exhibit more severe PD symptoms and experience faster disease progression.[93]

Compared to idiopathic PD, *GBA1*-PD tends to have earlier disease onset and faster cognitive decline.[94,95] *GBA1*-PD is characterized by strong Lewy pathology, that is widespread across the brain.[41]

The *GBA1* gene encodes the enzyme Glucocerebrosidase (GCase), which is located in the lysosome and is responsible for the degradation of glycolipids. In PD, pathogenic *GBA1* variants lead to a decrease in GCase activity and loss of function.[96] Reduced GCase activity has been shown to contribute to the accumulation of alpha-synuclein, a protein associated with PD pathogenesis, and may also induce neuroinflammation.[97] Additionally, the substrate of GCase is thought to interact with alpha-synuclein, promoting the formation of alpha-synuclein fibrils.[98]

VPS13C

VPS13C is one of the candidate genes from the PD GWAS. It has been implicated in the development of EOPD when rare homozygous or compound heterozygous variants are present.[99,100] Patients carrying these variants typically experience rapid disease progression and early cognitive dysfunction.[99]

The *VPS13C* gene is responsible for encoding a protein called vacuolar protein sorting 13C, which is partially localized in the mitochondrial membrane. Studies have shown that silencing of

VPS13C leads to mitochondrial dysfunction, which can disrupt cellular energy production and other mitochondrial functions.[99]

Furthermore, *VPS13C* has been implicated in the process of mitophagy, which is the selective removal of damaged or dysfunctional mitochondria through the Parkin/PINK1 pathway. Dysfunction in this pathway has been associated with impaired clearance of damaged mitochondria, leading to their accumulation and subsequent cellular damage. Studies have demonstrated that the silencing of *VPS13C* can disrupt Parkin/PINK1-dependent mitophagy, further contributing to mitochondrial dysfunction and potentially contributing to PD pathogenesis.[99]

MAPT

MAPT encodes the microtubule-associated protein tau, which is a neuronal protein involved in the formation of misfolded tau aggregates known as tauopathies.[101] Tauopathies can be found in conditions such as AD, PD, DLB, and MSA. More than half of AD autopsies have revealed the presence of Lewy bodies, and many PD patients also exhibit tau pathology.[102,103]

The *MAPT* locus is characterized by two main haplotypes: H1 and H2.[104] The H1 haplotype is present in all populations, while the H2 haplotype, which encompasses a 900kb inversion, is found in approximately 20% of Europeans and is associated with a reduced risk for PD.[105,106] The association between *MAPT* haplotypes and disease risk has been well-established.[107] However, the association of specific *MAPT* variants with disease remains unclear and has yielded controversial results in genotype-phenotype association studies.[108]

It has been observed that tau and alpha-synuclein frequently coexist in neurodegenerative disorders.[109] Tau pathology and alpha-synuclein may interact and co-localize, promoting the

misfolding of each other.[110] Tau is known to play a role in microtubule assembly, neuronal polarity, and axonal functions.[111] It is also involved in regulating neuronal plasticity, nucleolar organization, and providing DNA protection against oxidative stress.[112] Soluble tau can aggregate in the cytoplasm, forming neurofibrillary tangles, which can lead to neuronal dysfunction and death by disrupting microtubule assembly and transport.[113]

TMEM175

TMEM175 is identified as one of the most significant risk loci in the PD GWAS.[74] It encodes a transmembrane protein which functions as a lysosomal potassium or proton channel protein.[114] In the GWAS, two specific variants of *TMEM175* were highlighted: p.M393T as a risk variant and p.Q65P as a protective variant.[115]

One study examined the structure and molecular activity of *TMEM175* variants have provided insights into their functional implications.[115] The p.M393T variant is associated with reduced activity of the enzyme GCase, impairing its assembly, maturation, or trafficking within the lysosome. GCase plays a crucial role in the degradation of glycolipids, and its dysfunction has been implicated in PD.

On the other hand, structural analysis of the p.Q65P variant suggests that it leads to increased stability of the *TMEM175* transmembrane protein.[115] The exact mechanisms by which this variant confers a protective effect are not yet fully understood.

GCH1

GTP cyclohydrolase 1 (*GCH1*), is another candidate gene identified in the PD GWAS. *GCH1* is the rate-limiting enzyme in the synthesis of tetrahydrobiopterin, which is an essential cofactor for the production of dopamine.[116] Mutations in the *GCH1* gene can have implications for dopamine

synthesis and are associated with various disorders. In particular, rare mutations in *GCH1* are also associated with PD.[117]

HLA

Human leukocyte antigens (*HLA*) are cell-surface proteins involved in antigen presentation.[118] These proteins are encoded by genes located in the major histocompatibility complex (MHC), which is a highly polymorphic locus. The MHC contains two main classes of proteins: class I and class II, which have different roles in the immune system.

MHC class I proteins, encoded by *HLA-A*, *HLA-B*, and *HLA-C* genes, present foreign peptides derived from intracellular proteins to CD8+ T cells.[119] These proteins are expressed on the surface of all nucleated cells. On the other hand, MHC class II proteins are primarily found on antigen-presenting cells such as dendritic cells and B lymphocytes. *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1*, and *HLA-DRB1* are some of the highly polymorphic MHC class II genes. MHC class II proteins present antigens to CD4+ T cells, leading to T cell receptor activation and subsequent immune responses.

The *HLA* locus has been identified as a candidate locus in PD GWAS.[74] However, previous studies have yielded conflicting results, and the specific causal genes and variants within the *HLA* locus remain unclear.[74,120,121] It is worth noting that the *HLA* locus has also been associated with other neurodegenerative disorders, such as AD and amyotrophic lateral sclerosis.[122,123]

Further research is necessary to elucidate the role of *HLA* genes and their variants in the development and progression of PD, as well as their potential involvement in other neurodegenerative disorders.

Genetics of REM sleep behavior disorder and dementia with Lewy bodies

Genetic studies focusing on RBD and DLB have been conducted on smaller cohorts compared to PD. Despite the smaller sample sizes, these studies have provided valuable insights into the genetic landscape of synucleinopathies and RBD.

In a GWAS targeting LBD, which primarily included DLB patients and a smaller cohort of PD with dementia (PDD) patients, two novel risk loci were identified: *TMEM175* and *BIN1*.[124] Other studies have also demonstrated associations between DLB and genes such as *GBA1*, *APOE*, and *SNCA*.[125]

Genetic studies on RBD have revealed a partial genetic overlap between RBD and synucleinopathies. A RBD-specific GWAS identified associations between RBD and genes such as *GBA1*, *SNCA*, *TMEM175*, *SCARB2*, and *INPP5F*.[126] However, it is important to note that while some genes and variants, such as *LRRK2* and *APOE*, exhibited associations specific to PD and DLB, *GBA1*, *TMEM175* and *SNCA* showed associations with all three disorders.[127,128]

Furthermore, independent variants located at the 5' and 3' UTRs of *SNCA* were associated with different disorders. The 3' UTR variants were specifically associated with PD, whereas the 5' variants showed associations with DLB and RBD.

These findings highlight the complex genetic landscape of synucleinopathies, RBD, and DLB, indicating both shared and distinct genetic factors contributing to these disorders. Further research is needed to fully understand the precise genetic mechanisms and implications of these associations in different synucleinopathies.

Biological pathways involved in synucleinopathies

Autophagy lysosomal pathway

The lysosome plays a crucial role in the autophagy-lysosomal pathway (ALP), which is responsible for the degradation of various cellular components, including proteins, lipids, and organelles.[129] In PD, the ALP is particularly important for the degradation of alpha-synuclein, a protein whose accumulation is associated with the disease.[130]

Several genes implicated in synucleinopathies are directly involved in the ALP. For instance, *GBA1* encodes for GCase, a protein located on the inner surface of the lysosome. Reduced GCase activity can result in impaired clearance of alpha-synuclein, leading to its accumulation.[131]

TMEM175, a lysosomal or proton potassium channel protein, has been suggested to regulate the pH level within the lysosome. Alterations in lysosomal pH due to *TMEM175* variants could impact GCase activity and contribute to the pathogenesis of synucleinopathies.[115]

Another gene, *LRRK2*, is known to interact with the ALP. *LRRK2* is thought to be recruited to the lysosome, where it phosphorylates Rab proteins, which are involved in vesicular trafficking.[132] Dysregulation of Rab proteins by *LRRK2* mutations may disrupt lysosomal function and contribute to the accumulation of alpha-synuclein.[133]

Overall, the autophagy-lysosomal pathway is central to the pathophysiology of synucleinopathies, including PD and related disorders such as RBD. Dysfunction in this pathway, including impaired lysosomal degradation and defective mitophagy, can lead to the accumulation of alpha-synuclein and contribute to disease progression.

Mitophagy

Mitophagy is a cellular process that involves the degradation of dysfunctional or damaged mitochondria. [134] It is primarily associated with the Parkin/*PINK1* pathway, although other genes such as *LRRK2* and *SNCA* have also been implicated in its regulation.[135]

In a healthy mitochondrion, PINK1 is imported into the outer mitochondrial membrane and subsequently cleaved and degraded by the proteasome.[136] However, when mitochondria are damaged or depolarized, the import of PINK1 is inhibited, leading to its stabilization on the outer membrane.[136] Stabilized PINK1 then recruits Parkin to the mitochondria, which results in the ubiquitination of various mitochondrial proteins.[136] This ubiquitination serves as a signal for the selective recognition and targeting of damaged mitochondria for degradation via mitophagy.

In PD, alpha-synuclein aggregates have been shown to bind to mitochondria, disrupting mitochondrial function and causing mitochondrial membrane depolarization.[137] This can further impair the normal process of mitophagy. *LRRK2* has complex interactions with the Parkin/*PINK1* pathway and has been implicated in the regulation of mitophagy.[138]

In individuals with PD and mutations in the *GBA1* gene, post-mortem analyses have revealed increased mitochondrial stress and impaired mitophagy.[139] *GBA1* mutations are associated with reduced GCase activity, which can disrupt lysosomal function and impair the clearance of dysfunctional mitochondria through mitophagy.

Overall, mitochondrial dysfunction, disrupted mitophagy, and the interactions between genes such as Parkin/*PINK1*, *LRRK2*, *SNCA*, and *GBA1* play major roles in the pathogenesis of PD. Understanding these processes and their interplay may provide valuable insights into the mechanisms underlying the development and progression of the disease.

Neuroinflammation

Neuroinflammation in PD has been extensively studied and is characterized by the activation of microglia, which are immune cells in the central nervous system. This initial discovery of microglial activation in post-mortem brains of PD patients led to further investigations into the involvement of neuroinflammation in the disease.[140]

Multiple studies have shown increased levels of proinflammatory cytokines in the blood, cerebrospinal fluid, and brain tissue of PD patients.[141] These findings indicate a systemic and central nervous system inflammatory response in PD.

Neuroinflammation has also been observed in animal models of PD and in experimental models where alpha-synuclein is overexpressed or injected.[142] These models show increased microglial activation and production of proinflammatory cytokines, further supporting the role of neuroinflammation in PD.

Genetic evidence has also implicated the immune system in PD. Variants in genes associated with immune function, such as those in the *HLA* locus, *LRRK2*, *PRKN* and potentially *GBA1* have been found to contribute to neuroinflammation in PD.[143] The *HLA* locus is involved in antigen presentation and immune response, while *LRRK2* and *PRKN* have been linked to immune system dysfunction and microglial activation.

The presence of neuroinflammation in PD suggests a complex interplay between the immune system and neurodegeneration. Chronic inflammation and sustained activation of microglia can lead to neuronal damage and contribute to the progression of PD. Understanding the mechanisms underlying neuroinflammation in PD may provide new therapeutic targets for intervention and disease modification.

Clinical interventions in Parkinson's disease

The failure of clinical trials in PD has highlighted the need for improvements in various aspects of trial design and patient selection. Future clinical trials should aim to address these challenges to increase their chances of success.

One important consideration is the timing of intervention. PD is a complex and heterogeneous disease, and the underlying pathophysiological processes may vary among individuals. Identifying the optimal stage of disease progression for intervention is crucial. Early detection and intervention may be more effective in modifying the course of the disease, and clinical trials should aim to recruit patients at the right time to target specific disease mechanisms.

Genetic screening is another consideration for future clinical trials. As our understanding of the genetic risk factors for PD expands, genetic screening prior to recruitment can help identify patients who may benefit from specific interventions or stratify patients into subtypes for more targeted treatments. Genetic screening can also aid in the identification of potential responders and non-responders to specific therapies, enabling more personalized and precise approaches.

Precision medicine approaches, such as targeting specific genetic risk factors like *GBA1* and *LRRK2*, have been conducted in recent clinical trials.[144] These targeted therapies hold potential for more effective and tailored treatments for subsets of PD patients. Incorporating precision medicine principles into future clinical trials can help identify new therapeutic candidates and improve treatment outcomes.

In this thesis, exploring population applications and identifying new genetic candidates for future clinical trials across different chapters will contribute to the advancement of knowledge in this field. By investigating the broader population implications of genetic factors and identifying

potential therapeutic targets, this thesis can contribute to the development of precision medicine approaches in PD.

Hypothesis

I hypothesize that there are additional genetic risk factors in synucleinopathies that can be identified using bioinformatics and machine learning.

Objectives

Chapter 2: Investigating the Role of Heterozygous PRKN Variants in PD. In this chapter, the focus is on examining the risk associated with carrying heterozygous *PRKN* SNVs and CNVs in PD. The study aims to improve the selection criteria for clinical trials by understanding the impact of these variants on disease risk. Additionally, the chapter aims to provide a simple, fast, and cost-effective method for detecting biallelic *PRKN* variants, which can aid in the identification of individuals who may benefit from specific interventions or targeted therapies.

Chapter 3: Fine-mapping of the *HLA* locus in PD. This chapter involves the fine-mapping of the *HLA* locus in PD. Using bioinformatic methods, the study aims to impute *HLA* alleles with high accuracy, providing detailed information on the *HLA* allele, haplotype, or amino acid driving the association observed in the PD GWAS. By better understanding the role of *HLA* in PD, this research contributes to filling the knowledge gap regarding neuroinflammation in the disease. The findings may have implications for the development of new drug therapeutic possibilities targeting the *HLA* genes.

Chapter 4: Fine-mapping of the *HLA* locus in RBD and LBD. Similar to Chapter 3, this chapter focuses on fine-mapping the *HLA* locus, but in the context of two related synucleinopathies: RBD and LBD. The study examines *HLA* alleles, haplotypes, and amino acids across these disorders, aiming to identify any shared or distinct associations. By comparing the findings in *HLA* across these synucleinopathies, the chapter provides insights into the potential role of *HLA* in RBD and LBD, further enhancing our understanding of the immune component in these disorders.

Chapter 5: Gene prioritization of PD GWAS loci. In this chapter, the focus is on gene prioritization of PD GWAS loci. The study employs fine-mapping techniques to nominate candidate genes within each locus using machine learning methods. By integrating genetic, transcriptomic, and epigenetic data from brain tissues and specifically dopaminergic neurons, a machine learning model is trained to prioritize genes for further analysis. The results of this study can potentially identify novel genes and pathways associated with PD, opening avenues for potential drug targets and genetic discoveries in the field.

Overall, these chapters contribute to the understanding of PD genetics, and the identification of potential therapeutic targets. By investigating the role of *PRKN* variants, fine-mapping the *HLA* locus in PD, RBD, and LBD, and prioritizing candidate genes within the PD GWAS loci, these studies aim to advance our knowledge and pave the way for improved treatments and interventions in synucleinopathies.

Preface to Chapter 2

In this chapter, the focus is on exploring the population applications of genetic studies for clinical trials in PD. The controversial role of heterozygous *PRKN* variants in PD is investigated. The chapter begins by conducting systematic sequencing of rare variants and CNVs within the *PRKN* gene. By analyzing these rare variants, the study aims to shed light on the potential impact of heterozygous *PRKN* variants on PD risk.

To strengthen the findings, a meta-analysis of rare variants in *PRKN* is performed, pooling data from multiple studies. The chapter emphasizes the importance of genetic screening in the context of clinical trials. By understanding the genetic profile of individuals enrolled in clinical trials, researchers can improve participant selection criteria and enhance the precision of therapeutic interventions.

Overall, this chapter contributes to our understanding of the controversial role of heterozygous *PRKN* variants in PD and highlights the significance of genetic screening in the design and execution of clinical trials. The findings have implications for the development of personalized treatment approaches and the identification of individuals who may respond favorably to specific interventions based on their genetic profile.

Chapter 2: Analysis of heterozygous *PRKN* variants and copy number variations in Parkinson's disease

Eric Yu, BSc[1,2], Uladzislau Rudakou, BSc[1,2], Lynne Krohn, BSc[1,2], Kheireddin Mufti, BPharm[1,2], Jennifer A. Ruskey, MSc[2,3], Farnaz Asayesh, MSc[2,3], Mehrdad A. Estiar, MSc[1,2], Dan Spiegelman, MSc[2,3], Matthew Surface, BA[4], Stanley Fahn, MD[4], Cheryl H. Waters, MD[4], Lior Greenbaum, MD, PhD[5,6,7], Alberto J. Espay, MD, MSc[8], Yves Dauvilliers, MD, PhD[9], Nicolas Dupré, MD, FRCP[10,11], Guy A. Rouleau, MD, PhD, FRCPC, FRSC[1,2,3], Sharon Hassin-Baer, MD[7,12], Edward A. Fon, MD, FRCPC[2,3], Roy N. Alcalay, MD, MS[4,13] Ziv Gan-Or, MD, PhD[1,2,3].

Affiliations

1. Department of Human Genetics, McGill University, Montréal, Quebec, Canada
2. Montreal Neurological Institute and Hospital, McGill University, Montréal, Quebec, Canada
3. Department of Neurology and Neurosurgery, McGill University, Montréal, Quebec, Canada
4. Department of Neurology, College of Physicians and Surgeons, Columbia University Medical Center, New York, NY 10032, USA
5. The Danek Gertner Institute of Human Genetics, Sheba Medical Center, Tel Hashomer, Ramat Gan, Israel
6. The Joseph Sagol Neuroscience Center, Sheba Medical Center, Tel Hashomer, Ramat Gan, Israel
7. Sackler school of medicine, Tel-Aviv University, Tel-Aviv, Israel
8. UC Gardner Neuroscience Institute and Gardner Family Center for Parkinson's Disease and Movement Disorders, Cincinnati, OH, USA.
9. National Reference Center for Narcolepsy, Sleep Unit, Department of Neurology, Gui-de-Chauliac Hospital, CHU Montpellier, University of Montpellier, Inserm U1061, Montpellier, France
10. Division of Neurosciences, CHU de Québec, Université Laval, Québec City, Quebec, Canada
11. Department of Medicine, Faculty of Medicine, Université Laval, Québec City, Quebec, Canada
12. Movement Disorders Institute, Department of Neurology, Sheba Medical Center, Tel Hashomer, Ramat-Gan, Israel
13. Taub Institute for Research on Alzheimer's Disease and the Aging Brain, College of Physicians and Surgeons, Columbia University Medical Center, New York, NY, USA

Published: *Movement Disorders*. 2021 Jan;36(1):178-187.

Abstract

Background: Biallelic *PRKN* mutation carriers with Parkinson's disease (PD) typically have an earlier disease onset, slow disease progression and, often, different neuropathology compared to sporadic PD patients. However, the role of heterozygous *PRKN* variants in the risk of PD is controversial.

Objectives: We aimed to examine the association between heterozygous *PRKN* variants, including single nucleotide variants and copy-number variations, and PD.

Methods: We fully sequenced *PRKN* in 2,809 PD patients and 3,629 healthy controls, including 1,965 late onset (63.97±7.79 years, 63% men) and 553 early onset PD patients (43.33±6.59 years, 68% men). *PRKN* was sequenced using targeted next-generation sequencing with molecular inversion probes. Copy-number variations were identified using a combination of multiplex ligation-dependent probe amplification and ExomeDepth. To examine whether rare heterozygous single nucleotide variants and copy-number variations in *PRKN* are associated with PD risk and onset, we used optimized sequence kernel association tests and regression models.

Results: We did not find any associations between all types of *PRKN* variants and risk of PD. Pathogenic and likely-pathogenic heterozygous single nucleotide variants and copy-number variations were less common among PD patients (1.52%) than among controls (1.8%, false discovery rate-corrected p=0.55). No associations with age at onset and in stratified analyses were found.

Conclusions: Heterozygous single nucleotide variants and copy-number variations in *PRKN* are not associated with Parkinson's disease. Molecular inversion probes allow for rapid and cost-

effective detection of all types of *PRKN* variants, which may be useful for pre-trial screening and for clinical and basic science studies specifically targeting *PRKN* patients.

Introduction

Parkinson's disease (PD) is a common neurodegenerative disorder with a typical age at onset (AAO) ranging between 60-70 years.[1] However, a subgroup of patients has early onset PD (EOPD), typically defined as AAO < 50 years.[2] The most common genetic cause of EOPD are homozygous or compound heterozygous variants in the *PRKN* gene, found in 6.0-12.4% of individuals who present with PD symptoms before the age of 50.[3-5] *PRKN* has a high rate of single nucleotide variants (SNV) and copy number variations (CNVs), since it is located in a genomic region prone to rearrangements.[6, 7] *PRKN* encodes Parkin, an E3 ubiquitin protein ligase important in mitophagy.[8]

Neuropathological studies have demonstrated that individuals with biallelic *PRKN* variants diagnosed with PD do not have the typical PD neuropathology, as Lewy bodies are absent in most cases, and the neurodegenerative process is limited to the substantia nigra.[9, 10] It is therefore possible that patients with biallelic *PRKN* variants represent a distinct subgroup, or arguably a distinct disease with similar clinical features.[10] Since we are moving towards therapies targeting specific genetic defects in PD (such as *GBA* and *LRRK2*-targeting therapies), or a-synuclein accumulation[11] (which is mostly absent in *PRKN*-related patients),[9] it is crucial to properly identify these patients. However, the role of rare heterozygous *PRKN* SNVs and CNVs in PD has not been clearly established by association studies,[12] and it is currently controversial. For example, a previous study with 159 patients and 170 controls showed significant difference in heterozygous *PRKN* SNVs and CNVs between PD patients and controls, while larger studies suggested a lack of association.[13-15] Additional studies have also shown contradictory results in familial PD, EOPD

and late onset PD (LOPD) using SNVs and/or CNVs.[13-34] Therefore, the role of heterozygous *PRKN* variants remains controversial. Towards future clinical trials targeting *PRKN*, it will be crucial to determine whether heterozygous *PRKN* variants are associated with PD.

To investigate the potential effect of rare heterozygous SNVs and CNVs in PD, we applied a simple, fast and cost-effective method to detect both types of variants. Using targeted next generation sequencing and bioinformatic approaches, we fully sequenced *PRKN* to identify both SNVs and CNVs in a large cohort of PD, including LOPD and EOPD.

Methods

Study Population

A total of 2,809 unrelated and consecutively recruited PD patients and 3,629 controls from three cohorts were sequenced, including 1,965 LOPD patients (mean [SD], 63.97±7.79 years, 1,231 men [63%]) and 553 EOPD patients (mean [SD], 43.33±6.59, 374 men [68%]). Age and sex were not available for 291 patients, 88 controls and 22 patients, 4 controls, respectively. After excluding low sequencing quality samples and biallelic *PRKN* carriers, we performed statistical analysis on 6,090 individuals: 2,627 patients and 3,463 controls. The three cohorts are detailed in Table 1 and include: a) a cohort of European ancestry, confirmed by principal component analysis, collected at McGill University, including French-Canadian (mostly recruited through the Quebec Parkinson Network)[35] and French participants recruited in Quebec, Canada and Montpellier, France b) a cohort recruited at Columbia University, New York, as previously described,[36] primarily composed of individuals of self-reported European origin and Ashkenazi Jews, and c) a cohort collected at the Sheba Medical Center, Israel, of self-reported Ashkenazi Jewish ancestry, as previously described.[37] PD was diagnosed by movement disorder specialists according to the UK Brain Bank Criteria, without excluding patients with positive family history [38] or the Movement

Disorders Society Criteria.[39] Study protocols were approved by the relevant Institutional Review Boards and all patients signed informed consent before participating in the study.

Table 1. Study populations included in the analysis after quality control and exclusion of biallelic pathogenic/likely pathogenic *PRKN* SNV and CNV carriers.

	McGill University		Columbia University		Sheba Medical Center	
Variable	Patients (n=1,034)	Controls (n=2,451)	Patients (n=939)	Controls (n=491)	Patients (n=654)	Controls (n=521)
Age, y[a,b] (SD)	59.09 (10.58)	53.35 (14.17)	59.34 (11.56)	64.56 (9.87)	60.50 (11.82)	39.24 (14.38)
Early onset patients[c]/total with available AAO data, No. (%)	164/747 (22%)	NA	212/929 (23%)	NA	122/632 (19%)	NA
Male, No. (%)	654 (64%)	1,160 (47%)	611 (65%)	177 (36%)	408 (63%)	306 (59%)
Ashkenazi Jewish ancestry, No. (%)	0 (-)	0 (-)	210 (22%)	91 (19%)	654 (100%)	521 (100%)

[a] Data are presented as mean (SD).

[b] Data for age and sex are missing for 321 (12%) patients and 246 (7%).

[c] Early onset is defined as AAO < 50 years controls. Difference of age and sex between patients and controls were adjusted in our analysis.

Abbreviations: y, years; SD, standard error; AAO, age at onset; No, number; NA, not applicable;

Genetic analysis

***PRKN* sequencing**

All samples were sequenced at McGill University, Canada using the same method. A total of 50 genes were captured using molecular inversion probes (MIPs) and sequenced as previously described.[40] In brief, probes that specifically target the coding sequences of the genes of interest were designed, followed by capture and PCR amplification of the targeted regions. After adding barcodes, samples were pooled and sequenced at the McGill University and Génome Québec

Innovation Centre with Illumina HiSeq 2500/4000. The full protocol is available upon request. Alignment (GRCh37/hg19), quality control and variant calls were done using the Burrows-Wheeler Aligner (BWA),[41] Genome Analysis Toolkit (GATK v3.8),[42] and ANNOVAR [43] as previously described.[44] Only rare variants (minor allele frequency, MAF, < 0.01) according to the public database Genome Aggregation Database (GnomAD) [45] with a minimum coverage of 30x were included in the analysis. Samples with more than 10% missingness were excluded. The script for these analyses can be found at https://github.com/gan-orlab/MIPVar. We examined all rare exonic variants using the Integrative Genomics Viewer (IGV v 2.7).[46] All variants were classified using Varsome [47] according to the American College of Medical Genetics and Genomics (ACMG) standards and guidelines into five categories: pathogenic, likely pathogenic, uncertain significance, likely benign and benign.

Detection and validation of copy number variations

There are four general types of methods to infer CNVs from next-generation sequencing.[48] Because MIPs target only a small portion of the genome, most CNV breakpoints will not be sequenced. Therefore, only read-depth based methods can be applied for MIPs since other types of methods utilize reads that span breakpoints. In order to detect CNVs, we examined two methods based on read depth for the MIP data, ExomeDepth v1.1.10 [49] and panelcn.MOPS v1.4.0 in R.[50] When using ExomeDepth, each test sample is compared to the best set of reference samples out of 3,629 controls, chosen by the software according to the correlation of the coverage for each probe between the test sample and the reference samples. A filter for samples with correlation above 0.97 per the suggestion of the developer was applied to remove false positives. Panelcn.MOPS also selects the best set of reference samples according to correlation and includes several quality control (QC) steps, such as a minimum user defined depth of coverage per probe. Probes are

marked as low quality if their read count shows high variance across the test sample and selected references. To validate CNVs, we performed multiplex ligation-dependent probe amplification (MLPA) using the SALSA MLPA P051-D2 Parkinson probemix 1 kit according to the manufacturer instructions (MRC Holland), which is the gold standard for *PRKN* CNV detection.

Quality Control of MIPs for CNV detection

The highest performing parameters were achieved by excluding probes from genes in our library where the average coverage was below 100X in more than 15% of the coding and untranslated regions of the genes. Probes with average coverage below 100X, and samples with average coverage across all genes less than 50X were also excluded. Figure 1 details the numbers of patients and controls in each cohort after different stages of quality control.

Statistical Analysis

The associations between rare heterozygous SNVs (MAF < 0.01), heterozygous CNVs and PD were tested using optimized sequence kernel association tests (SKAT-O v1.3.2 in R)[51] in all cohorts separately, adjusted for age, sex and ancestry as needed. The initial analysis was performed after excluding biallelic carriers of pathogenic and likely pathogenic mutations and adjusting for age, sex, ethnicity and the presence of *GBA* and *LRRK2* variants (Figure 1, yellow). Rare variants were grouped by: a) CADD score (CADD>12.37), which represent the top 2% of variants predicted to be deleterious, b) functional variants, which include stop gain, nonsynonymous, splice-site and frameshift variants, c) nonsynonymous variants, and d) loss-of-function variants, which include frameshift, splice-site and stop gain variants. A meta-analysis of the results from the three cohorts was performed using MetaSKAT (MetaSKAT v0.80, R)[52] for heterozygous SNVs, CNVs, and both combined, according to the five ACMG categories (pathogenic, likely pathogenic, uncertain significance, likely benign and benign). Since the age- and sex-adjusted

model removes samples without available data on age and sex, we also performed an unadjusted model to avoid this exclusion (Figure 1, blue). We have also repeated all analyses after several additional filtering and adjusting stages, including: adjusting for all *GBA* and *LRRK2* p.Gly2019Ser variant carriers (Supplementary Table 1), excluding these *GBA* and *LRRK2* variant carriers, analyzing only samples with early onset PD (defined as AAO < 50 years), and excluding samples with CNVs in which phasing was not possible (n=8, for example, in a sample with a reported deletion of exons 3-4 the deletion could be on the same allele, or each exon can be deleted on a different allele, Figure 1, grey). The association between heterozygous SNVs, CNVs and AAO of Parkinson's disease was also calculated using linear regression adjusted for sex and ancestry as needed in all cohorts separately. Here too, patients carrying *GBA* variants or the *LRRK2* p.Gly2019Ser variant (Supplementary Table 1) were excluded and all analyses were repeated. METAL [53] was used to performed fixed-effect meta-analysis on all cohorts in the AAO analysis. Since we have performed multiple interdependent analyses, we used a false discovery rate (FDR) correction for multiple comparisons with a FDR-corrected $q<0.05$ considered as statistically significant.

Figure 1. Flow chart of different analysis phases. The flow chart detail the total numbers of patients and controls included in different phases of the analysis. In red, the total number of samples sequenced. In green, the total numbers of samples which passed the quality control phase. In blue, the total numbers of samples after exclusion of 9 patients with biallelic pathogenic and likely pathogenic mutations in *PRKN*. In yellow, the total number of samples included in the analysis aadjusted for age, sex, ethnicity and the presence of *GBA* and *LRRK2* variants. In grey, the total number of samples included in the analysis after excluding additional samples with potentially pathoigenic biallelic copy number variations that could not be phased, i.e. samples with

deletions of consecutive exons, for which we could not determine if they occur on the same allele or if they are biallelic.

Results

Identification of *PRKN* SNVs and CNVs

The average coverage of *PRKN* (NM_004562) across all samples was 988X, with 98% of nucleotides covered at >30X, and 94% covered at >100X. We identified 199 rare SNVs in 237 patients and 300 controls in the main analysis (Table 2), including nonsynonymous, frameshift deletions and splice site variants in *PRKN* across all cohorts (the specific variants are detailed in Supplementary Table 2).

To identify CNVs, we first aimed to examine which calling method is best suited to properly call CNVs from our MIP targeted sequencing panel. For this purpose, we screened for CNVs in 510 samples using MLPA, the gold standard for CNV detection in *PRKN*. We specifically enriched these samples with EOPD patients to increase the chances to detect CNVs. Out of the 510 samples, 46 carried CNVs in *PRKN* (32 patients and 14 controls). The 32 patients included four homozygous *PRKN* deletion carriers, 17 heterozygous deletion carriers and 11 duplication carriers. Subsequently, we have examined which method (ExomeDepth or panel.cnMOPS) has the highest performance. Except for one deletion for which the MIPs data did not pass QC due to low coverage call rate, deletions and duplications in *PRKN* were identified with 97% sensitivity and 95% specificity using ExomeDepth. In contrast, using the best parameters, panel.cnMOPS had 98% sensitivity but only 54% specificity using samples that passed QC when compared to MLPA. The parameters and CNV call rates for each method are detailed in Supplementary Table 3. Due to its superior performance, we applied ExomeDepth on all cohorts, and identified a total of 62 carriers

of CNVs in patients and controls. Supplementary Table 4 details all carriers of CNVs, including heterozygous and bi-allelic carriers of other CNVs or other SNVs.

Heterozygous *PRKN* SNVs and CNVs are not associated with Parkinson's disease

To examine the association of rare (MAF < 0.01) heterozygous SNVs and CNVs on risk of PD, we took two approaches. First, we performed a SKAT-O in each cohort to determine whether there is a burden of heterozygous *PRKN* variants of different types. "Pathogenic" variants included pathogenic and likely pathogenic variants, while "non-benign" variants included pathogenic, likely pathogenic and variants of uncertain significance. All CNVs were considered as pathogenic loss-of-function variations. No statistically significant associations were found in any of the SKAT-O analyses (Table 2). Second, we performed a series of meta-analyses by collapsing in each cohort SNVs alone, CNVs alone, and combined. In these analyses too, adjusted for age, sex and ethnicity, no association between heterozygous carriage of *PRKN* mutations and PD was found (Table 2). Pathogenic and likely pathogenic variants were less frequent in patients (1.52%) than in controls (1.8%, $p = 0.55$, Table 2), suggesting lack of association with risk of PD. In order to avoid the possibility that the exclusion of samples without available data on age and sex had biased the results, we have also performed an unadjusted analysis including all samples. Additional analyses with and without *GBA* and *LRRK2* p.Gly2019Ser variants, with and without CNVs of unknown phasing, and including only samples patients with AAO < 50 have also been performed. In these analyses too, there were no statistically significant differences between patients and controls (Supplementary Table 5-6).

Table 2. Rare *PRKN* heterozygous SNV and CNV analysis for risk of Parkinson's disease using SKAT-O adjusted for age, sex, ethnicity and *GBA* and *LRRK2* status.

	Carriers in patients, No. (%)	Carriers in controls, No. (%)	*P*-value[a]
McGill University (n = 2964)			
SNV	91 (12.2)	230 (10.4)	0.594
CNV	4 (0.537)	15 (0.676)	0.608
SNV & CNV	93 (12.5)	238 (10.7)	0.621
Patho SNV	11 (1.48)	22 (0.991)	0.675
Patho SNV & CNV	15 (2.01)	37 (1.67)	0.675
No Benign SNV	67 (8.99)	152 (6.85)	0.507
No Benign SNV & CNV	69 (9.26)	160 (7.21)	0.552
Columbia University (n = 1,420)			
SNV	101 (10.9)	45 (9.16)	0.614
CNV	13 (1.4)	11 (2.24)	0.507
SNV & CNV	109 (11.7)	52 (10.6)	0.588
Patho SNV	5 (0.538)	5 (1.02)	0.507
Patho SNV & CNV	18 (1.94)	16 (3.26)	0.502
No Benign SNV	76 (8.18)	31 (6.31)	0.507
No Benign SNV & CNV	85 (9.15)	38 (7.74)	0.507
Sheba Medical Center (n = 1,139)			
SNV	45 (7.12)	25 (4.93)	0.507
CNV	2 (0.316)	4 (0.789)	0.584
SNV & CNV	46 (7.28)	27 (5.33)	0.507
Patho SNV	0 (0)	1 (0.197)	0.679
Patho SNV & CNV	2 (0.316)	5 (0.986)	0.584
No Benign SNV	42 (6.65)	23 (4.54)	0.502
No Benign SNV & CNV	43 (6.8)	25 (4.93)	0.507
Meta-Analysis (n = 5,523)			
SNV	237 (10.3)	300 (9.35)	0.507
CNV	19 (0.824)	30 (0.933)	0.507
SNV & CNV	248 (10.7)	317 (9.84)	0.507
Patho SNV	16 (0.695)	28 (0.87)	0.646
Patho SNV & CNV	35 (1.52)	58 (1.8)	0.553
No Benign SNV	185 (8.02)	206 (6.4)	0.461
No Benign SNV & CNV	197 (8.54)	223 (6.93)	0.461

Abbreviations: SNV—Single Nucleotide Variant; CNV—Copy Number Variation; Patho—pathogenic and likely pathogenic variants; No Benign—analysis excluding benign and likely benign variants.

[a] P-values shown are after FDR correction (q value <= 0.05).

Heterozygous *PRKN* SNVs and CNVs are not associated with AAO of Parkinson's disease

The association between rare heterozygous SNVs and CNVs on AAO of PD was examined using linear regression in each cohort alone on the same groups of mutations mentioned in the previous association study. After adjusting for sex, ancestry, and the presence of *GBA* and *LRRK2* variants, we found no association in any analyses. We also performed meta-analysis by collapsing each cohort which yielded no statistically significant results (Table 3). When examining CNVs, the meta-analysis shows an earlier AAO in heterozygous *PRKN* CNV carriers (3.6 years younger compared to non-carriers), but the association was not statistically significant after correction for multiple comparisons. This difference in AAO was mainly driven by an effect of CNVs in the Columbia cohort, which was almost 8 years younger in carriers of CNVs (average AAO of 51.85 years) compared to non-carriers of CNVs (59.44 years). This difference was not statistically significant after correction for multiple comparisons as well. Larger studies for AAO of heterozygous *PRKN* carriers are needed to further study these findings. Association analyses between different types of heterozygous *PRKN* variants and AAO of PD, including with and without *LRRK2* and *GBA* variant carriers, with and without ambiguous phasing (see methods), and in AAO < 50 can be found in Supplementary Tables 7-8. In all analyses, there were no statistically significant associations.

Table 3. Rare *PRKN* heterozygous SNV and CNV analysis for Age at Onset of Parkinson's disease adjusted age, sex, ethnicity and *GBA* and *LRRK2* status.

	AAO of carriers, y mean (SD)	AAO of non-carriers, y mean (SD)	Coeff (95% CI)	*P*-value[a]
McGill University (n=745)				
SNV	61.04 (10.54)	58.83 (10.56)	2.19 (-0.127,4.5)	0.49
CNV	61.75 (14.31)	59.09 (10.56)	1.87 (-8.57,12.3)	0.91
SNV & CNV	60.96 (10.71)	58.84 (10.54)	2.07 (-0.217,4.37)	0.52
Patho SNV	57.73 (12.96)	59.12 (10.54)	-1.27 (-7.57,5.02)	0.91
Patho SNV & CNV	58.8 (12.94)	59.11 (10.53)	-0.44 (-5.85,4.97)	0.96
No Benign SNV	61.04 (10.29)	58.91 (10.59)	2.08 (-0.565,4.73)	0.55
No Benign SNV & CNV	60.93 (10.52)	58.91 (10.57)	1.94 (-0.678,4.55)	0.60
Columbia University (n=929)				
SNV	59.2 (12.82)	59.35 (11.4)	0.22 (-2.15,2.59)	0.96
CNV	51.85 (14.01)	59.44 (11.49)	-6.68 (-12.9,-0.411)	0.43
SNV & CNV	58.64 (13.02)	59.43 (11.36)	-0.371 (-2.67,1.93)	0.92
Patho SNV	60.4 (18.88)	59.33 (11.52)	1.21 (-8.87,11.3)	0.96
Patho SNV & CNV	54.22 (15.43)	59.44 (11.46)	-4.5 (-9.84,0.84)	0.52
No Benign SNV	61.74 (11.99)	59.12 (11.5)	2.81 (0.126,5.49)	0.43
No Benign SNV & CNV	60.51 (12.67)	59.22 (11.44)	1.56 (-0.994,4.11)	0.67
Sheba Medical Center (n=632)				
SNV	55.84 (14.12)	60.86 (11.57)	-5.02 (-8.59,-1.46)	0.20
CNV	71.5 (3.536)	60.47 (11.83)	11.3 (-5.13,27.7)	0.62
SNV & CNV	56.13 (14.1)	60.85 (11.57)	-4.72 (-8.25,-1.18)	0.20
Patho SNV	NA	60.5 (11.82)	NA	0.84
Patho SNV & CNV	71.5 (3.536)	60.47 (11.83)	11.3 (-5.13,27.7)	0.62
No Benign SNV	56.12 (14.27)	60.82 (11.58)	-4.69 (-8.38,-1)	0.21
No Benign SNV & CNV	56.42 (14.23)	60.8 (11.59)	-4.37 (-8.02,-0.721)	0.25
Meta-Analysis (n=2,306)				
SNV	59.17 (12.44)	59.62 (11.21)	0.122 (-1.41, 1.65)	0.96
CNV	56 (14.57)	59.61 (11.3)	-2.89 (-8.1, 2.32)	0.71
SNV & CNV	58.72 (12.69)	59.65 (11.19)	-0.122 (-1.63, 1.38)	0.96
Patho SNV	58.56 (14.45)	59.59 (11.31)	0.506 (-1.33, 2.34)	0.86
Patho SNV & CNV	57.17 (14.36)	59.62 (11.28)	-1.8 (-5.57, 1.98)	0.78
No Benign SNV	60.07 (12)	59.53 (11.26)	0.961 (-0.751, 2.67)	0.71
No Benign SNV & CNV	59.4 (12.4)	59.56 (11.23)	0.523 (-1.14, 2.19)	0.84

Abbreviations: SNV—Single Nucleotide Variant; CNV—Copy Number Variation; AAO—Age At Onset; Coeff—Regression Coefficient; CI—Confidence Interval; Patho—pathogenic and likely pathogenic variants; No Benign—analysis excluding benign and likely benign variants. NA—Not Applicable.

[a] P-values shown are after FDR correction (q value <= 0.05).

Table 4: Parkinson's disease patients with biallelic pathogenic and likely pathogenic *PRKN* SNVs and CNVs.

Sample	SNV	CNV	Sex	AAS	AAO
McGill cohort					
S29243	p.Gly430Asp	Exon 9 duplication	M	38	28
S21069	p.Gln34ArgfsTer5, p.Arg42Pro		F	73	29
S06128	p.Gln34ArgfsTer5	Exon 3, 4 deletion	M	44	39
S29254		Homozygous exon 3 deletion	M	57	NA
S29263	p.Arg275Trp	Exon 3 deletion	F	41	NA
Columbia cohort					
S23082		Homozygous exon 3, 4 deletion	F	54	16
S22919	p.Gln34ArgfsTer5	Exon 3, 4 deletion	M	52	19
S23324	p.Arg275Trp	Exon 3 deletion	F	70	28
S33346		Homozygous exon 3, 4 deletion	M	37	37

Sample are sorted by age at onset with one SNV or CNV per locus at each line.

Abbreviations: SNV—Single nucleotide Variants; CNV—Copy Number Variation; AAS—Age At Sample; AAO—Age At Onset; M—Male; F—Female; NA—Not Available.

Identification of *PRKN*-associated parkinsonism patients

Overall, we were able to identify 9 patients with pathogenic or likely pathogenic homozygous and compound heterozygous *PRKN* SNVs and/or CNVs (Table 4). The most common pathogenic SNV in our cohort was p.Gln34ArgfsTer5 mutation, found in 3 (33%) *PRKN* patients, and the most

common CNV was heterozygous deletion of exon 3, found in 7 (77%) *PRKN* patients. The average AAO of PD in biallelic *PRKN* SNV/CNV carriers was 28.0 ±7.82 years old.

Discussion

In the current study, we found that the frequencies of heterozygous SNVs and CNVs in *PRKN* are similar in PD patients and controls. These results do not support a role for heterozygous *PRKN* variants in the risk of PD or its AAO. Of note, in one cohort (Columbia), the average AAO of CNV carriers was about 8 years younger compared to non-carriers (Table 3), yet in the other cohorts there was no difference between CNV carriers and non-carriers. Additional studies on AAO in heterozygous *PRKN* carriers are required to conclusively determine whether or not they are associated with earlier AAO. Since the *PRKN* region is prone to genetic variance,[6] including multiple SNVs and CNVs, properly genotyping all types of *PRKN* variants could be challenging. Using a simple, fast and cost-effective method, we were able to successfully detect all CNVs, SNVs and indels. With MIPs, deep coverage can be achieved, and the probes always target the exact same region, as opposed to whole-exome or whole-genome sequencing where there is no full overlap between all the reads. When the coverage is high, it provides an advantage that allows for more accurate calls of CNVs as well as SNVs and indels. Using this approach, we have identified 199 rare *PRKN* variants and 62 participants with *PRKN* CNVs, with very high sensitivity and specificity (97% and 95%, respectively, when compared to the gold standard MLPA method). Our approach can therefore be used for large-scale screening of PD cohorts, with only validation of detected *PRKN* CNVs with MLPA, instead of fully screening all patients with MLPA. Of note, we identified 9 patients with pathogenic and likely pathogenic biallelic *PRKN* variants. This number of patients is lower than previously reported in EOPD. It is possible that in Ashkenazi Jewish Parkinson's disease patients (comprising the entire Sheba cohort and a large portion of the

Columbia cohort), the frequency of *PRKN* variants is lower, as evident by the lack of such patients in the Sheba cohort. This is also supported by the Columbia cohort, in which all biallelic *PRKN* patients are of European ancestry and none among the Ashkenazi Jewish origin.

There have been multiple studies analyzing the role of heterozygous *PRKN* mutations with conflicting results, shown in Supplementary Table 9. These conflicts may arise from different screening approaches. Some studies first sequenced all patients for rare SNVs and/or CNVs, then sequenced only for selected variants in controls. This approach will create a bias, as the controls may carry other pathogenic *PRKN* variants. Other studies sequenced all patients and controls for heterozygous SNVs and/or CNVs more systematically, and the majority of them were negative. Systematic analysis, as was done in the current study, will avoid misrepresenting the genetic landscape of the study population. Our results do not support an association between heterozygous SNVs and CNVs in *PRKN* and PD, which is supported by other systematic studies of *PRKN* as shown in Supplementary Table 9.[14-17, 24, 33] These results also emphasize the need for determining the pathogenicity of different *PRKN* variants, as many variants are currently defined as variants of unknown significance. Having a reliable assay for Parkin activity, as previously suggested, would provide an experimental way to assess pathogenicity of *PRKN* variants.[54]

To further study the potential effect of heterozygous *PRKN* variants, previous studies have compared the rate of 18F-dopa uptake in biallelic *PRKN* patients, asymptomatic heterozygous *PRKN* mutation carriers and healthy controls.[55, 56] These studies have suggested that some *PRKN* heterozygous carriers may have reduced uptake of ^{18}F-dopa, especially in the caudate and putamen. A follow-up longitudinal study by one of these groups, however, suggested that this reduction is subclinical, and that the rate of progression is very slow and unlikely to lead to clinical parkinsonism manifestations.[57]

In recent years, treatments that target specific genes and proteins implicated by human genetic studies, such as *SNCA* (a-synuclein), *GBA* and *LRRK2*, are being tested in clinical trials.[58] Therefore, identifying patients that may benefit from these trials, or conversely, patients that are less likely to benefit, is crucial. Neuropathological studies on brains of patients with *PRKN*-associated parkinsonism have demonstrated that the vast majority of patients with biallelic *PRKN* mutations do not have accumulation of a-synuclein and the typical Lewy bodies that are seen in PD.[59] Since a-synuclein does not accumulate, it is likely that treatment targeting a-synuclein will not be efficient for these patients, who should therefore be excluded from these clinical trials. Furthermore, the neurodegenerative process in *PRKN*-associated Parkinsonism is limited to the substantia nigra and locus coeruleus, and does not spread to other brain regions.[60] Since we did not detect an association between heterozygous *PRKN* variants and PD, we recommend that heterozygous carriers of *PRKN* variants should not be excluded from such trials, as it is likely that the presence of heterozygous *PRKN* variants in PD patients is due to chance. Clinically, patients with *PRKN*-associated Parkinsonism are also different, as they have early onset disease, slowly progressing and typically without or with very limited non-motor symptoms.[59] Therefore, it is important to identify these patients, and our method for rapid and cost-effective detection of *PRKN* variants would be useful for pre-trial screening and for clinical and basic science studies specifically targeting *PRKN* patients.

Although this study examined heterozygous mutations systematically, there are several limitations. The error rate of ExomeDepth CNV detection could affect the results of the association study because not all samples were analysed using MLPA. Furthermore, potentially pathogenic intronic variants have not been examined since intronic regions were not sequenced. In addition, our cohorts were not matched for age and sex. Our controls are on average younger and our patients

are predominantly composed of men, yet age and sex were adjusted for when possible. The missing age at onset of patients underpowers our AAO study, however, because data were missing at random, its effect on our results is likely minimal. Another limitation is that in a case-control set-up, phasing cannot be performed, and patients with two variants are considered as compound heterozygous carriers. Since all patients with two mutations had AAO<50, it is likely that indeed they are all compound heterozygous, but we cannot rule out that they carry two variants on the same allele. In addition, individuals with CNVs in consecutive exons are considered as heterozygous carriers, while in fact they can have separate deletions of each exon in different alleles. To examine whether inclusion of these patients affected the results, we repeated the analysis after excluding them, which did not substantially change the results (Supplementary Tables 5-6). An additional limitation of our study is that it includes predominantly individuals of European and Ashkenazi Jewish ancestries. While we adjusted for ancestry in the analysis, studies in additional ancestries are required to determine if heterozygous *PRKN* variants may have a role in PD in other populations.

To conclude, our findings do not support a role for heterozygous *PRKN* variants in PD, and additional large-scale studies are required for a definite conclusion. Our study and the methods we have used provide a framework and a cost-effective method for rapidly screening for all types of *PRKN* variants, which will be useful in future genetic and clinical studies, and for stratification or patient selection for clinical trials.

References

1. Pagano G, Ferrara N, Brooks DJ, Pavese N. Age at onset and Parkinson disease phenotype. Neurology 2016;86(15):1400-1407.

2. Schrag A, Quinn N, Ben-Shlomo Y. Heterogeneity of Parkinson's disease. Journal of Neurology, Neurosurgery & Psychiatry 2006;77(2):275-276.

3. Bonifati V. Autosomal recessive parkinsonism. Parkinsonism Relat Disord 2012;18:S4-S6.

4. Kilarski LL, Pearson JP, Newsway V, et al. Systematic review and UK-based study of PARK2 (parkin), PINK1, PARK7 (DJ-1) and LRRK2 in early-onset Parkinson's disease. Mov Disord 2012;27(12):1522-1529.

5. Alcalay RN, Caccappolo E, Mejia-Santana H, et al. Frequency of known mutations in early-onset Parkinson disease: implication for genetic counseling: the consortium on risk for early onset Parkinson disease study. Arch Neurol 2010;67(9):1116-1122.

6. Ambroziak W, Koziorowski D, Duszyc K, et al. Genomic instability in the PARK2 locus is associated with Parkinson's disease. J Appl Genet 2015;56(4):451-461.

7. La Cognata V, Morello G, D'Agata V, Cavallaro S. Copy number variability in Parkinson's disease: assembling the puzzle through a systems biology approach. Hum Genet 2017;136(1):13-37.

8. Brüggemann N, Klein C. Parkin type of early-onset Parkinson disease. GeneReviews®[Internet]: University of Washington, Seattle, 2013.

9. Mata IF, Lockhart PJ, Farrer MJ. Parkin genetics: one model for Parkinson's disease. Hum Mol Genet 2004;13(suppl_1):R127-R133.

10. Schneider SA, Alcalay RN. Neuropathology of genetic synucleinopathies with parkinsonism: Review of the literature. Mov Disord 2017;32(11):1504-1523.

11. Sardi SP, Cedarbaum JM, Brundin P. Targeted therapies for Parkinson's disease: from genetics to the clinic. Mov Disord 2018;33(5):684-696.

12. Nalls MA, Blauwendraat C, Vallerga CL, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. The Lancet Neurology 2019;18(12):1091-1102.

13. Lesage S, Lohmann E, Tison F, et al. Rare heterozygous parkin variants in French early-onset Parkinson disease patients and controls. J Med Genet 2008;45(1):43-46.

14. Kay D, Stevens C, Hamza T, et al. A comprehensive analysis of deletions, multiplications, and copy number variations in PARK2. Neurology 2010;75(13):1189-1194.

15. Kay DM, Moran D, Moses L, et al. Heterozygous parkin point mutations are as common in control subjects as in Parkinson's patients. Ann Neurol 2007;61(1):47-54.

16. Bandrés-Ciga S, Price TR, Barrero FJ, et al. Genome-wide assessment of Parkinson's disease in a Southern Spanish population. Neurobiol Aging 2016;45:213. e213-213. e219.

17. Benitez BA, Davis AA, Jin SC, et al. Resequencing analysis of five Mendelian genes and the top genes from genome-wide association studies in Parkinson's Disease. Mol Neurodegener 2016;11(1):29.

18. Bras J, Guerreiro R, Ribeiro M, et al. Analysis of Parkinson disease patients from Portugal for mutations in SNCA, PRKN, PINK1 and LRRK2. BMC Neurol 2008;8(1):1.

19. Brooks J, Ding J, Simon-Sanchez J, Paisan-Ruiz C, Singleton A, Scholz S. Parkin and PINK1 mutations in early-onset Parkinson's disease: comprehensive screening in publicly available cases and control. J Med Genet 2009;46(6):375-381.

20. Camargos ST, Dornas LO, Momeni P, et al. Familial Parkinsonism and early onset Parkinson's disease in a Brazilian movement disorders clinic: Phenotypic characterization and frequency of SNCA, PRKN, PINK1, and LRRK2 mutations. Movement Disorders 2009;24(5):662-666.

21. Clark LN, Afridi S, Karlins E, et al. Case-control study of the parkin gene in early-onset Parkinson disease. Arch Neurol 2006;63(4):548-552.

22. Fiala O, Zahorakova D, Pospisilova L, et al. Parkin (PARK 2) mutations are rare in Czech patients with early-onset Parkinson's disease. PLoS One 2014;9(9).

23. Klein C, Djarmati A, Hedrich K, et al. PINK1, Parkin, and DJ-1 mutations in Italian patients with early-onset parkinsonism. European journal of human genetics 2005;13(9):1086-1093.

24. Lincoln SJ, Maraganore DM, Lesnick TG, et al. Parkin variants in North American Parkinson's disease: cases and controls. Movement disorders: official journal of the Movement Disorder Society 2003;18(11):1306-1311.

25. Macedo MG, Verbaan D, Fang Y, et al. Genotypic and phenotypic characteristics of Dutch patients with early onset Parkinson's disease. Movement disorders: official journal of the Movement Disorder Society 2009;24(2):196-203.

26. Moura KCV, Campos Junior M, de Rosso ALZ, et al. Genetic analysis of PARK2 and PINK1 genes in Brazilian patients with early-onset Parkinson's disease. Dis Markers 2013;35(3):181-185.

27. Sironi F, Primignani P, Zini M, et al. Parkin analysis in early onset Parkinson's disease. Parkinsonism Relat Disord 2008;14(4):326-333.

28. Spataro N, Roca-Umbert A, Cervera-Carles L, et al. Detection of genomic rearrangements from targeted resequencing data in Parkinson's disease patients. Movement Disorders 2017;32(1):165-169.

29. Wang Y, Clark LN, Louis ED, et al. Risk of Parkinson disease in carriers of parkin mutations: estimation using the kin-cohort method. Arch Neurol 2008;65(4):467-474.

30. Camacho JLG, Jaramillo NM, Gómez PY, et al. High frequency of Parkin exon rearrangements in Mexican-mestizo patients with early-onset Parkinson's disease. Movement disorders 2012;27(8):1047-1051.

31. Huttenlocher J, Stefansson H, Steinberg S, et al. Heterozygote carriers for CNVs in PARK2 are at increased risk of Parkinson's disease. Hum Mol Genet 2015;24(19):5637-5643.

32. Pankratz N, Dumitriu A, Hetrick KN, et al. Copy number variation in familial Parkinson disease. PLoS One 2011;6(8).

33. Simon-Sanchez J, Scholz S, del Mar Matarin M, et al. Genomewide SNP assay reveals mutations underlying Parkinson disease. Hum Mutat 2008;29(2):315-322.

34. Hopfner F, Mueller SH, Szymczak S, et al. Private variants in PRKN are associated with late-onset Parkinson's disease. Parkinsonism Relat Disord 2020;75:24-26.

35. Gan-Or Z, Rao T, Leveille E, et al. The Quebec Parkinson Network: A Researcher-Patient Matching Platform and Multimodal Biorepository. J Parkinsons Dis 2020;10:301-313.

36. Alcalay RN, Levy OA, Waters CC, et al. Glucocerebrosidase activity in Parkinson's disease with and without GBA mutations. Brain 2015;138(Pt 9):2648-2658.

37. Ruskey JA, Greenbaum L, Ronciere L, et al. Increased yield of full GBA sequencing in Ashkenazi Jews with Parkinson's disease. Eur J Med Genet 2019;62(1):65-69.

38. Hughes AJ, Daniel SE, Kilford L, Lees AJ. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. Journal of Neurology, Neurosurgery & Psychiatry 1992;55(3):181-184.

39. Postuma RB, Berg D, Stern M, et al. MDS clinical diagnostic criteria for Parkinson's disease. Mov Disord 2015;30(12):1591-1601.

40. Ross JP, Dupre N, Dauvilliers Y, et al. Analysis of DNAJC13 mutations in French-Canadian/French cohort of Parkinson's disease. Neurobiol Aging 2016;45:212. e213-212. e217.

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

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

43. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38(16):e164-e164.

44. Rudakou U, Ruskey JA, Krohn L, et al. Analysis of common and rare VPS13C variants in late-onset Parkinson disease. Neurol Genet 2020;6(1):385.

45. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016;536(7616):285-291.

46. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. Nat Biotechnol 2011;29(1):24.

47. Kopanos C, Tsiolkas V, Kouris A, et al. VarSome: the human genomic variant search engine. Bioinformatics 2018;35(11):1978-1980.

48. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics 2013;14(11):S1.

49. Plagnol V, Curtis J, Epstein M, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics 2012;28(21):2747-2754.

50. Povysil G, Tzika A, Vogt J, et al. panelcn. MOPS: Copy-number detection in targeted NGS panel data for clinical diagnostics. Hum Mutat 2017;38(7):889-897.

51. Lee S, Emond MJ, Bamshad MJ, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet 2012;91(2):224-237.

52. Lee S, Teslovich TM, Boehnke M, Lin X. General framework for meta-analysis of rare variants in sequencing association studies. Am J Hum Genet 2013;93(1):42-53.

53. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 2010;26(17):2190-2191.

54. Yi W, MacDougall EJ, Tang MY, et al. The landscape of Parkin variants reveals pathogenic mechanisms and therapeutic targets in Parkinson's disease. Hum Mol Genet 2019;28(17):2811-2825.

55. Pavese N, Moore RY, Scherfler C, et al. In vivo assessment of brain monoamine systems in parkin gene carriers: a PET study. Exp Neurol 2010;222(1):120-124.

56. Khan NL, Scherfler C, Graham E, et al. Dopaminergic dysfunction in unrelated, asymptomatic carriers of a single parkin mutation. Neurology 2005;64(1):134-136.

57. Pavese N, Khan NL, Scherfler C, et al. Nigrostriatal dysfunction in homozygous and heterozygous parkin gene carriers: An 18F-dopa PET progression study. Movement disorders: official journal of the Movement Disorder Society 2009;24(15):2260-2266.

58. Sardi SP, Simuni T. New Era in disease modification in Parkinson's disease: Review of genetically targeted therapeutics. Parkinsonism Relat Disord 2019;59:32-38.

59. Schneider SA, Alcalay RN. Neuropathology of genetic synucleinopathies with parkinsonism: review of the literature. Mov Disord 2017;32(11):1504-1523.

60. Dawson TM, Dawson VL. The role of parkin in familial and sporadic Parkinson's disease. Mov Disord 2010;25(S1):S32-S39.

Preface to Chapter 3

In Chapter 3, I performed genetic fine-mapping on the *HLA* locus in PD. The *HLA* locus has been the subject of controversial results in relation to PD, and this study aims to provide further insights.

Fine-mapping involves analyzing genetic variations in a specific genomic region with a higher resolution, allowing for a more detailed examination of the genetic factors contributing to a particular trait or disease. By focusing on the *HLA* locus in PD, I aim to identify the specific *HLA* residues that are associated with the disease. This information can help in understanding the underlying mechanisms and potential therapeutic targets related to the *HLA* locus in PD.

By conducting genetic fine-mapping on the *HLA* locus, this study aims to contribute to a better understanding of the role of *HLA* in PD and potentially uncover novel insights into the disease's etiology and progression. This research can provide valuable information for the development of targeted therapies and personalized treatment approaches based on an individual's *HLA* profile.

Chapter 3: Fine mapping of the *HLA* locus in Parkinson's disease in Europeans

Eric Yu, BSc,[1,2] Aditya Ambati, PhD,[3] Maren Stolp Andersen, MD,[4,5] Lynne Krohn, MSc,[1,2] Mehrdad A. Estiar, MSc,[1,2] Prabhjyot Saini, MSc,[1,2] Konstantin Senkevich, MD, PhD,[2,6] Yuri L. Sosero, MD,[1,2] Ashwin Ashok Kumar Sreelatha, MSc, MTech,[7] Jennifer A. Ruskey, MSc,[2,6] Farnaz Asayesh, MSc,[2,6] Dan Spiegelman, MSc,[2,6] Mathias Toft, MD, PhD,[4,5] Marte K. Viken, PhD,[8,9] Manu Sharma, PhD,[7] Cornelis Blauwendraat, PhD,[10] Lasse Pihlstrøm, MD, PhD,[4] Emmanuel Mignot, MD, PhD*,[3] Ziv Gan-Or, MD, PhD*.[1,2,6]

*Equal contribution

Affiliations

1. Department of Human Genetics, McGill University, Montréal, QC, Canada. 2. The Neuro (Montreal Neurological Institute-Hospital), McGill University, Montreal, QC, Canada. 3. Stanford Center For Sleep Sciences and Medicine, Department of Psychiatry and Behavioral Sciences, Stanford University, Palo Alto, California, United States of America. 4. Department of Neurology, Oslo University Hospital, Oslo, Norway. 5. Institute of Clinical Medicine, University of Oslo, Oslo, Norway. 6. Department of Neurology and Neurosurgery, McGill University, Montréal, QC, Canada. 7. Centre for Genetic Epidemiology, Institute for Clinical Epidemiology and Applied Biometry, University of Tübingen, Tübingen, Germany 8. Department of Medical Genetics, University of Oslo and Oslo University Hospital, Oslo, Norway. 9. Department of Immunology, Oslo University Hospital, Oslo, Norway. 10. Molecular Genetics Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, United States of America.

Published: *npj Parkinsons Dis*. 2021 Sep 21;7(1):84

Abstract

We fine mapped the leukocyte antigen (*HLA)* region in 13,770 Parkinson's disease (PD) patients, 20,214 proxy-cases and 490,861 controls of European origin. Four *HLA* types were associated with PD after correction for multiple comparisons, *HLA-DQA1**03:01, *HLA-DQB1**03:02, *HLA-DRB1**04:01 and *HLA-DRB1**04:04. Haplotype analyses followed by amino acid analysis and conditional analyses suggested that the association is protective and primarily driven by three specific amino acid polymorphisms present in most *HLA-DRB1**04 subtypes - 11V, 13H and 33H (OR=0.87, 95%CI=0.83-0.90, $p<8.23x10^{-9}$ for all three variants). No other effects were present after adjustment for these amino acids. Our results suggest that specific *HLA-DRB1* variants are associated with reduced risk of PD, providing additional evidence for the role of the immune system in PD. Although effect size is small and has no diagnostic significance, understanding the mechanism underlying this association may lead to identification of new targets for therapeutics development.

Introduction

Although Parkinson's disease (PD) is primarily a neurodegenerative disorder, the role of the immune system in the pathophysiology of PD is increasingly recognized based on animal and human studies.[1-3] The immune system can be involved in the initiation of PD, as well as in its progression, and that this involvement can be peripheral and central.[3,4]

Neuropathological studies have shown evidence for microglia activation in brains of patients. However, it was initially unclear whether this activation was a part of the disease process, a consequence, or an epiphenomenon.[5] Genetic evidence also links the immune system with PD, since genes such as *LRRK2*, the human leukocyte antigen (*HLA*) locus and possibly *BST1*, all associated with PD[6] and have a role in the immune system.[7-9]

The *HLA* region on chromosome 6 includes genes that encode components of the major histocompatibility complex (MHC).[8] Several genome-wide association studies (GWASs) have shown an association between the *HLA* locus and risk of PD. In the latest GWAS, an association with *HLA-DRB5* has been reported, with a potential effect of the rs112485576 single nucleotide polymorphism (SNP) on the expression of *HLA-DRB5*.[6] Previous studies have suggested different associations with *HLA-DQA2*, *HLA-DQB1*, *HLA-DRA*, *HLA-DRB1*, *HLA-DRB5* and with haplotypes within the *HLA* region in Europeans.[10-15]

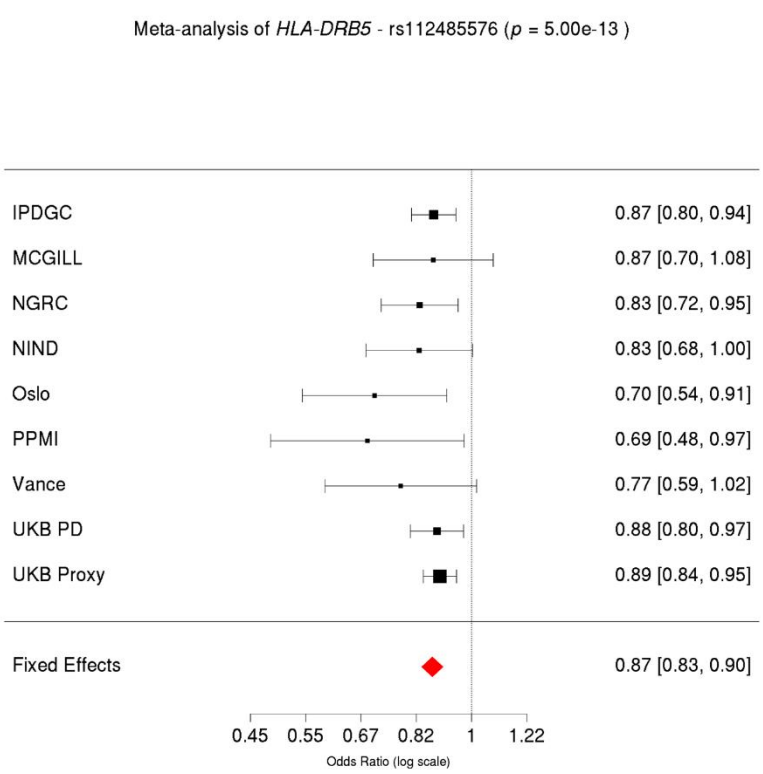
In this study, we performed the largest *HLA* alleles, haplotypes and amino acid analyses in PD on 12,137 patients, 14,422 proxy patients and 351,953 controls. We further performed conditional analyses to fine map and identify specific drivers of the association with PD in the *HLA* region.

Results

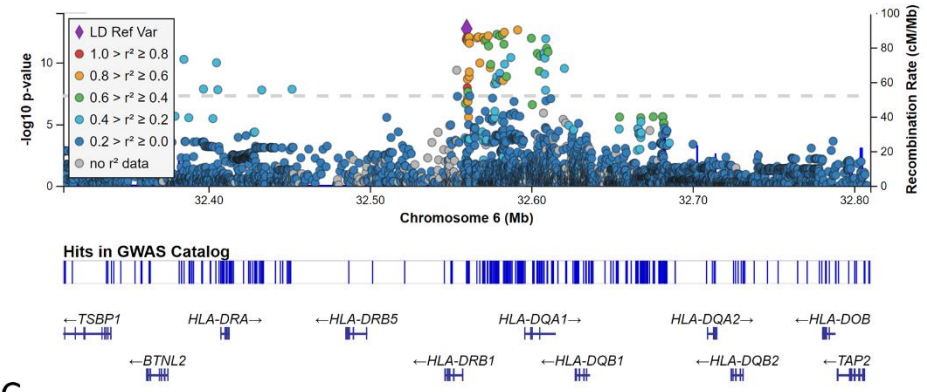
Meta-analysis of *HLA* types in Parkinson's disease suggests a single association

After standard QC, a total of 12,137 patients, 14,422 proxy patients and 351,953 controls were included in the analysis (Supplementary Table 1). As shown in Figure 1A, our SNP-level meta-analysis validated the previous association for SNP, rs112485576, in the *HLA* locus (In the current analysis: OR=0.87, 95% CI= 0.83-0.90, $p=5.00x10^{-13}$, in the previous meta analysis: OR=0.85, 95% CI= 0.82-0.87, $p=6.96x10^{-28}$).[6] No residual HLA effects were found after adjustment of this SNP (Figure 1B, C), indicating that the association of this locus was primarily driven by a single genetic risk factor. We also validated previous associations for the SNPs rs17425622, rs2395163, rs3129882, rs9275326 in the *HLA* locus (Supplementary Figure 2-5).

A


Meta-analysis of HLA-DRB5 - rs112485576 (p = 5.00e-13)
IPDGC
0.87 [0.80, 0.94]
MCGILL
0.87 [0.70, 1.08]
NGRC
0.83 [0.72, 0.95]
NIND
0.83 [0.68, 1.00]
Oslo
0.70 [0.54, 0.91]
PPMI
0.69 [0.48, 0.97]
Vance
0.77 [0.59, 1.02]
UKB PD
0.88 [0.80, 0.97]
UKB Proxy
0.89 [0.84, 0.95]
Fixed Effects
0.87 [0.83, 0.90]
0.45 0.55 0.67 0.82 1 1.22
Odds Ratio (log scale)

B


LD Ref Var
1.0 > r² ≥ 0.8
0.8 > r² ≥ 0.6
0.6 > r² ≥ 0.4
0.4 > r² ≥ 0.2
0.2 > r² ≥ 0.0
no r² data
-log10 p-value
Recombination Rate (cM/Mb)
32.40
32.50
32.60
32.70
32.80
Chromosome 6 (Mb)
Hits in GWAS Catalog
←TSBP1
HLA-DRA→
←HLA-DRB5
HLA-DQA1→
HLA-DQA2→
←HLA-DOB
←BTNL2
←HLA-DRB1
←HLA-DQB1
←HLA-DQB2
←TAP2

C

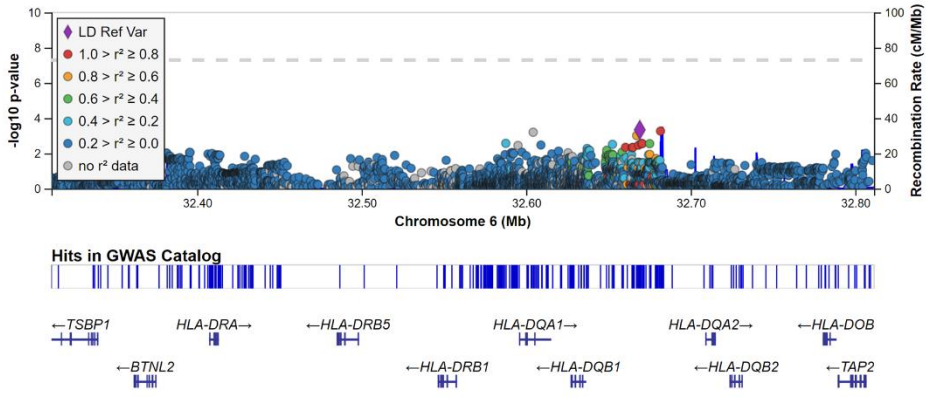

LD Ref Var
1.0 > r² ≥ 0.8
0.8 > r² ≥ 0.6
0.6 > r² ≥ 0.4
0.4 > r² ≥ 0.2
0.2 > r² ≥ 0.0
no r² data
-log10 p-value
Recombination Rate (cM/Mb)
32.40
32.50
32.60
32.70
32.80
Chromosome 6 (Mb)
Hits in GWAS Catalog
←TSBP1
HLA-DRA→
←HLA-DRB5
HLA-DQA1→
HLA-DQA2→
←HLA-DOB
←BTNL2
←HLA-DRB1
←HLA-DQB1
←HLA-DQB2
←TAP2

Figure 1. Validation of previously associated top *HLA* locus SNP (rs112485576) in our cohort. **a)** Forest plot describing the effect size and 95% confidence interval of rs112485576 for each cohort and fixed-effect meta-analysis. **b-c)** Two LocusZoom plots highlighting the significant variants before (**b**) and after (**c**) the conditional analysis on rs112485576. Dashed lines correspond to the significance threshold. Linkage disequilibrium values are shown with respect to the most significant SNP in the locus.

We next performed a meta-analysis association study of all *HLA* types with carrier frequency above 1%. After *HLA* imputation, a total of 141 different *HLA* types across 10 *HLA* loci were included (setting the Bonferroni corrected threshold for statistical significance on $\alpha=3.55x10^{-4}$; 0.05/141). Following these analyses, we found four *HLA* alleles that were associated with PD (Table 1, results for other *HLA* alleles are detailed in Supplementary Table 5): *HLA-DQA1**03:01, *HLA-DQB1**03:02, *HLA-DRB1**04:01, *HLA-DRB1**04:04. These four alleles are all located within a small genomic segment and have similar odds ratio ranging between 0.84-0.89. Three of the four alleles have similar carrier frequencies, indicating that they could be part of the same haplotypes, with the fourth potentially representing a sub-haplotype (Table 1).

HLA haplotype analysis

For haplotype analysis, we allowed for up to three genes to be included in each haplotype, since including more than three genes generated multiple haplotypes with low allele frequency that could not be analyzed at the current sample size. A total of 84 different HLA haplotypes (Supplementary Table 6) with allele frequency >1% were identified, setting the cut-off Bonferroni corrected p value for statistical significance at $\alpha=5.95x10^{-4}$. Three different HLA haplotypes were associated with PD after correction for multiple comparisons: *DQA1**03:01~*DQB1**03:02, *DRB1**04:01~*DQA1**03:03 and *DRB1**04:04~*DQA1**03:01. Upon further examination, this

association was found to be driven by several well-known sub-haplotypes, *DRB1**04:04~*DQA1**03:01~*DQB1**03:02 and *DRB1**04:01~*DQA1**03:01/3~*DQB1**03:01/2 (Supplementary Table 6). Because both *DQA1**03:01 and *DQA1**03:03 as well as *DQB1**03:01 and *DQB1**03:03 are present within the extended *DRB1**04:01 haplotype, it is likely that these associations are driven by *DRB1*.

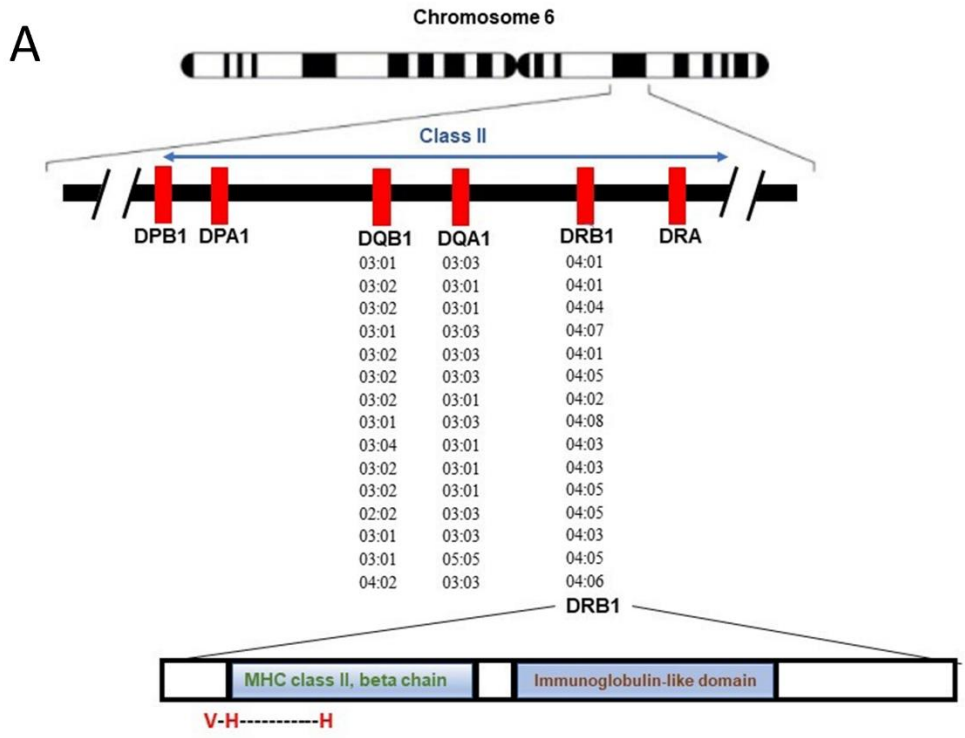
Meta-analysis of the association of HLA amino acid changes with Parkinson's disease

To further identify the specific source of the association in the *HLA* locus, we performed an analysis of 636 amino acid changes in the *HLA* genes, setting the cut-off Bonferroni corrected p value for statistical significance at $\alpha=7.86\text{x}10^{-5}$. Ten amino acid changes were significantly associated with reduced risk of PD (Supplementary Table 7). The top three associated variants are linked amino-acids 11V, 13H and 33H (Table 3) present in all *DRB1**04 subtypes, complementing the HLA haplotype analysis. Four other variants, 26S, 47Q, 56R and 76V, in the *DQA1* gene, are in perfect LD with each other ($r^2=1$, D'=1, Supplementary Table 7) and in partial LD ($r^2=0.38$, D'=0.85) with the *DRB1* variants. The association of these *DQA1* variants is weaker than the *DRB1* variants in terms of both effect size and statistical association (Supplementary Table 7). Three other variants, 71T, 74E, and 75L, are in the *DQB1* gene, are also in perfect LD with each other ($r^2=1$, D'=1, Supplementary Table 7) and in partial LD ($r^2=0.16$, D'=0.99) with *DRB1* 13H and 33H.

Table 1. Meta-analyses of *HLA* alleles association

Allele	Freq Cases	Freq Controls	OR (95% CI)	P-value[a]	Direction	HetPVal
*DQA1**03:01	0.168	0.185	0.86 (0.81-0.90)	3.54e-08	---------	0.85
*DQB1**03:02	0.182	0.199	0.87 (0.83-0.92)	7.62e-07	---------	0.58
*DRB1**04:01	0.187	0.221	0.89 (0.84-0.94)	2.82e-05	---------	0.87
*DRB1**04:04	0.068	0.082	0.84 (0.77-0.91)	8.21e-05	---------	0.85

Abbreviations. Freq Cases: Frequency of allele in patients; Freq Controls: Frequency of allele in controls; OR (95% CI): Odds ratio and 95% confidence interval; Direction: Direction of beta for each cohort; HetPVal: P-value of heterogeneity. [a] Bonferroni correction for multiple comparisons set the threshold for statistical significance to $\alpha=3.55x10^{-4}$.


A
Chromosome 6
Class II
DPB1
DPA1
DQB1
DQA1
DRB1
DRA
03:01 03:03 04:01
03:02 03:01 04:01
03:02 03:01 04:04
03:01 03:03 04:07
03:02 03:03 04:01
03:02 03:03 04:05
03:02 03:01 04:02
03:01 03:03 04:08
03:04 03:01 04:03
03:02 03:01 04:03
03:02 03:01 04:05
02:02 03:03 04:05
03:01 03:03 04:03
03:01 05:05 04:05
04:02 03:03 04:06
DRB1
MHC class II, beta chain
Immunoglobulin-like domain
V-H----------H

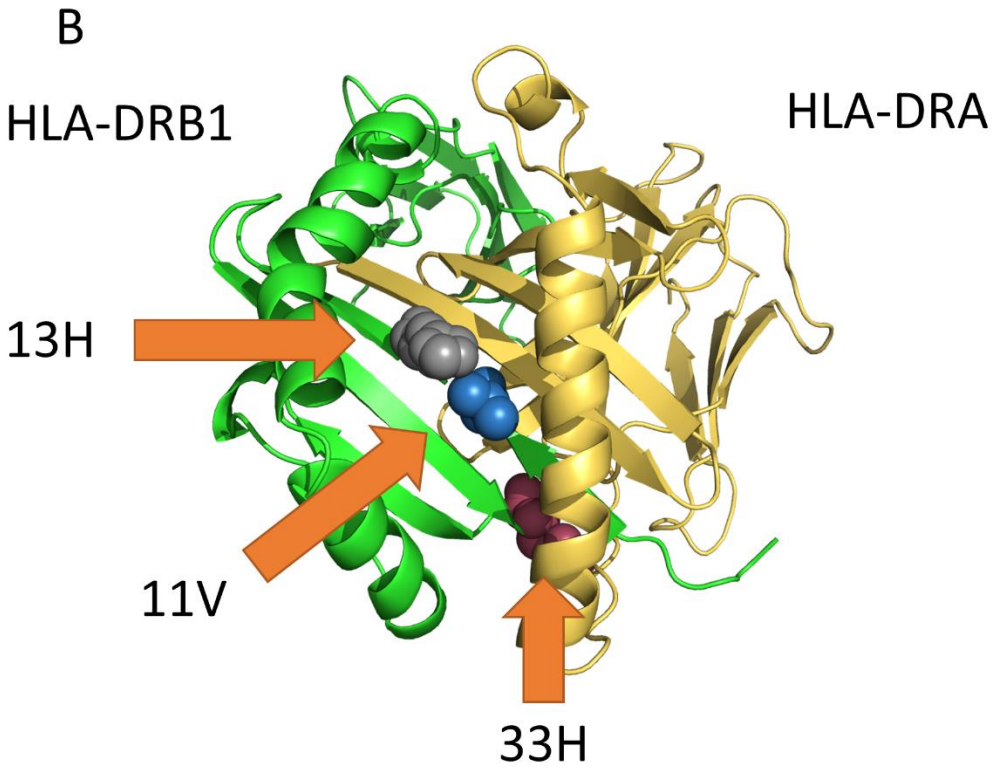

B
HLA-DRB1
HLA-DRA
13H
11V
33H

Figure 2. Association of the *HLA-DRB1* alleles and location of associated amino acids. a) The location of the *HLA* locus, alleles and amino acids associated with Parkinson's disease in the current study. **b)** 3D model of HLA-DRB1 – HLA-DRA and the location of the 11V, 13H and 33H amino acids associated with PD (highlighted by arrows). The model was generated with PyMol v. 2.4.1 (pdb 4is6).

Conditional analyses confirm that *DRB104 amino acid variants likely drive the association of the *HLA* locus with PD**

To further determine the specific genes or variants that drive these associations, we performed a set of conditional analyses, and re-analyzed the allele types, haplotypes and amino acid associations with PD. We conditioned the *HLA* type regression model on the following: rs112485576 (the top GWAS hit in the *HLA* locus), *DQA1**03:01, *DQA1**03:03 and the *DRB1* variant 13H. We have also adjusted for the PD PRS, to examine a potential polygenic effect (Supplementary Tables 5, 7). While the adjustment for the *DRB1* variant 13H completely eliminated the associations in the *DQA1* gene, adjustment for *DQA1**03:01 and *DQA1**03:03 did not completely eliminate the association of the *DRB1* gene (Supplementary Table 5), again supporting this gene and these specific amino acids (11V, 13H and 33H, Table 3) as the drivers of the association in the *HLA* locus. Adjustment for PRS did not change the results. It is also worth noting that the *DRB1* variants 11V (r^2=0.96, D'=0.99), 13H (r^2=0.99, D'=0.99), and 33H (r^2=0.99, D'=0.99) are in LD with rs112485576.

Table 2. Meta-analyses of HLA haplotype association

Haplotype	Freq Cases	Freq Controls	OR (95% CI)	P-value[a]	Direction	HetPVal
*DQA1**03:01-*DQB1**03:02	0.157	0.173	0.87 (0.82-0.93)	7.21e-05	+----+---	0.62
*DRB1**04:04-*DQA1**03:01	0.067	0.081	0.83 (0.76-0.91)	9.82e-05	+--------	0.34
*DRB1**04:01-*DQA1**03:03	0.115	0.143	0.87 (0.81-0.94)	4.96e-04	+--------	0.71

Abbreviations. Freq Cases: Frequency of allele in patients; Freq Controls: Frequency of allele in controls; OR (95% CI): Odds ratio and 95% confidence interval; Direction: Direction of beta for each cohort; HetPVal: P-value of heterogeneity.

[a] Bonferroni correction for multiple comparisons set the threshold for statistical significance to $\alpha=5.95x10^{-4}$.

Table 3. Meta-analyses of HLA amino acid changes association.

Amino Acid	Freq Cases	Freq Controls	OR (95% CI)	P-value[a]	Direction	HetPVal
DRB1 13H	0.289	0.331	0.87 (0.83-0.91)	4.32e-09	---------	0.60
DRB1 33H	0.289	0.331	0.87 (0.83-0.91)	4.32e-09	---------	0.60
DRB1 11V	0.302	0.342	0.87 (0.83-0.91)	8.22e-09	---------	0.45

Abbreviations. Freq Cases: Frequency of allele in patients; Freq Controls: Frequency of allele in controls; OR (95% CI): Odds ratio and 95% confidence interval; Direction: Direction of beta for each cohort; HetPVal: P-value of heterogeneity. P-value of heterogeneity.

[a] Bonferroni correction for multiple comparisons set the threshold for statistical significance to $\alpha=7.86x10^{-5}$.

Discussion

In the current study, we performed a thorough analysis of the *HLA* region and examined its association with PD in the European population using a total of 12,137 patients, 14,422 proxy

patients and 351,953 controls. Following a series of regression models and conditional analyses, our results indicate that the drivers of the association in the *HLA* region are three amino acid changes specific of *HLA-DRB1**04 subtypes, 11V, 13H and 33H (Figure 2). Two of these amino acid changes, 13H and 33H are in perfect LD, and 11V is in very strong LD with the other two variants. This study agrees with a smaller HLA sequencing study[12] in 1,597 PD cases and 1,606 controls which also observed a protective effect of *DRB1**04 and the same amino acids, although it also reported additional associations with *DRB1**01:01 and *DRB1*10:01* which were not confirmed in the current study. Interestingly, the V-H-H motif at position 11V, 13H and 33H are central to the DRB1*04:01 heterodimer and contribute to peptide binding, notably through pocket P6[16] (Figure 2).

Previous studies on the *HLA* genomic region in PD have reported associations of different genes and HLA types with PD, including *HLA-DQA2*, *HLA-DQB1*, *HLA-DRA*, *HLA-DRB1* and *HLA-DRB5*.[10-15] The suggested association with *HLA-DRB5* reported in the most recent PD GWAS is based on an expression quantitative trait locus (eQTL) analysis, as the top associated SNP in this region, rs112485576, was also associated with differential expression of *HLA-DRB*.[6] A previous study of 2,000 PD patients and 1,986 controls has implicated a non-coding variant (rs3129882) within *HLA-DRA* as driving the association with PD and suggested that this variant affects the expression of *HLA-DR* and *HLA-DQ* genes.[11] Similarly, another study suggested that the same variant in *HLA-DRA* (rs3129882) is associated with differential expression of MHC-II on immune cells.[17] While our study does not rule out this possibility, since the main variants driving the association are amino acid changes in *DRB1**04 that will affect epitope binding ability, it is likely that the effect on PD risk is through these variants and not due to modified expression. Additional functional studies will be required to study this hypothesis.

The current study adds further support to the hypothesis suggesting an involvement of the peripheral and central immune system in PD. On top of the *HLA* locus, several other genes with potential roles in the immune system, including *LRRK2* and potentially *BST1*,[7,9] have been implicated in PD.[6] In the periphery, there are notable changes in the immune system of PD patients compared to controls, as peripheral monocytes have differential expression of immune related proteins and markers.[3] Whether these changes are drivers of the disease or a result of the disease is still undetermined, but accumulating evidence suggest that they can be part of the pathogenic process of PD. In the central nervous system, pathological studies suggest that microglia cells may have a central role in PD.[18] Microgliosis is a prominent pathological finding in post-mortem brains of PD patients, and evidence suggest that microglia activation occurs early in the disease process and may be involved in the pathogenesis of PD.[3] The specific contribution of *HLA* to these processes is still unclear and needs to be further studied.

One intriguing possibility that may directly involve *HLA* with PD is the potential interaction of HLA-DRB1*04 with α-synuclein, notably an epitope surrounding p.S129. Recent data has shown that α-synuclein fragments can bind MHC and increase T cell reactivity.[19] This activity is proinflammatory, involves both CD4 and CD8 cells and may occur before the onset of motor symptoms,[19,20] suggesting an involvement of inflammation in early PD pathogenesis. Studying specific α-synuclein fragments has suggested that two major regions of α-synuclein may be associated with increased T cell reactivity in PD, with preferential CD4 activity: an N-terminal region involving amino acid p.Y39 and a C-terminal region surrounding amino acid p.S129, two important residues undergoing phosphorylation.[19] The phosphorylation of p.S129 is particularly interesting as it is well known in promoting aggregation.[21,22] Further analysis focused on the p.Y39 region suggested an association with α-synuclein-specific p.Y39 T cell responses and HLA

DRB1*15:01 and DRB5*01:01 presentation. This association was abolished by phosphorylation, which reduced binding of p.Y39-phosphorylated α-synuclein.[19,20] Other experiments by these authors have also suggested CD8+ T cells responses mediated by HLA-A11*01 presentation of epitopes in the same N-terminal region of α-synuclein. However, in the current study we could not confirm an association of these HLA types with PD.

More interestingly in the context of our work, increased CD4+ T cell response to both p.S129 phosphorylated and unphosphorylated α-synuclein was also demonstrated, suggesting involvement of DQB1*05:01 and DQB1*04:02, as these alleles strongly bound these α-synuclein epitopes.[19] These HLA alleles, however, are not associated with risk of PD in the current study. However, the authors reported in the supplementary data that DRB1*04:01 was also a selective and strong binder of the same α-synuclein epitope with p.S129, but only when the epitope was unphosphorylated.[19] Notably, no other DRB1 alleles that were assayed in this study[19] had increased binding affinity to α-synuclein epitope with p.S129, except for the DRB1*04:01 allele that was a strong binder only when unphosphorylated. Binding register analysis using Immune Epitope Database (IEDB) MHC-II Binding Prediction (http://tools.iedb.org/mhcii/) suggests that this epitope binds a 9 amino acid AYEMPSEEG core, with p.S129 at P6 position, a position postulated to be important based on our HLA-DR amino acid analysis presented above. As CD4+ T cell responses are generally stronger when epitopes are presented by HLA-DR versus HLA-DQ,[23] HLA-DRB1*04 responses to the p.S129 unphosphorylated form of α-synuclein could be dominant in individuals with HLA-DRB1*04, explaining the protective effect of this HLA subtype in PD. Additional experiments will be needed to further explore this hypothesis.

Our study has several limitations. First, this study was performed on European populations, and the results may be limited to this population only. Additional studies in other populations are

required. Several studies on *HLA* types and PD have been performed in Asian populations,[24-27] and the GWAS risk variant rs112485576 has a similar OR (0.85) in the largest Asian GWAS to date,[28] yet larger studies are required, as well as studies in other populations. An additional potential limitation of our study is its use of imputation rather than fully sequenced HLA types. Given the very high performance of the imputation tool when compared to full sequencing (Supplementary Table 2), the potential effect of imputation inaccuracies is likely small and should be diluted in our large sample size. In addition, we cannot rule out that other, rarer HLA types that were not included in the current analysis may also have a role in PD. An additional limitation of the current study is that by adjusting for sex we eliminate potential sex-specific effects. It is possible that specific HLA types are relevant in one sex more or less than the other, and this should be studied in larger, sex-stratified cohorts.

To conclude, our results suggest a role for the *HLA-DRB1* gene in susceptibility for PD, and provide further evidence for the importance of the immune system in PD. Since the effect is small, it does not merit routine HLA typing in PD, but understanding the mechanism underlying this association may lead to better understanding of PD in general and offer new targets for future immune-related treatment.

Methods

Study population

This study was designed as a meta-analysis of multiple cohorts, including a total of 13,770 PD patients, 20,214 proxy-patients and 490,861 controls, as detailed in Supplementary Table 1. In brief, we included cohorts and datasets from eight independent sources: International Parkinson's Disease Genomics Consortium (IPDGC) NeuroX dataset (dbGap phs000918.v1.p1, including datasets from multiple independent cohorts as previously described),[29] McGill University

(McGill),[30] National Institute of Neurological Disorders and Stroke (NINDS) Genome-Wide genotyping in Parkinson's Disease (dbGap phs000089.v4.p2),[31] NeuroGenetics Research Consortium (NGRC) (dbGap phs000196.v3.p1),[11] Oslo Parkinson's Disease Study (Oslo), Parkinson's Progression Markers Initiative (PPMI), Vance (dbGap phs000394) and PD cases and proxy-cases from the UK Biobank (UKB). Proxy-cases are first degree relatives of PD patients, thus sharing ~50% of the patients' genetic background and eligible to serve as proxies, as previously described.[32] All cohorts were previously included in the most recent PD GWAS.[6] Study protocols were approved by the relevant Institutional Review Boards and all patients signed informed consent before participating in the study.

Pre-imputation genotype quality control

In order to include only high-quality samples and SNPs, standard quality control (QC) was performed on all datasets individually using PLINK v1.9.[33] Standard GWAS QC was done to filter out samples and SNPs with low call rate, heterozygote outliers along with gender mismatch as previously described.[6] SNPs deviating from Hardy-Weinberg equilibrium were removed. Only samples of European ancestry clustering with HapMap v3 using principal component analysis were included as shown in Supplementary Figure 1. In order to exclude related individuals, we examined relatedness in each dataset separately, followed by relatedness test across all datasets combined, to exclude individuals who were included in more than one dataset. All individuals with pi_hat >0.125 were excluded using GCTA v1.26.0.[34]

UK Biobank quality control

For the analysis of the UKB data, unrelated participants of European ancestry (field 22006), with low missingness rate (field 220027) were included after exclusion of heterozygosity outliers as previously described.[6]. PD patients from the UK Biobank were included based on self-report (field

20002) or based on their International Classification of disease diagnosis code (ICD-10, code G20, field 41270). From the remaining participants, proxy-cases were defined as first degree relatives (parents or siblings, field 20112-20114) of patients with PD. Principal components were calculated using flashpca[35] after excluding related individuals as described above. The control group was divided randomly to two groups of controls: one was included in the GWAS comparing PD patients from UKB and controls, and the second was included in the GWAS comparing proxy-cases from UKB and controls. This division was done proportionally to the size of each GWAS.

Imputation

For SNP imputation of each dataset, we used the Michigan Imputation Server on the 1,000 Genomes Project panel (Phase 3, Version 5) using Minimac3 and SHAPEIT v2.r790. Imputed UK Biobank genotyped data v3 were downloaded in July 2019. All variants with an imputation quality (r^2) of >0.30 were labeled as soft calls and >0.80 were labeled as hard calls. Soft calls were only used together with the hard calls for polygenic risk score calculation (see below); hard calls were used for all other analyses.

Association analysis of common variants on chromosome 6

Prior to determining HLA types, we performed a simple association test of all SNPs located on chromosome 6, to verify that we identified the same hit in the HLA region as previously described.[6] For this purpose, we generated summary statistics of chromosome 6 for each dataset, and used logistic regression with an additive model adjusting for age at onset for patients and age at enrollment of controls, sex and population stratification (first 10 principal components) with PLINK v2.00a2LM (25 Oct 2019).[33] The UK Biobank data was analyzed similarly using logistic regression adjusting for age, sex, the first 10 principal components and Townsend index to account for additional potential population stratification confounders. Finally, to harmonize effects in cases

and proxy-cases, summary statistics for proxy-cases were rescaled based on genome-wide association study by proxy (GWAX) as previously described.[32] To meta-analyze the different datasets, we performed a fixed-effect meta-analysis using METAL with an inverse-variance-based model.[36]

HLA locus analysis

HLA imputation

To impute specific HLA types for each individual, we inferred two field resolution HLA alleles using HIBAG v1.22.0, a statistical method for HLA type imputation in R.[37] HIBAG was shown to be as accurate or more accurate in Europeans compared to other types of HLA imputation tools.[38] HIBAG provided a reference panel for Europeans (n = 2,572) with high imputation accuracy for *HLA-A*, *HLA-B*, *HLA-C*, class I genes, and *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1*, and *HLA-DRB1*, class II genes. *HLA-DRB3*, *HLA-DRB4* and *HLA-DRB5* imputation models were trained using HIBAG[37] on European origin sample training set (n = 3,267) genotyped on the Illumina Infinium PsychArray-24 chip and fully sequenced at 8-digit resolution for HLA loci. These models were validated in a test set (n=886) with high accuracy (Supplementary Table 2). Imputation accuracy for European *DRB1**04 alleles was determined for *DRB1**04:01 *DRB1**04:02 *DRB1**04:03, *DRB1**04:04, *DRB1**04:05, *DRB1**04:07, *DRB1**04:08. Alleles with an imputation probability of <0.5 were defined as undetermined and individuals with two or more undetermined alleles were excluded from the analysis (Supplementary Table 1 details the numbers included for each allele in each cohort after all quality control steps). To further examine imputation accuracy, the results of the *DRB1* imputation were compared against high throughput HLA sequencing in 380 PD samples from Oslo. The combined frequency of seven different *DRB1**04 alleles detected in sequence data was 0.15 with the 04:01 and 04:04 alleles being the most common (Supplementary Table 3).

Imputation accuracy for DRB1*04 alleles was very high at 2-digit resolution (Supplementary Table 4).

Statistical analysis

To examine the association of HLA alleles with PD, we used R v3.6 to perform logistic regression, adjusting for age at onset, sex and the first 10 principal components. The UK Biobank dataset was also adjusted for Townsend index. Haplotype analyses were performed using haplo.stats in R with logistic regression as stated above. Only haplotypes with posterior probability >0.2 and carrier frequency of >1% were included in the analysis. Amino acid association analyses were performed using HIBAG after converting P-coded alleles to amino acid sequences for exon 2, 3 of HLA class I genes and exon 2 of class II genes. Amino acid associations were tested using logistic regression as described above. A polygenic risk score (PRS) was calculated using PRSice v 2.2.11 without linkage disequilibrium (LD) clumping or P thresholding.[39] The beta weights from the summary statistics of the 90 genome-wide significant variants in the latest PD GWAS[6] were used in the PRS. To make sure that all possible variants were included in the PRS analysis, we also performed imputation using the Haplotype Reference Consortium panel (HRC) (Version r1.1 2016) with Minimac4 and Eagle v2.4. Ambiguous variant (rs6658353) and rs112485576 from the HLA region were excluded from the PRS calculation. To examine whether secondary hits exist in the HLA region, we adjusted for significant HLA variants, HLA alleles, HLA amino acid changes and PRS, by introducing significant findings from the first analyses as covariates in the regression models. Statistical analyses were only performed on alleles, haplotypes and amino acid changes with more than 1% carrier frequency. *P* value significance levels were adjusted using Bonferroni correction. Meta-analysis was performed as described above. All missing data were excluded from the analyses.

Code availability

The scripts used in this analysis is available at https://github.com/gan-orlab/HLA_HIBAG/.

Data availability

Anonymized data will be shared by request from any qualified investigator.

References

1 Sanchez-Guajardo, V., Barnum, C. J., Tansey, M. G. & Romero-Ramos, M. Neuroimmunological processes in Parkinson's disease and their relation to α-synuclein: microglia as the referee between neuronal processes and peripheral immunity. *ASN neuro* **5**, AN20120066 (2013).

2 Wang, Q., Liu, Y. & Zhou, J. Neuroinflammation in Parkinson's disease and its potential as therapeutic target. *Translational Neurodegeneration* **4**, 19, doi:10.1186/s40035-015-0042-0 (2015).

3 Tansey, M. G. & Romero-Ramos, M. Immune system responses in Parkinson's disease: Early and dynamic. *Eur J Neurosci* **49**, 364-383, doi:10.1111/ejn.14290 (2019).

4 Devos, D. *et al.* Colonic inflammation in Parkinson's disease. *Neurobiology of disease* **50**, 42-48 (2013).

5 Zhang, W. *et al.* Aggregated α-synuclein activates microglia: a process leading to disease progression in Parkinson's disease. *The FASEB Journal* **19**, 533-542 (2005).

6 Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol* **18**, 1091-1102, doi:10.1016/s1474-4422(19)30320-5 (2019).

7 Malavasi, F. *et al.* CD38 and CD157 as receptors of the immune system: a bridge between innate and adaptive immunity. *Mol Med* **12**, 334-341, doi:10.2119/2006-00094.Malavasi (2006).

8 Shiina, T., Hosomichi, K., Inoko, H. & Kulski, J. K. The HLA genomic loci map: expression, interaction, diversity and disease. *J Hum Genet* **54**, 15-39, doi:10.1038/jhg.2008.5 (2009).

9 Wallings, R. L., Herrick, M. K. & Tansey, M. G. LRRK2 at the Interface Between Peripheral and Central Immune Function in Parkinson's. *Front Neurosci* **14**, 443, doi:10.3389/fnins.2020.00443 (2020).

10 Ahmed, I. *et al.* Association between Parkinson's disease and the HLA-DRB1 locus. *Mov Disord* **27**, 1104-1110, doi:10.1002/mds.25035 (2012).

11 Hamza, T. H. *et al.* Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. *Nat Genet* **42**, 781-785, doi:10.1038/ng.642 (2010).

12 Hollenbach, J. A. *et al.* A specific amino acid motif of HLA-DRB1 mediates risk and interacts with smoking history in Parkinson's disease. *Proc Natl Acad Sci U S A* **116**, 7419-7424, doi:10.1073/pnas.1821778116 (2019).

13 Wissemann, W. T. *et al.* Association of Parkinson disease with structural and regulatory variants in the HLA region. *The American Journal of Human Genetics* **93**, 984-993 (2013).

14 Bandres-Ciga, S. *et al.* The Genetic Architecture of Parkinson Disease in Spain: Characterizing Population-Specific Risk, Differential Haplotype Structures, and Providing Etiologic Insight. *Mov Disord* **34**, 1851-1863, doi:10.1002/mds.27864 (2019).

15 Hill-Burns, E. M., Factor, S. A., Zabetian, C. P., Thomson, G. & Payami, H. Evidence for more than one Parkinson's disease-associated variant within the HLA region. *PLoS One* **6**, e27109, doi:10.1371/journal.pone.0027109 (2011).

16 Kamoun, M. *et al.* HLA Amino Acid Polymorphisms and Kidney Allograft Survival. *Transplantation* **101**, e170-e177, doi:10.1097/TP.0000000000001670 (2017).

17 Kannarkat, G. T. *et al.* Common Genetic Variant Association with Altered HLA Expression, Synergy with Pyrethroid Exposure, and Risk for Parkinson's Disease: An

Observational and Case-Control Study. *NPJ Parkinsons Dis* **1**, doi:10.1038/npjparkd.2015.2 (2015).

18 Levesque, S. *et al.* Reactive microgliosis: extracellular micro-calpain and microglia-mediated dopaminergic neurotoxicity. *Brain* **133**, 808-821, doi:10.1093/brain/awp333 (2010).

19 Sulzer, D. *et al.* T cells from patients with Parkinson's disease recognize α-synuclein peptides. *Nature* **546**, 656-661, doi:10.1038/nature22815 (2017).

20 Lindestam Arlehamn, C. S. *et al.* α-Synuclein-specific T cell reactivity is associated with preclinical and early Parkinson's disease. *Nat Commun* **11**, 1875, doi:10.1038/s41467-020-15626-w (2020).

21 Fayyad, M. *et al.* Investigating the presence of doubly phosphorylated α-synuclein at tyrosine 125 and serine 129 in idiopathic Lewy body diseases. *Brain Pathol* **30**, 831-843, doi:10.1111/bpa.12845 (2020).

22 Li, X. Y. *et al.* Phosphorylated Alpha-Synuclein in Red Blood Cells as a Potential Diagnostic Biomarker for Multiple System Atrophy: A Pilot Study. *Parkinsons Dis* **2020**, 8740419, doi:10.1155/2020/8740419 (2020).

23 Grifoni, A. *et al.* Characterization of Magnitude and Antigen Specificity of HLA-DP, DQ, and DRB3/4/5 Restricted DENV-Specific CD4+ T Cell Responses. *Front Immunol* **10**, 1568, doi:10.3389/fimmu.2019.01568 (2019).

24 Chang, K. H., Wu, Y. R., Chen, Y. C., Fung, H. C. & Chen, C. M. Association of genetic variants within HLA-DR region with Parkinson's disease in Taiwan. *Neurobiol Aging* **87**, 140 e113-140 e118, doi:10.1016/j.neurobiolaging.2019.11.002 (2020).

25 Ma, Z. G., Liu, T. W. & Bo, Y. L. HLA-DRA rs3129882 A/G polymorphism was not a risk factor for Parkinson's disease in Chinese-based populations: a meta-analysis. *Int J Neurosci* **125**, 241-246, doi:10.3109/00207454.2014.926349 (2015).

26 Sun, C. *et al.* HLA-DRB1 alleles are associated with the susceptibility to sporadic Parkinson's disease in Chinese Han population. *PLoS One* **7**, e48594, doi:10.1371/journal.pone.0048594 (2012).

27 Zhao, Y. *et al.* Association of HLA locus variant in Parkinson's disease. *Clin Genet* **84**, 501-504, doi:10.1111/cge.12024 (2013).

28 Foo, J. N. *et al.* Identification of Risk Loci for Parkinson Disease in Asians and Comparison of Risk Between Asians and Europeans: A Genome-Wide Association Study. *JAMA Neurol* **77**, 746-754, doi:10.1001/jamaneurol.2020.0428 (2020).

29 Nalls, M. A. *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat Genet* **46**, 989-993, doi:10.1038/ng.3043 (2014).

30 Gan-Or, Z. *et al.* The Quebec Parkinson Network: A Researcher-Patient Matching Platform and Multimodal Biorepository. *Journal of Parkinson's disease* **10**, 301-313, doi:10.3233/jpd-191775 (2020).

31 Simón-Sánchez, J. *et al.* Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat Genet* **41**, 1308-1312, doi:10.1038/ng.487 (2009).

32 Liu, J. Z., Erlich, Y. & Pickrell, J. K. Case-control association mapping by proxy using family history of disease. *Nat Genet* **49**, 325-331, doi:10.1038/ng.3766 (2017).

33 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, doi:10.1186/s13742-015-0047-8 (2015).

34 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82, doi:10.1016/j.ajhg.2010.11.011 (2011).

35 Abraham, G. & Inouye, M. Fast principal component analysis of large-scale genome-wide data. *PLoS One* **9**, e93766, doi:10.1371/journal.pone.0093766 (2014).

36 Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-2191 (2010).

37 Zheng, X. *et al.* HIBAG—HLA genotype imputation with attribute bagging. *The pharmacogenomics journal* **14**, 192-200 (2014).

38 Karnes, J. H. *et al.* Comparison of HLA allelic imputation programs. *PLoS One* **12**, e0172444, doi:10.1371/journal.pone.0172444 (2017).

39 Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* **8**, giz082 (2019).

Preface to Chapter 4

In Chapter 4, I plan to investigate the role of the *HLA* locus in both RBD and DLB. Although the association between *HLA* and RBD and DLB is not yet clear, previous studies have indicated the presence of neuroinflammation in these conditions.

Neuroinflammation, characterized by the activation of immune cells in the central nervous system, has been implicated in various neurodegenerative disorders, including RBD and DLB. The *HLA* locus, with its involvement in immune response and antigen presentation, is a potential candidate for further investigation in relation to these disorders.

This study aims to conduct the largest *HLA* fine-mapping analyses of RBD and DLB to date. By analyzing the genetic variations within the *HLA* locus, there will be a better understanding of its potential role in RBD and DLB. The fine-mapping approach allows for a more detailed examination of specific *HLA* alleles, variants, or residues that may be associated with these conditions.

Through this research, I aim to contribute to the current knowledge on the involvement of the *HLA* locus in RBD and DLB. By uncovering potential associations and identifying specific *HLA* variants or residues, we may gain valuable insights into the underlying mechanisms of neuroinflammation and its impact on these disorders. This information can potentially guide future studies, help develop targeted therapies, and improve our understanding of RBD and DLB pathogenesis.

Chapter 4: *HLA* in isolated REM sleep behavior disorder and Lewy body dementia.

Eric Yu, BSc[1,2], Lynne Krohn, PhD[1,2], Jennifer A. Ruskey, MSc[2,3], Farnaz Asayesh, MSc[2,3], Dan Spiegelman, MSc[2,3], Zalak Shah, PhD[4], Ruth Chia, PhD[5], Isabelle Arnulf, MD, PhD[6] , Michele T.M. Hu, MBBS, PhD[7,8], Jacques Y. Montplaisir, MD, PhD[9,10], Jean-François Gagnon, PhD[9,11], Alex Desautels, MD, PhD[9,12], Yves Dauvilliers, MD, PhD[13], Gian Luigi Gigli, MD[14,15], Mariarosaria Valente, MD[14,15], Francesco Janes, MD, PhD[14], Andrea Bernardini, MD[14], Birgit Högl, MD[16], Ambra Stefani, MD, PhD[16], Abubaker Ibrahim, MD[16], Anna Heidbreder, MD[16,17], Karel Sonka, MD, PhD[18], Petr Dusek, MD, PhD[18], David Kemlink, MD, PhD[18], Wolfgang Oertel, MD[19], Annette Janzen, MD[19], Giuseppe Plazzi, MD[20,21], Elena Antelmi, MD, PhD[22], Michela Figorilli, MD, PhD[23], Monica Puligheddu, MD, PhD[23], Brit Mollenhauer, MD[24,25], Claudia Trenkwalder, MD[24,25], Friederike Sixel-Döring, MD[19,24], Valérie Cochen De Cock, MD, PhD[26,27], Luigi Ferini-Strambi, MD[28], Femke Dijkstra, MD[29,30,31], Mineke Viaene, MD, PhD[29,30], Beatriz Abril, MD[32], Bradley F. Boeve, MD[33], Guy A. Rouleau, MD, PhD, FRCPC, FRSC[1,2,3], Ronald B. Postuma, MD, MSc[2,3,9], Sonja W. Scholz, MD PhD[4,34] for the International LBD Genomics Consortium, Ziv Gan-Or, MD, PhD[1,2,3].

Affiliations

1. Department of Human Genetics, McGill University, Montréal, QC, Canada.
2. The Neuro (Montreal Neurological Institute-Hospital), McGill University, Montréal, QC, Canada.
3. Department of Neurology and Neurosurgery, McGill University, Montréal, QC, Canada.
4. Neurodegenerative Diseases Research Unit, National Institute of Neurological Disorders and Stroke, Bethesda, MD, USA.
5. Neuromuscular Diseases Research Section, National Institute on Aging, Bethesda, MD, USA.
6. Sleep Disorders Unit, Pitié Salpêtrière Hospital, Paris Brain Institute and Sorbonne University, Paris, France.
7. Oxford Parkinson's Disease Centre (OPDC), University of Oxford, Oxford, United Kingdom.
8. Division of Neurology, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, United Kingdom.
9. Center for Advanced Research in Sleep Medicine, Centre Intégré Universitaire de Santé et de Services Sociaux du Nord-de-l'Île-de-Montréal – Hôpital du Sacré-Coeur de Montréal, Montréal, QC, Canada.
10. Department of Psychiatry, Université de Montréal, Montréal, QC, Canada.
11. Department of Psychology, Université du Québec à Montréal, Montréal, QC, Canada.
12. Department of Neurosciences, Université de Montréal, Montréal, QC, Canada.
13. National Reference Center for Narcolepsy, Sleep Unit, Department of Neurology, Gui-de-Chauliac Hospital, CHU Montpellier, University of Montpellier, Inserm U1061, Montpellier, France.
14. Clinical Neurology Unit, Department of Neurosciences, University Hospital of Udine, Udine, Italy.

15. Department of Medicine (DAME), University of Udine, Udine, Italy.
16. Sleep Disorders Clinic, Department of Neurology, Medical University of Innsbruck, Innsbruck, Austria.
17. Department for Sleep Medicine and Neuromuscular disease, University Hospital Muenster, Muenster, Germany.
18. Department of Neurology and Centre of Clinical Neuroscience, Charles University, First Faculty of Medicine and General University Hospital, Prague, Czech Republic.
19. Department of Neurology, Philipps University, Marburg, Germany.
20. Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio-Emilia, Modena, Italy.
21. IRCCS, Institute of Neurological Sciences of Bologna, Bologna, Italy.
22. Neurology Unit, Movement Disorders Division, Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona, Verona, Italy.
23. Department of Medical Sciences and Public Health, Sleep Disorder Research Center, University of Cagliari, Cagliari, Italy.
24. Paracelsus-Elena-Klinik, Kassel, Germany.
25. Department of Neurosurgery, University Medical Centre Göttingen, Göttingen, Germany.
26. Sleep and Neurology Unit, Beau Soleil Clinic, Montpellier, France.
27. EuroMov Digital Health in Motion, University of Montpellier IMT Mines Ales, Montpellier, France.
28. Department of Neurological Sciences, Università Vita-Salute San Raffaele, Milan, Italy.
29. Laboratory for Sleep Disorders, St. Dimpna Regional Hospital, Geel, Belgium.
30. Department of Neurology, St. Dimpna Regional Hospital, Geel, Belgium.
31. Department of Neurology, University Hospital Antwerp, Edegem, Antwerp, Belgium.
32. Sleep disorder Unit, Carémeau Hospital, University Hospital of Nîmes, France.
33. Department of Neurology, Mayo Clinic, Rochester, MN, USA.
34. Department of Neurology, Johns Hopkins University Medical Center, Baltimore, MD, USA.

Preprint: https://doi.org/10.1101/2023.01.31.23284682

Abstract

Background and Objectives

Isolated/idiopathic REM sleep behavior disorder (iRBD) and Lewy body dementia (LBD) are synucleinopathies that have partial genetic overlap with Parkinson's disease (PD). Previous studies have shown that neuroinflammation plays a substantial role in these disorders. In PD, specific residues of the human leukocyte antigen (*HLA*) were suggested to be associated with a protective effect. This study examined whether the *HLA* locus plays a similar role in iRBD, LBD and PD.

Methods

We performed HLA imputation on iRBD genotyping data (1,072 patients and 9,505 controls) and LBD whole-genome sequencing (2,604 patients and 4,032 controls) using the multi-ethnic HLA reference panel v2 from the Michigan Imputation Server. Using logistic regression, we tested the association of HLA alleles, amino acids and haplotypes with disease susceptibility. We included age, sex and the top 10 principal components as covariates. We also performed an omnibus test to examine which HLA residue positions explain the most variance.

Results

In iRBD, *HLA-DRB1**11:01 was the only allele passing FDR correction (OR=1.57, 95% CI=1.27-1.93, p=2.70e-05). We also discovered associations between iRBD and *HLA-DRB1* 70D (OR=1.26, 95%CI=1.12-1.41, p=8.76e-05), 70Q (OR=0.81, 95% CI=0.72-0.91, p=3.65e-04) and 71R (OR=1.21, 95% CI=1.08-1.35, p=1.35e-03). In *HLA-DRB1*, position 71 ($p_{omnibus}$=0.00102) and 70 ($p_{omnibus}$=0.00125) were associated with iRBD. We found no association in LBD.

Discussion

This study identified an association between *HLA-DRB1* 11:01 and iRBD, distinct from the previously reported association in PD. Therefore, the *HLA* locus may play different roles across

synucleinopathies. Additional studies are required better to understand HLA's role in iRBD and LBD.

Introduction

Isolated/idiopathic REM sleep behavior disorder (iRBD) is a prodromal synucleinopathy characterized by enactment of dreams, vocalization and absence of muscle atonia during REM sleep.[1] iRBD is one of the strongest predictors for certain neurodegenerative disorders, as approximately 80% of patients will convert to Parkinson's disease (PD), Lewy body dementia (LBD) or multiple system atrophy (MSA) after 10-15 years on average following iRBD diagnosis.[2]

Previous evidence has shown that iRBD and synucleinopathies share a partial genetic overlap.[3] While specific loci (*SNCA, GBA, TMEM175*) were shared between these traits, distinct loci such as *LRRK2* and *MAPT* for PD and *APOE* LBD were also identified.[3] Furthermore, while the *SNCA* locus is important in PD, LBD and iRBD, the association with *SNCA* is driven by different variants for the different traits.[3] Similar phenomenon occurs in the *SCARB2* locus, where different variants are associated with PD or RBD.[3] Understanding the shared genes and pathways and the genetic differences will lead to better characterization of these disorders. For instance, microglial activation, a form of neuroinflammation, was found in all these disorders.[4-6] However, the role of the immune system in their pathophysiology is poorly understood.

Recently, a fine-mapping study of the human leukocyte antigen (*HLA*) locus in PD demonstrated a strong association of HLA-DRB1 amino acids 11V, 13H and 33H with reduced PD risk.[7] Located on chromosome 6, the *HLA* locus is a highly polymorphic region with complicated linkage patterns. *HLA* plays an essential role in the adaptive immune system by presenting antigens to T-cells.

Since the role of the *HLA* locus is unknown in iRBD and LBD, this study aims to examine whether *HLA* variants may affect the risk for these disorders. We analyzed the association of different *HLA* alleles, haplotypes and amino acids in two cohorts of iRBD and LBD patients.

Table 1: Study population after quality control.

Variable	Isolated REM sleep behavior disorder		Lewy body dementia	
	Patients (n = 1,072)	Controls (n = 9,505)	Patients (n = 2,604)	Controls (n = 4,032)
Age (years), (SD)	60.54 (11.06)	63.49 (16.59)	74.36 (11.76)	72.63 (16.99)
Male, number (%)	860 (80.22)	4824 (50.75)	1656 (63.59)	1967 (48.78)

SD, standard deviation; n, number

Methods

Study population

iRBD and LBD cohorts from two previous genome-wide association studies (GWAS) were included in this analysis (Table 1).[3,8] iRBD patients were diagnosed according to the International Classification of Sleep Disorders (2nd or 3rd Edition) with video polysomnography. LBD was diagnosed according to consensus criteria, as described elsewhere.[8-10] The iRBD cohort is composed of 1,072 patients and 9,505 controls with genotyping data from the OmniExpress GWAS chip (Illumina inc.). The control group includes six publicly available cohorts: controls from the International Parkinson's Disease Genomics Consortium (IPDGC) NeuroX dataset (dbGap phs000918.v1.p1), National Institute of Neurological Disorders and Stroke (NINDS) Genome-Wide genotyping in Parkinson's Disease (dbGap phs000089.v4.p2), NeuroGenetics Research Consortium (NGRC) (dbGap phs000196.v3.p1), Parkinson's Progression Markers Initiative (PPMI) and Vance (dbGap phs000394).

The LBD cohort consisted of 2,604 patients and 4,032 controls with whole-genome sequencing data as described elsewhere.[8] Study participants signed informed consent forms and the Institutional Review Board at McGill University approved the study protocol.

Quality control

We performed standard GWAS quality control steps for both cohorts using PLINK v1.90. We excluded variants that were heterozygosity outliers (|F| > 0.15), sample call rate outliers (<0.95) and samples failing sex checks were also excluded. We determined genetic ancestry by merging samples with HapMap3 and clustering with principal components analysis (PCA). We only selected samples of European ancestry. A relatedness check was performed with GCTA[11] to remove third-degree relatives or closer ones. Then, we performed several variant-level filtrations, such as removing call rate outliers (<0.95) and variants with significantly different missingness between cases and controls ($p<0.0001$). We also excluded variants that failed PLINK –test-mishap ($p<0.0001$) and deviated from Hardy-Weinberg equilibrium ($p<0.0001$) in controls.

HLA imputation

Samples passing quality control were imputed on the Michigan Imputation Server with the four-digit multi-ethnic HLA reference panel v2[12] using Minimac4 and phased with Eagle v2.4. This reference panel is composed of five global populations (n=20,349). Only alleles with imputation score (r^2) above 0.8 were included. We determined *HLA* haplotypes using haplo.stats R package (https://analytictools.mayo.edu/research/haplo-stats/), which employs an Expectation–maximization (EM) algorithm.

Power calculations

We performed power calculations online for each cohort using CaTS to compute statistical power. (https://csg.sph.umich.edu/abecasis/gas_power_calculator/). We assumed a prevalence of 1% for

iRBD[13] and 4% for LBD[14]. In iRBD, we had enough statistical power (>0.8) to detect an association (p=0.0005) with an odd ratio of 1.6 with a minor allele frequency (MAF) of 0.05. In LBD, we had enough statistical power (>0.8) to detect an association (p=0.0005) with an odd ratio of 1.4 with a MAF of 0.05.

Statistical analysis

We performed logistic regression with an additive model on each *HLA* allele, haplotype and amino acid after adjusting for age at onset, sex and the top 10 principal components. We also performed an Omnibus test using the OMNIBUS_LOGISTIC module from HLA-TAPAS.[12] All rare associations (carrier frequency < 1%) were excluded. A 5% false-discovery rate (FDR) for multiple testing was applied.

Data availability

Anonymized data not published within this article will be made available by request from any qualified investigator.

Code availability

All scripts used in this study can be found at https://github.com/gan-orlab/HLA_syn.

Results

After *HLA* imputation, we examined the association of *HLA* alleles, haplotypes and amino acids. *HLA-DRB1**11:01 was the only allele passing FDR correction (OR=1.57, 95% CI=1.27-1.93, p=2.70e-05, Table 2). In addition, *HLA-DRB1* 70D, an amino acid present in *DRB1**11:01, was associated with iRBD (OR=1.26, 95%CI=1.12-1.41, p=8.76e-05). We also found association with 70Q (OR=0.81, 95% CI=0.72-0.91, p=3.65e-04 and 71R (OR=1.21, 95% CI=1.08-1.35, p=1.35e-03). In *HLA-DRB1*, positions 71 ($p_{omnibus}$=0.00102) and 70 ($p_{omnibus}$=0.00125) were the most

associated with iRBD. *DRB1**11:01 also tags three haplotypes: *DQA1**05:01~*DQB1**03:01~*DRB1**11:01 (OR=1.40, 95%CI=1.16-1.70, *p*=5.17e-04), *DQA1**05:01~*DRB1**11:01 (OR=1.41, 95%CI=1.16-1.72, *p*=5.43e-04), *DQB1**03:01~*DRB1**11:01 (OR=1.36, 95%CI=1.13-1.64, *p*=1.04e-03).

Table 2: HLA association in isolated REM sleep behavior disorder

	MAF in cases	**MAF in controls**	**OR**	**95% CI**	***p***	***p* (FDR)**
Alleles						
*HLA-DRB1**11:01	0.0726	0.0472	1.57	1.27-1.93	2.70e-05	2.75e-03
Amino acids						
HLA-DRB1 70D	0.505	0.444	1.26	1.12-1.41	8.76e-05	2.09e-02
HLA-DRB1 70Q	0.440	0.503	0.81	0.72-0.91	3.65e-04	4.41e-02
HLA-DRB1 71R	0.545	0.496	1.21	1.08-1.35	1.35e-03	4.41e-02
Haplotype						
*DQA1**05:01~*DQB1**03:01~*DRB1**11:01	0.0924	0.0657	1.40	1.16-1.70	5.17e-04	2.85e-02
*DQA1**05:01~*DRB1**11:01	0.0933	0.0652	1.41	1.16-1.72	5.43e-04	2.85e-02
*DQB1**03:01~*DRB1**11:01	0.0989	0.0707	1.36	1.13-1.64	1.04e-03	3.64e-02

MAF, minor allele frequency; OR, odds ratio; CI, confidence interval; *p*, p-value; FDR, false discovery rate for each group

When we repeated the analysis at one-field (two-digit) resolution, e.g., treating *DRB1**11:01 and 11:04 as the same, the association of *DRB1**11 was not significant (*p*=0.004, Supplementary Table #1), suggesting that it is specifically the *DRB1**11:01 allele associated with RBD. For LBD, no association was statistically significant after correction for multiple comparisons. We also examined the association of HLA-DRB1 33H, which was previously reported to be associated with PD (Supplementary Table #3).[7] The MAFs of HLA-DRB1 33H in iRBD cases and controls were 0.125 vs. 0.149, respectively (*p*=0.499). Meanwhile, the DRB1 33H allele frequency in both LBD cases and its controls was 0.145.

Discussion

This study shows an association between *DRB1**11:01, DRB1 70D, 70Q and 71R on iRBD. We also identified HLA-DRB1 positions 71 and 70 via an omnibus test, which suggests that residues at those positions explain a large amount of variance. HLA-DRB1 position 70-74 is a strong risk factor for rheumatoid arthritis and is referred to as a "shared epitope" (SE).[15] The SE, in combination with DRB1 11V, was associated with a protective effect for PD.[16] The SE is composed of a Q/R-K/R-RAA sequence with important antigen-binding grooves. However, 11:01 does not have the SE and there was no association between alleles with the SE (01:01, 01:02, 04:01, 04:04, 04:05, 04:08, 10:01)[16] and iRBD. These findings indicate that the effects of position 70 and 71 may be independent of the SE. Additional studies examining the role of HLA-DRB1 in PD and iRBD will be necessary.

In addition, DRB1 33H, a variant also associated with PD, was not significantly associated with iRBD or LBD. However, the difference in carrier frequency between iRBD cases and controls for DRB1 33H, similar to that seen in PD, suggests that our study may lack the power to detect this association in iRBD. A recent study has suggested a shared mechanism between PD, AD, amyotrophic lateral sclerosis and HLA-DRB1*04, harboring the 33H amino acid change.[17] This subtype was associated with decreased neurofibrillary tangles in post-mortem brains. It also binds to a K311 acetylated Tau PHF6 sequence.[17] These results exemplify the possibility of different HLA types with specific genetic variants that may affect the binding of substrates relevant for neurodegenerative disorders and activating inflammatory response.

We could not replicate the association of a previous study of HLA antigens with 25 iRBD cases. This study showed a significant association between iRBD and DQB1*05 and DQB1*06.[18] The most likely explanation for the discrepancy is that the previous study had reduced power to

detect a true effect. Another study has suggested that HLA-DR expression was associated with iRBD.[19] Fine-mapping and colocalization studies for these findings will be required once larger datasets of iRBD become available. Whether the mechanism underlying the associations with PD and iRBD is through functional effects of specific amino acid changes or due to different expressions of *HLA* genes in various brain tissues is still to be determined.

Although the role of the immune system in synucleinopathies is still unclear, some potential mechanisms of effect may exist. The varying effects of HLA between prodromal and clinical stages could be associated with HLA presenting different antigens in different brain regions. In LBD [20] and iRBD [21], activated CD4+ T-cells were shown to be dysregulated and associated with neuronal damage.

Another possibility is that the varying effects between iRBD and PD originate in the gastrointestinal tract.[22] For example, constipation, a common symptom in the early stages of PD, can aggravate or be caused by gut inflammation. In iRBD patients, one study showed a prevalence of constipation between 18-41%.[22] Gut bacterial antigens can be exposed from aging-related depletion of the gut lining. [23] HLA alleles may induce an immune response to self-proteins from these antigens.

Our study has several limitations. First, future replication studies with larger cohorts would be needed to increase statistical power since we do not have a replication cohort. Note that we used the most extensive available cohorts for iRBD and LBD.[3,8] Due to the polygenicity of the HLA locus, various populations have different HLA allele frequencies. This study was done only on samples with European ancestry, and multi-ancestry analysis could provide more refined evidence on the role of HLA in synucleinopathies. The cohorts used in the study were also not matched for age and sex. However, we adjusted for these variables in the analysis.

To conclude, we found an alternative *HLA* association of iRBD compared to PD and LBD. More experimental evidence is necessary to characterize the genetic landscape of synucleinopathies and the role of the immune system.

Reference

1. Dauvilliers Y, Schenck CH, Postuma RB, et al. REM sleep behaviour disorder. *Nature reviews Disease primers*. 2018;4(1):1-16.

2. Postuma RB, Iranzo A, Hu M, et al. Risk and predictors of dementia and parkinsonism in idiopathic REM sleep behaviour disorder: a multicentre study. *Brain*. Mar 1 2019;142(3):744-759. doi:10.1093/brain/awz030

3. Krohn L, Heilbron K, Blauwendraat C, et al. Genome-wide association study of REM sleep behavior disorder identifies polygenic risk and brain expression effects. *Nature Communications*. 2022;13(1):1-16.

4. Stokholm MG, Iranzo A, Østergaard K, et al. Assessment of neuroinflammation in patients with idiopathic rapid-eye-movement sleep behaviour disorder: a case-control study. *The Lancet Neurology*. 2017/10/01/ 2017;16(10):789-796. doi:https://doi.org/10.1016/S1474-4422(17)30173-4

5. Hirsch EC, Vyas S, Hunot S. Neuroinflammation in Parkinson's disease. *Parkinsonism & related disorders*. 2012;18:S210-S212.

6. Amin J, Holmes C, Dorey RB, et al. Neuroinflammation in dementia with Lewy bodies: a human post-mortem study. *Translational psychiatry*. 2020;10(1):1-11.

7. Yu E, Ambati A, Andersen MS, et al. Fine mapping of the HLA locus in Parkinson's disease in Europeans. *npj Parkinson's Disease*. 2021;7(1):1-7.

8. Chia R, Sabir MS, Bandres-Ciga S, et al. Genome sequencing analysis identifies new loci associated with Lewy body dementia and provides insights into its genetic architecture. *Nat Genet*. Mar 2021;53(3):294-303. doi:10.1038/s41588-021-00785-3

9. Emre M, Aarsland D, Brown R, et al. Clinical diagnostic criteria for dementia associated with Parkinson's disease. *Mov Disord*. Sep 15 2007;22(12):1689-707; quiz 1837. doi:10.1002/mds.21507

10. McKeith IG, Boeve BF, Dickson DW, et al. Diagnosis and management of dementia with Lewy bodies: Fourth consensus report of the DLB Consortium. *Neurology*. 2017;89(1):88-100.

11. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. Jan 7 2011;88(1):76-82. doi:10.1016/j.ajhg.2010.11.011

12. Luo Y, Kanai M, Choi W, et al. A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. *Nature Genetics*. 2021/10/01 2021;53(10):1504-1516. doi:10.1038/s41588-021-00935-7

13. Haba-Rubio J, Frauscher B, Marques-Vidal P, et al. Prevalence and determinants of rapid eye movement sleep behavior disorder in the general population. *Sleep*. 2017;41(2)doi:10.1093/sleep/zsx197

14. Kane JPM, Surendranathan A, Bentley A, et al. Clinical prevalence of Lewy body dementia. *Alzheimer's Research & Therapy*. 2018/02/15 2018;10(1):19. doi:10.1186/s13195-018-0350-6

15. Gregersen PK, Silver J, Winchester RJ. The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis Rheum*. Nov 1987;30(11):1205-13. doi:10.1002/art.1780301102

16. Hollenbach JA, Norman PJ, Creary LE, et al. A specific amino acid motif of HLA-DRB1 mediates risk and interacts with smoking history in Parkinson's disease. *Proc Natl Acad Sci U S A.* Apr 9 2019;116(15):7419-7424. doi:10.1073/pnas.1821778116

17. Guen YL, Luo G, Ambati A, et al. Protective association of HLA-DRB1*04 subtypes in neurodegenerative diseases implicates acetylated Tau PHF6 sequences. *medRxiv*. 2021:2021.12.26.21268354. doi:10.1101/2021.12.26.21268354

18. Schenck CH, Garcia-Rill E, Segall M, Noreen H, Mahowald MW. HLA class II genes associated with REM sleep behavior disorder. *Annals of neurology*. 1996;39(2):261-263.

19. Farmen K, Nissen SK, Stokholm MG, et al. Monocyte markers correlate with immune and neuronal brain changes in REM sleep behavior disorder. *Proceedings of the National Academy of Sciences*. 2021;118(10):e2020858118.

20. Gate D, Tapp E, Leventhal O, et al. CD4⁺ T cells contribute to neurodegeneration in Lewy body dementia. *Science*. 2021;374(6569):868-874. doi:doi:10.1126/science.abf7266

21. De Francesco E, Terzaghi M, Storelli E, et al. CD4+ T-cell Transcription Factors in Idiopathic REM Sleep Behavior Disorder and Parkinson's Disease. *Movement Disorders*. 2021;36(1):225-229. doi:https://doi.org/10.1002/mds.28137

22. Knudsen K, Fedorova TD, Hansen AK, et al. Objective intestinal function in patients with idiopathic REM sleep behavior disorder. *Parkinsonism & Related Disorders*. 2019/01/01/ 2019;58:28-34. doi:https://doi.org/10.1016/j.parkreldis.2018.08.011

23. Negi S, Singh H, Mukhopadhyay A. Gut bacterial peptides with autoimmunity potential as environmental trigger for late onset complex diseases: In–silico study. *PloS one*. 2017;12(7):e0180518.

Preface to Chapter 5

In Chapter 5, I nominated candidate genes within each PD GWAS locus. While current PD studies often focus on well-established genes, this study takes a different approach by utilizing a machine learning model to prioritize genes from lesser-studied loci.

To accomplish this, I collected and integrated various types of data, including genetic, transcriptomic, and epigenetic information. This diverse dataset was used to train a machine learning model capable of predicting the most likely candidate genes within the PD GWAS loci. By combining multi-omic data, I aimed to improve the accuracy and reliability of the predictions.

Post hoc analyses were conducted to validate the identified candidate genes and pathways. These analyses involved assessing the functional relevance of the nominated genes, investigating their expression patterns, and exploring their potential involvement in biological pathways and processes associated with PD. By performing these additional analyses, I aimed to strengthen the validity of the nominated genes and provide further insights into their potential roles in PD.

By incorporating machine learning techniques and integrating various omics data, this study aimed to shed light on lesser-studied genes within the PD GWAS loci. The identification and validation of candidate genes and pathways within these loci have the potential to expand our understanding of the underlying mechanisms of PD and contribute to the development of novel therapeutic targets.

It is worth noting that the success and significance of the study's findings will depend on the quality and representativeness of the data used for training the machine learning model, as well as the rigor of the post hoc analyses conducted to validate the nominated genes and pathways.

Chapter 5: Identification of novel variants, genes and pathways potentially linked to Parkinson's disease using machine learning.

Eric Yu, BSc[1,2], Roxanne Larivière, PhD[3,4], Rhalena A. Thomas, PhD[3,4], Lang Liu, MBI[1,2], Konstantin Senkevich, MD, PhD[2,4], Shady Rahayel, PhD[2,5], Jean-Francois Trempe, MD, PhD[6], Edward A. Fon, MD, FRCPC[3,4], Ziv Gan-Or, MD, PhD[* 1,2,4].

1. Department of Human Genetics, McGill University, Montreal, Quebec, Canada.
2. The Neuro (Montreal Neurological Institute-Hospital), Montreal, Quebec, Canada.
3. Early Drug Discovery Unit (EDDU), Montreal Neurological Institute-Hospital (The Neuro), Montreal, Quebec, Canada.
4. Department of Neurology and Neurosurgery, McGill University, Montreal, Quebec, Canada.
5. Centre for Advanced Research in Sleep Medicine, Hôpital du Sacré-Cœur de Montréal, Montreal, Quebec, Canada
6. Department of Pharmacology and Therapeutics and Centre de Recherche en Biologie Structurale, McGill University, Montreal, Quebec, Canada.

Abstract

There are 78 loci associated with Parkinson's disease (PD) in the most recent genome-wide association study (GWAS), yet the specific genes driving these associations are mostly unknown. Herein, we aimed to nominate the top candidate gene from each PD locus, and identify variants and pathways potentially involved in PD. We trained a machine learning model to predict PD-associated genes from GWAS loci using genomic, transcriptomic, and epigenomic data from brain tissues and dopaminergic neurons. We nominated candidate genes in each locus, identified novel pathways potentially involved in PD, such as the inositol phosphate biosynthetic pathway (*INPP5F*, *IP6K2*, *ITPKB*, *PPIP5K2*). Specific common coding variants in *SPNS1* and *MLX* may be involved in PD, and burden tests of rare variants further support that *CNIP3*, *LSM7*, *NUCKS1* and the polyol biosynthetic pathway are associated with PD. Functional studies are needed to further study the involvements of these genes and pathways in PD.

Introduction

Genome-wide association studies (GWAS) have nominated many variants associated with complex traits. In Parkinson's disease (PD), the most recent GWAS revealed 90 independent risk variants across 78 genomic loci. [1] Although many single-nucleotide polymorphisms (SNPs) are in novel genomic loci, well-established PD genes discovered decades ago, such as *LRRK2*, *PINK1*, *PARK7*, *SNCA*, *GBA1*, *PRKN* and *MAPT* still account for the vast majority of research on Parkinson's disease.

Several disadvantages of GWAS limit additional functional analyses. First, above 90% of GWAS significant SNPs are in noncoding regions. [2] These SNPs are often passenger variants due to complex linkage disequilibrium (LD). Second, the causal gene associated with the causal SNPs remains unclear in most GWAS loci. Downstream GWAS analyses were developed to overcome

these challenges by prioritizing causal genes at GWAS loci. For example, fine-mapping and colocalization methods aim to nominate causal SNPs, and methods such as transcriptome-wide association studies nominate gene-trait association. These models use LD structure, and gene expression panels to discover causal SNP/genes.[3-5] Although these methods suggest causal variants and genes, additional biological evidence is often required to pair causal variants with causal genes. Multi-omic analyses can integrate a diverse range of comprehensive cellular and biological datasets such as genomic, transcriptomic and epigenetic datasets. Using this approach, platforms such as Open Target Genetics performed systematic analyses of gene prioritization across all publicly available GWASs.[6] However, Open Target Genetics lacks relevant disease-specific tissues such as dopaminergic neurons and microglia for PD. Using a similar approach, we may discover additional pathways and genetic targets involved in PD.

In this study, we leveraged disease-relevant multi-omic datasets relevant to PD in our machine-learning model (Figure 1). We trained this model on well-established PD genes to nominate causal genes from PD GWAS loci.

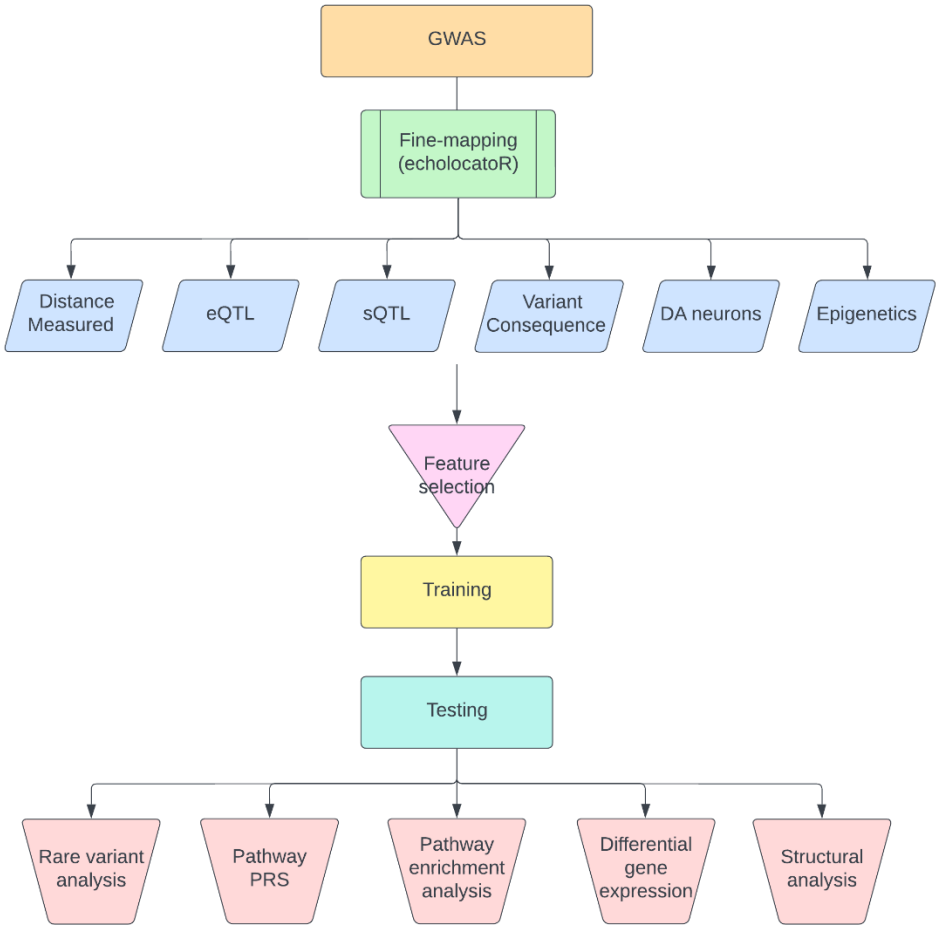


Figure 1: Workflow summary.

This figure describes the analyses performed in this study.

Results

Machine learning model nominates PD-associated genes in each PD locus

We used well-established PD-associated genes from the PD GWAS (*GBA1*, *LRRK2*, *SNCA*, *GCH1*, *MAPT*, *TMEM175*, *VPS13C*) as positive labels, and the other genes from the same loci (n=205) were used as negative labels (i.e. genes that are unlikely to be involved in PD). The machine learning model identified the best predictive features, and then each gene received a probability score to be the gene driving the association in each locus (Supplementary Table 1). Overall, after

removing the redundant features, the model performance increased from 0.66 to 0.82. We used average precision as an evaluation function to maximize the sensitivity of the model. We then nominated the top-scoring genes in each locus (Supplementary Table 1, Figure 2). Two genes, *MAPT* and *TOX3*, were nominated twice in neighboring loci that harbor them, bringing the total number of genes nominated in this model to 76 genes in 78 loci. 48 of the 76 genes (63%) had a probability score higher than 0.75. Of note, five genes (*NEK1*, *FDFT1*, *PSD*, *BAG3* and *SLC2A13*) that were ranked second in their respective loci also had a probability score > 0.75. However, the nominated genes in their loci (*CLCN3*, *CTSB*, *GBF1*, *INPP5F* and *LRRK2*, respectively) all had probability scores >0.94. In seven other loci the top nominated genes had an especially low probability score (<0.3), including *RBMS3*, *HIST1H2BL*, *TRIM40*, *EHMT2*, *RPS12*, *MICU3* and *ITGA8*.

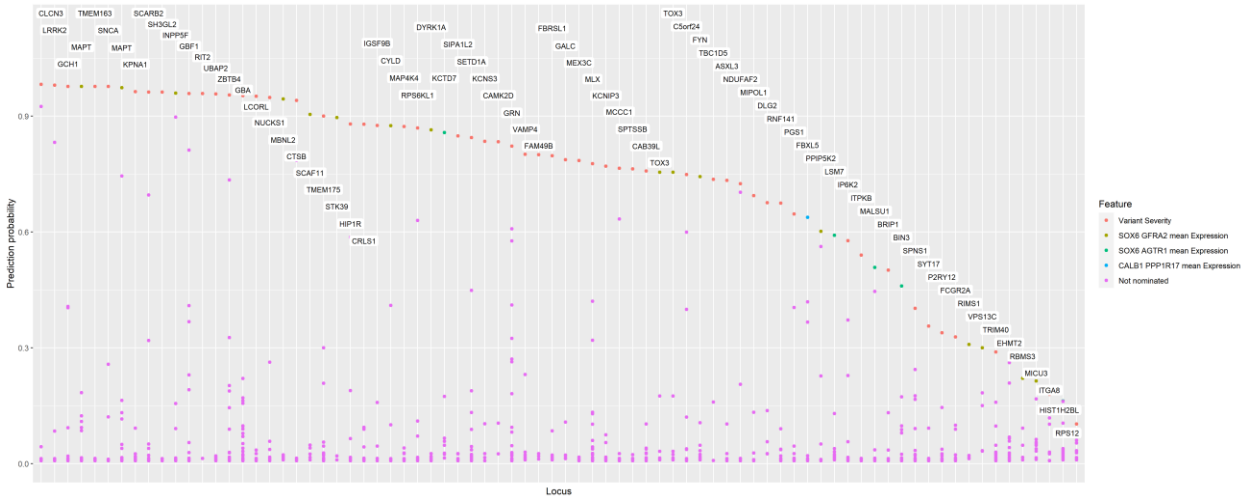


Figure 2: Probability score of the Parkinson's disease GWAS candidate genes

The probability score from the machine learning model for each locus in the Parkinson's disease sorted in descending order. For each gene, the top non-distance feature was used to color the data. Variant severity is calculated from the Variant Effect Predictor (VEP) IMPACT rating

which is based on the classification of the severity of the variant consequence. SOX6 GFRA2, SOX6 AGTR1, CALB1 PPP1R17 correspond to gene expression from subclusters nominated by Kamath et al.

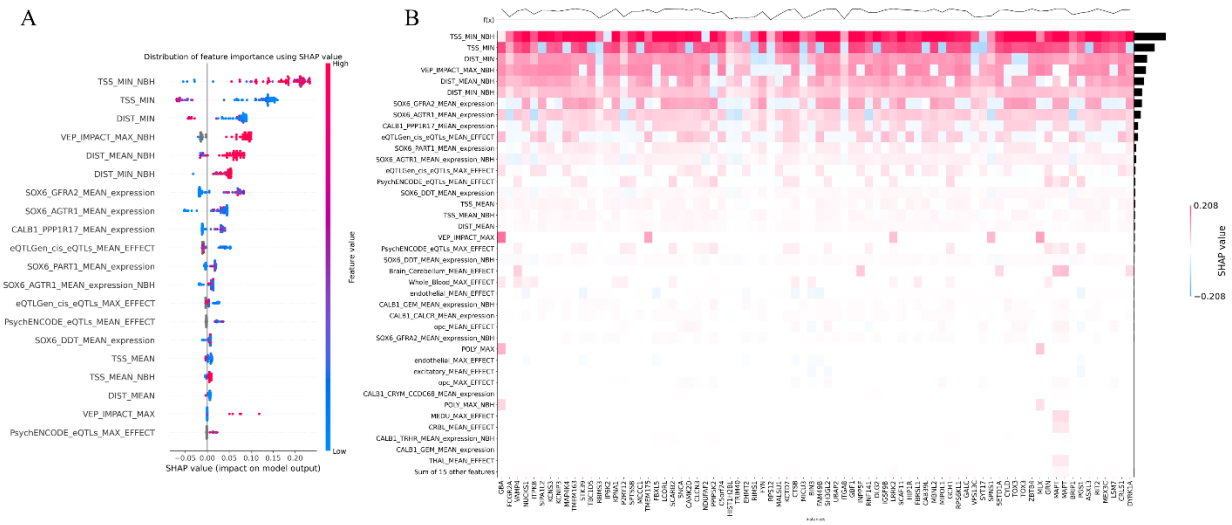


Figure 3: Feature importance for the Parkinson's disease GWAS gene prioritization model

A) Bee-swarm plot of feature importance using SHAP values along with the distribution of genes based on feature value B) Heatmap of feature importance using SHAP value for the top candidate gene in each locus. The plot at the top represents the probability score of each gene. The bar plot on the right shows the relative importance of each feature. Abbreviations for each feature can be found in Supplementary Table 2.

Gene expression in specific PD-associated dopaminergic neuron subtypes is an important feature predicting PD-relevant genes

Next, we sought to determine which features of the model contributed the most to the prediction, by using Shapley Additive exPlanations (SHAP) values.[7,8] SHAP values provide, for each gene, the relative contribution of each feature to the selection of that gene. The most important features for the scoring of each gene are shown in Figure 3A. Distance-related features, such as distance from the top-associated SNP in the locus to the transcription start site or distance to the beginning of the gene, were the most important features in our model, as expected.[6] Then, the next most important feature was the Variant Effect Predictor (VEP) value, followed by additional distance measures. Interestingly, the following top features were expression is a specific dopaminergic cell subtypes, marked by the expression of the genes *GFRA2* and *AGTR1*. The latter is a specific subtype of dopaminergic neurons shown by Kamath et al. to be selectively degenerated in brains of PD patients.[9] The remaining features include expression in other dopaminergic cell subpopulation, expression quantitative trait loci (eQTLs) and others. Epigenetic features were not predictive in our model. As shown in Figure 3B, all nominated genes had at least one of the distance features contributing to their selection. Among the candidate genes in each locus, missense SNPs contributed to the score of two candidate genes: *SPNS1* (p.L563V, rs7140) and *MLX* (p.Q139R, rs665268), on top of the known contribution of missense variants in *GBA1*, *LRRK2* and *GCH1*. *SPNS1* and *MLX* have not been previously implicated in PD, and the important features for these two genes are shown in Figure 4.

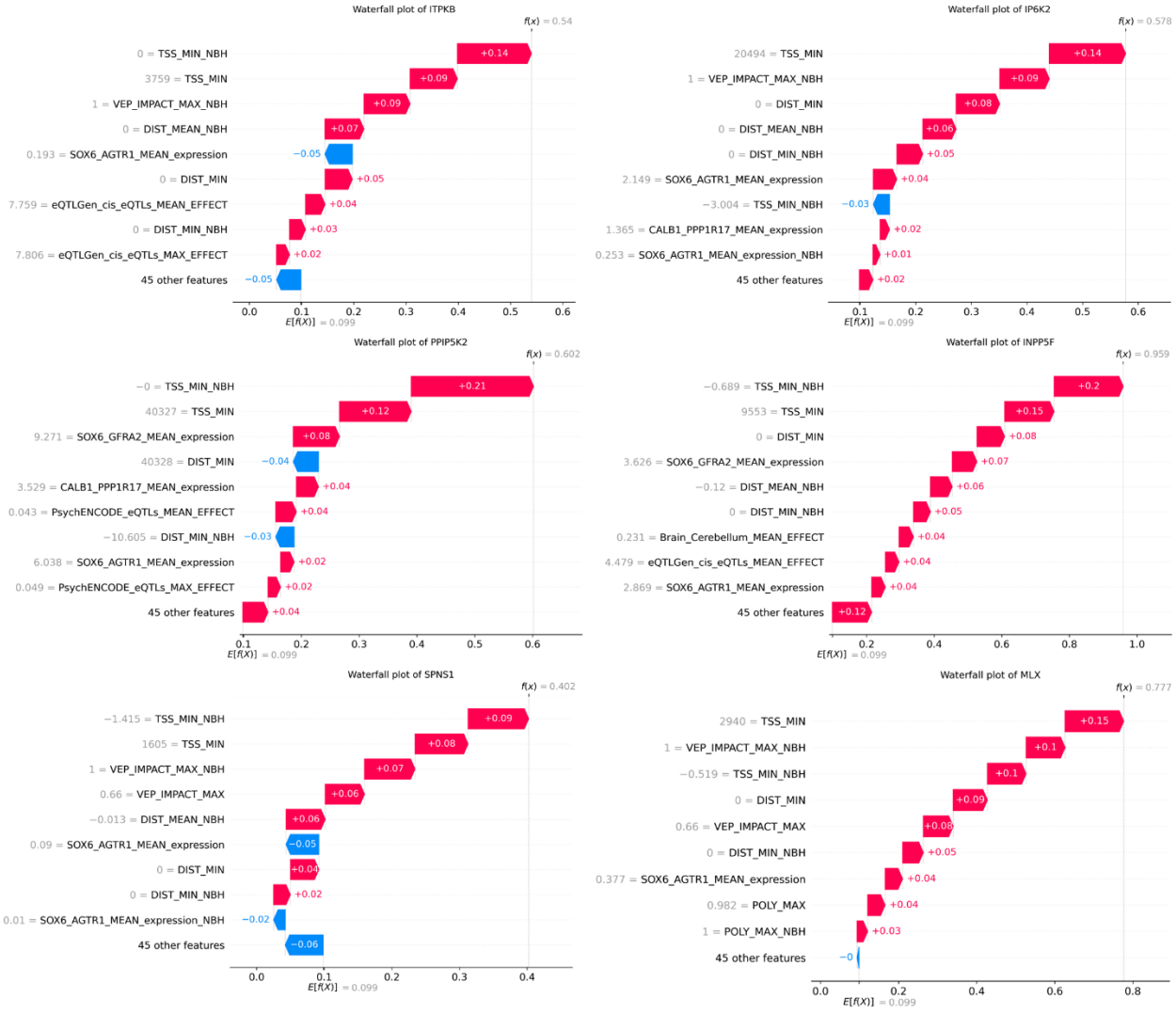


Figure 4: Waterfall plots for Parkinson's disease GWAS candidate genes

Importance of the top 10 features using SHAP values for different candidate genes. E[f(x)] is the base score for each gene. f(x) is the final score after accounting for all features. Abbreviations for each feature can be found in Supplementary Table 2.

Differential gene expression of genes from the inositol phosphate biosynthetic pathway and *MLX1* in PD

To further establish the importance of the nominated genes in PD, we examined whether they could be differentially expressed in PD, using expression data from single-cell RNAseq and bulk RNAseq datasets from Kamath et al[9] and FOUNDIN-PD.[10] of the genes of interest, *INPP5F* (average log fold change[FC] = −7.22, $p = 2.90e-31$) and *MLX* (average log FC = -1.80, $p = 2.23e-4$) were associated with PD in the data from Kamath et al (Supplementary Table 3).[9] In FOUNDIN-PD[10], we found differential expression of *INPP5F* (average log FC = 0.070, $p = 1.89e-19$) and *IP6K2* (average log FC = −0.076, $p = 1.35e-35$) in scRNA data from dopaminergic neurons by comparing PD and control (Supplementary Table 4). Results from the bulk RNAseq analysis can be found in Supplementary Table 5.

Structural analysis of *SPNS1* and *MLX*

Since nonsynonymous variants in *SPNS1* and *MLX* were identified as major contributors to their selection as the nominated genes in their loci, we aimed to examine the potential consequences of these variants by performing *in silico* structural analyses. *SPNS1* encodes a transporter for phospholipids at the lysosome membrane.[11] It mediates the efflux of lysophosphatidylcholine and lysophosphatidylethanolamine out of the lysosome. The SNP rs7140 is located in the 3'-untranslated region (UTR) of the canonical splice variant 1 transcript, which produces the 528 a.a. isoform that has been investigated functionally[11] (Uniprot #Q9H2V7). This canonical isoform has also been observed in numerous proteomics datasets in gpmDB (https://gpmdb.thegpm.org/index.html). However, six other potential isoforms generated by alternative splicing have been predicted, including a 538 a.a. fragment with an alternative C-terminus where the rs7140 SNP is located within the coding region (Uniprot #H3BR82). The

rs7140 variant results in the p.L512M mutation in this isoform. To determine the impact of this mutation on the function of this *SNPS1* isoform, we inspected the 3D structure model generated by AlphaFold.[12] Leu512 is located in the unstructured C-terminus of this membrane-bound protein, on the lumenal side of the membrane (Figure 5A). The role of the C-terminus in this isoform of *SPNS1* remains unclear, and thus the impact of the p.L512M mutation is unknown.

The Max-like protein (MLX) is at the heart of a transcriptional network pathway involved in energy metabolism and cell signalling.[13,14] It interacts with at least 6 other related proteins including the MAD family of transcriptional repressors and the Mondo family of transcriptional activators. These proteins contain basic/helix-loop-helix/leucine zipper (bHLHZ) domains that form heterodimers and interact with DNA carrying the CACGTG E-box motif. To understand the impact of the p.Q223R *MLX* mutation on its activity, we modeled the structure of MLX heterodimers with both the MAD and Mondo families using AlphaFold. MLX dimerizes with MAD1,[14] and thus we superposed its bHLHZ domain on the MAD1-MAX-DNA complex crystal structure[15] to generate the ternary complex models. The model shows that Gln223 in MLX is at the end of the dimerization "zipper" helix (Figure 5B). The mutation p.Q223R induces the formation of a salt bridge with Glu139 in MAD1, which could strengthen the interaction. This could then downregulate the interaction of MAD1 with MAX through competition, and thus affect the extent of the transcriptional repression. Glu139 is not conserved in other MAD-related proteins such as MXI1 and MAD3/4. Furthermore, the model of MLX interacting with MLXIP, a protein of the Mondo family also known as MondoA,[16] shows that the mutation may negatively affect the formation of this heterodimer by introducing a charge next to a hydrophobic sidechain (Figure 5C). The nuclear localization of Mondo proteins is dependent on their interaction with MLX,[13] and thus

the mutation may down regulate activation by the Mondo family while strengthening repression via MAD1.

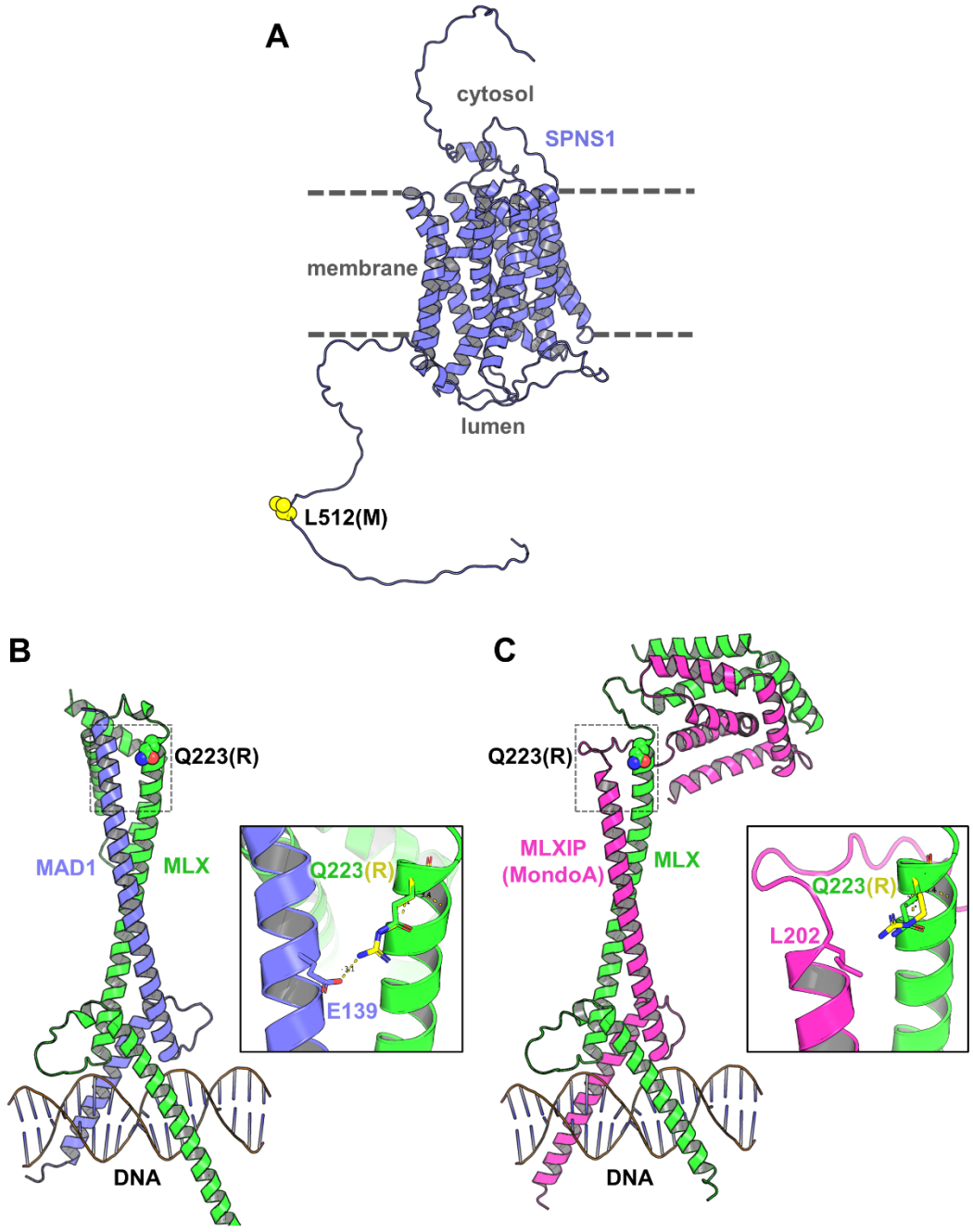


Figure 5: Structural analysis of SPNS1 p.L512M and MLX p.Q223R

A) Alphafold prediction of the structure of the lysophospholipid transporter SPNS1 (alternative isoform, Uniprot #H3BR82). The mutation p.L512M would take place in the lumen of the lysosome. B) AlphaFold model of the MAD1-MLX heterodimer superposed on the structure of the MAD1-MAX-DNA complex (PDB 1NLW). The inset is a zoom on the MLX p.Q223R mutation, displaying the effect that the mutation may have on the interaction with the MAD1 protein. C) AlphaFold model of the MLXIP-MLX heterodimer superposed on the structure of the MAD1-MAX-DNA complex, as described above. Note that AlphaFold also predicts an interaction between the C-termini of MLXIP and MLX (but not MAD1 and MLX).

Gene enrichment nominates the inositol phosphate biosynthetic pathway as a novel pathway involved in PD

We further aimed to examine if the nominated genes highlight specific pathways and mechanisms that may be involved in PD. We performed a pathway enrichment analysis by examining over-representation of the nominated genes in biological processes and cellular components using the top genes in each locus. Among the biological processes passing false discovery rate (FDR) correction, the inositol phosphate biosynthetic process (GO:0032958) and polyol biosynthetic process (GO:0046173) were strongly enriched (Figure 6A). Inositol is associated with 4 candidate genes: *ITPKB*, *IP6K2*, *PPIP5K2* and *INPP5F*. Feature importance of *ITPKB*, *IP6K2*, *PPIP5K2* and *INPP5F* are shown in Figure 4. Cellular components such as exocytic vesicle (GO:0070382), and dendritic tree (GO:0097447) were also identified in the gene enrichment analysis (Figure 6B).

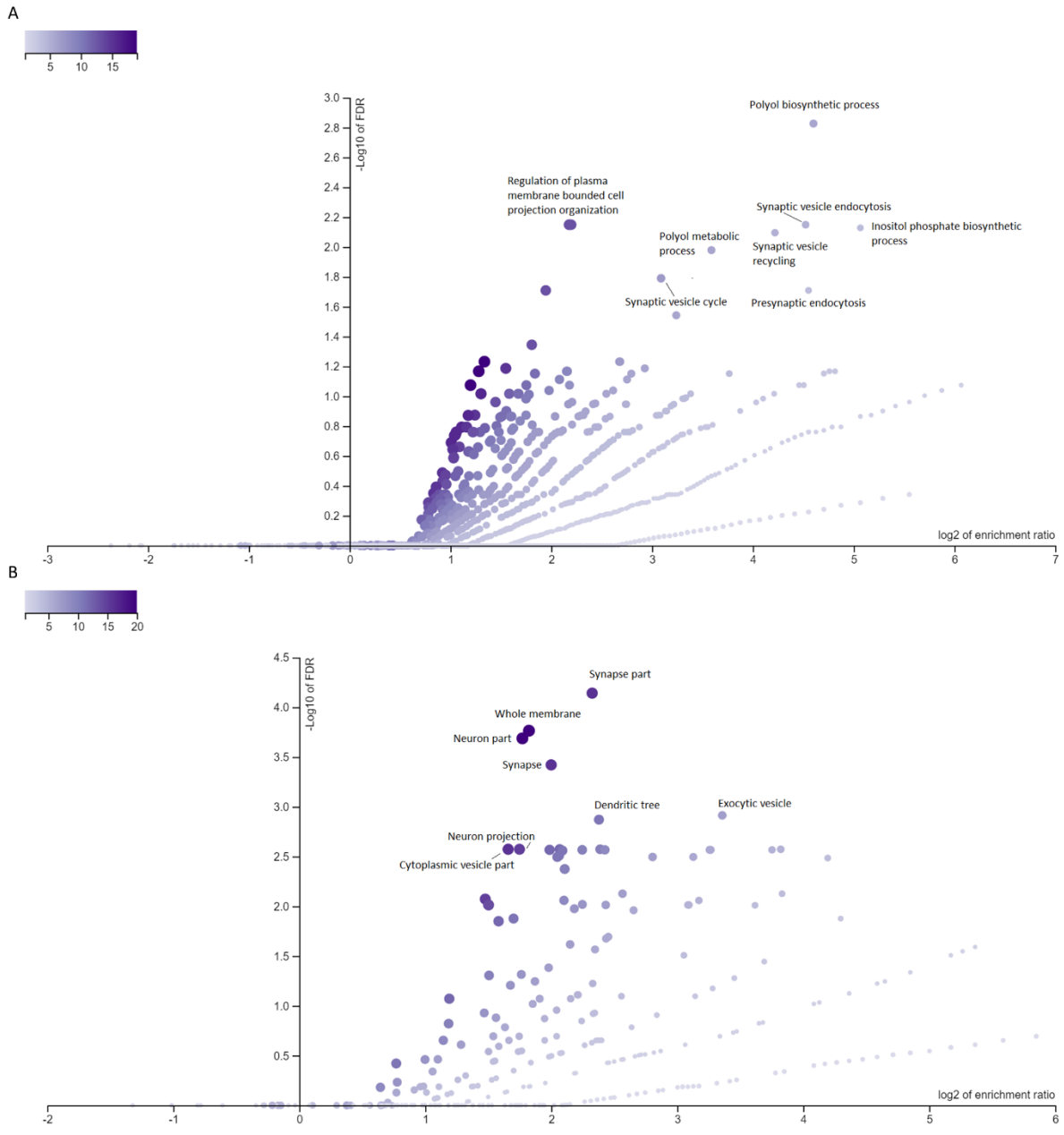


Figure 6: Volcano plots of Gene Ontology biological processes and cellular components.

Volcano plots of gene-set enrichment analysis using WebGestalt showing the log of the FDR versus the enrichment ratio. *P*-value are calculated using a hypergeometric test. All named pathways after significant after FDR correction.

Pathway specific polygenic risk score of the inositol phosphate biosynthetic pathway is associated with PD

To further study the association between the putative, novel PD pathways and PD status, pathway-specific polygenic risk scores (PRS) were calculated. The association between these PRS and PD was examined in six PD cohorts, followed by a meta-analysis as detailed in the Methods section. One outlier cohort was excluded due to heterogeneity. The pathway specific PRS were first calculated using all the genes in that pathway. Then, to further validate that the specific pathway is indeed important in PD, we excluded the genes nominated by our machine learning pathway and re-calculated the PRS. By removing these genes, that has GWAS significant signals, we could examine the residual effect of the remaining of the pathway. The inositol phosphate biosynthetic pathway was associated with PD even after excluding the genes nominated in our analysis (OR 1.06, 95% CI 1.03-1.09, p=7.01E-05), as well as other related pathways (Table 1). Forest plots of pathway specific PRS can be found in Supplementary Figure 1.

Table 1: Meta-analyzes of pathway-specific polygenic risk scores

Pathway-specific PRS	OR	95% CI	P	Het P
POLYOL_BIOSYNTHETIC_PROCESS	1.20	1.17-1.24	2.07E-42	1.91E-05
INOSITOL_PHOSPHATE_BIOSYNTHETIC_PROCESS	1.15	1.12-1.18	2.36E-25	1.97E-02
POLYOL_BIOSYNTHETIC_PROCESS_filtered	1.09	1.06-1.12	1.04E-09	1.12E-02
INOSITOL_PHOSPHATE_BIOSYNTHETIC_PROCESS_filtered	1.06	1.03-1.09	1.31E-05	1.45E-01

PRS: Polygenic risk score, OR: odds ratio, CI: confidence interval; P: p-value, Het: Heterogeneity, filtered: excluded Parkinson's disease top gene, GOBP_INOSITOL_PHOSPHATE_BIOSYNTHETIC_PROCESS: GeneOntology inositol phosphate biosynthetic process (GO:0032958), GOBP_POLYOL_BIOSYNTHETIC_PROCESS: GeneOntology polyol biosynthetic process (GO:0046173).

Table 2: Meta-analysis of rare variant analysis of putative causal genes

Set	P	FDR P
GBA_Rarefunctional	2.04E-12	6.22E-10
GBA_Rarenonsyn	3.38E-11	5.15E-09
GBA_RareLOF	1.22E-06	1.24E-04
GBA_RareCADD	2.32E-06	1.77E-04
LSM7_RareLOF	3.69E-06	2.25E-04
KCNIP3_RareLOF	1.12E-05	5.69E-04
GCH1_RareLOF	2.02E-05	8.80E-04
LRRK2_RareCADD	6.07E-05	2.31E-03
Polyol_Rarefunctional	1.59E-04	5.38E-03
Polyol_Rarenonsyn	2.86E-04	8.74E-03
NUCKS1_RareCADD	4.13E-04	1.14E-02
Polyol_RareLOF	1.54E-03	3.91E-02
SYT17_Rarenonsyn	4.61E-03	9.37E-02
P2RY12_RareLOF	4.38E-03	9.37E-02
CYLD_RareLOF	4.48E-03	9.37E-02
SYT17_Rarefunctional	7.39E-03	1.38E-01
LCORL_RareLOF	7.66E-03	1.38E-01
CAMK2D_RareLOF	8.62E-03	1.46E-01
FBRSL1_RareLOF	1.12E-02	1.80E-01
CTSB_RareLOF	1.20E-02	1.82E-01
KPNA1_RareCADD	1.35E-02	1.96E-01
ASXL3_RareLOF	1.52E-02	2.10E-01
KPNA1_RareLOF	1.76E-02	2.33E-01
LRRK2_Rarefunctional	2.57E-02	3.14E-01
MICU3_RareLOF	2.56E-02	3.14E-01
VAMP4_Rarenonsyn	2.93E-02	3.43E-01
MBNL2_RareCADD	3.04E-02	3.43E-01
LRRK2_Rarenonsyn	3.28E-02	3.57E-01
KPNA1_Rarefunctional	3.46E-02	3.64E-01
LSM7_Rarefunctional	3.58E-02	3.64E-01
HIP1R_Rarenonsyn	3.93E-02	3.87E-01
KPNA1_Rarenonsyn	4.23E-02	3.91E-01
HIP1R_Rarefunctional	4.22E-02	3.91E-01

Set: variant set across genes/pathway, P: p-value, FDR P: false discovery rate p-value, Rarefunctional: rare functional variants, Rarenonsyn: rare nonsynonymous variants, RareLOF: rare loss-of-function variants, RareCADD: rare variants with CADD score above 15.

Association of rare variants with nominated PD genes

In order to further establish the potential role of the nominated genes in PD, we performed rare variant burden tests in all the genes nominated by our model. As expected, genes that are known to harbor rare PD coding mutations including *GBA1*, *LRRK2* and *GCH1* were associated with PD (Table 2, Supplementary Table 6). Three additional genes, including two genes that have not been previously implicated in PD (*KCNIP3* and *LSM7*) showed burden of rare variants after FDR correction for multiple comparisons. We then examined the genes from the polyol biosynthetic pathway and found that rare variants in this pathway were also associated with PD (SKAT-O p=1.58E-04), further supporting its role in PD.

Discussion

In this study, we nominated genes that potentially drive the associations with PD for each of the 78 PD GWAS loci, using multi-omic data and machine learning. Our nominated genes include many genes that have not been studied in the context of PD, as well as genes with coding variants such as *SPNS1* and *MLX that could be further studied*. Furthermore, our gene enrichment, pathway specific PRS and rare variant analyses strongly support an involvement of the inositol phosphate biosynthetic pathway in PD.

Four genes nominated by our machine learning model belong to the inositol phosphate biosynthetic pathway: *ITPKB*, *IP6K2* and *PPIP5K2 and SNCA*,[17] showing a strong enrichment of this pathway. In addition, *INPP5F* is another gene nominated by our analysis that is involved in inositol processing through a parallel pathway. Our findings that the PRS of this pathway, even

after excluding the aforementioned genes, is associated with PD, and that rare variants in genes from this pathway are also associated with PD, provide additional support for the importance of this pathway in PD. *ITPKB* encodes for a ubiquitous kinase that phosphorylates inositol 1,4,5-trisphosphate (IP3) to inositol 1,3,4,5 tetrakisphosphate (IP4) using a Ca2+/Calmodulin-dependent mechanism. IP3 is a secondary messenger that stimulates calcium release from the endoplasmic reticulum (ER). In primary neurons, *ITPKB* expression change was shown to increase or reduce levels of a-synuclein aggregation.[18] *ITPKB* knockdown also leads to the accumulation of calcium in mitochondria which can inhibit autophagy. *ITPKB* mRNA levels were also shown to be correlated with *SNCA* expression in cortex and a-synuclein protein levels in A53T or A30P mutants.[19] *IP6K2* and *PPIP5K2* interact with similar molecules. *IP6K2* converts inositol hexakisphosphate (IP6) to 5-diphosphoinositol pentakisphosphate (5-IP7) or 1-diphosphoinositol pentakisphosphate (1-IP7) to bis-diphosphoinositol tetrakisphosphate (1,5-IP8) while *PPIP5K2* convert 5-IP7 to 1,5-IP8 and IP6 to 1-IP7. [20] *IP6K2* was implicated in cell death, apoptosis and neuroprotection.[21] In mice, *IP6K2* was found to regulate mitophagy through interaction with *PINK1*.[21] *PPIP5K2* has not been previously implicated in PD. It was associated with hearing loss and colorectal carcinoma.[22,23] *INPP5F* is involved with a different inositol pathway, it encodes Sac2, which converts phosphoinositides such as PI(4,5)P2 to phosphatidylinositol during endocytosis.[24]

Inositol phosphate has been suggested to be involved in obesity, insulin resistance and energy metabolism.[25] [3H]Inositol 1,4,5-trisphosphate binding sites were found to be reduced in certain brain regions of PD patients such as the caudate nucleus, putamen, and pallidum.[26] Additionally, IP6 was shown to be associated with PD. IP6 has a neuroprotective effect on dopaminergic cells by preventing 6-OHDA-induced apoptosis.[27] IP6 inhibits the activity of β-

secretase 1 (BACE1), an enzyme that cleaves amyloid-β precursor protein into toxic Aβ peptides.[28] Paraquet-induced neurodegeneration in *Drosophila* was suggested to also be mediated by inositol phosphates.[29] Previous studies have also suggested that different stereoisomers of inositol such as *scyllo*-inositol can inhibit PD[30] or decrease *myoinositol* in PD patients.[31,32] Recent studies on inositol investigated the effect of *SYNJ1*, an autosomal recessive form of early-onset parkinsonism.[33] SYNJ1 is a lipid phosphatase of phosphatidylinositol-3,4,5-trisphosphate (PIP3).[34] SYNJ1 knockout cell models were associated with an increase of a-synuclein and PIP3 levels. PIP3 dysregulation was suggested to promote a-synuclein aggregation and the risk of PD. Based on the evidence from the candidate inositol genes and previous work on inositol, inositol could potentially be a therapeutic target for PD. In 1999, a clinical trial on inositol was conducted on nine PD patients.[35] The treatment with inositol compared with placebo did not improve clinical outcomes. However, we cannot rule out inositol and inositol phosphates as potential therapeutic targets as only nine patients were recruited for this trial.

SPNS1 and *MLX* were found to be the top causal gene in their respective locus with putative causal missense SNPs: rs7140 and rs665268. Rs7140 corresponds to p.Leu563Val on the *SPNS1* transcript variant X1. We found that *SPNS1* is also associated with lower expression in *SOX6_ATGR1* dopaminergic cell subpopulation. This subcluster was previously highlighted to be the most susceptible to neurodegeneration in PD.[9] *SPNS1* encodes a sphingolipid transmembrane transporter in the lysosome. The autophagy-lysosomal pathway has been well-established to be crucial in PD pathogenesis, especially the lysosomal sphingolipid metabolism pathway, which includes well established PD-associated genes such as *GBA1*, *GALC*, *SMPD1* and others.[36,37] SPNS1 deficiency results in lipid accumulation in the lysosome and impaired lysosomal function.[11]

MLX encodes a Max-like protein X which belongs to a family of transcription factors regulating glucose metabolism. Rs665268 is a missense variant (p.Gln139Arg) that was found to be associated with Takayasu's arteritis, an autoimmune systemic vasculitis.[38] *MLX* was also reported to be associated with age at onset of Alzheimer's disease in females.[39] This variant was suggested to affect two important PD pathways by increasing oxidative stress and suppressing autophagy in immune cells. Although *SPNS1* and *MLX* have not been previously implicated in PD, the role of *SPNS1* and *MLX* in PD needs to be further studied.

There are several limitations to this study. The GWAS on which this analysis is based on is of European populations, therefore our results are potentially restricted to this population only. In addition, the training set for the machine learning model is limited to a small set of known or highly likely PD genes with the assumption of one causal gene per locus. The study also lacked samples for a testing set due to the previous issue. Since these limitations may introduce some bias, we used different strategies such as controlling for an imbalanced dataset and choosing balanced accuracy as an evaluation function to maximize the performance of the model. Lastly, the meta-analysis of rare variants can also be somewhat biased due to case/control imbalance. Larger studies will be required to validate our findings.

Our results nominate multiple genes that have not been thoroughly studied in PD and provide foundation for future functional studies of these genes. As larger PD GWASs will nominate more SNPs and loci, prioritizing causal genes will be crucial to understand the underlying biological mechanisms and disease pathophysiology through additional studies. Future gene prioritization studies will also be able to leverage larger datasets with more positive labels as new PD genes get discovered, and therefore increase the accuracy of the predictions.

Methods

General structure of the study

Figure 1 depicts the design of the study. In brief, we aimed to nominate the most likely gene to be involved in PD from each GWAS locus from the most recent PD GWAS.[1] To do so, we first defined all the genes and SNPs that are within these loci (detailed in the next paragraph). Then, in order to nominate the top genes in each locus, we used a machine learning approach. Using a literature search and a consensus between the authors of the paper, we identified seven genes from different loci that are well-established in PD (*GBA1*, *LRRK2*, *SNCA*, *GCH1*, *MAPT*, *TMEM175*, *VPS13C*) and very likely to be the driving gene in their respective loci. We then acquired data for multiple features, including different distance measures from top SNPs, different QTLs, expression in relevant tissues and cell types and predictions of variant consequences (78 features were used after removal of redundant features). Using the seven well-established PD genes which received positive labels, and 212 genes in the same loci that received negative labels (i.e. not likely to drive the association with PD, since the PD-driving gene is already well-established), we trained the machine learning model, and created a prediction score for each gene in each locus. The top-score gene in each locus is the nominated gene to be associated with PD. We then performed multiple post hoc analyses to further validate and explore our results: burden tests for rare variants in the top-scoring genes, pathway enrichment and pathway PRS analyses, differential expression analyses and structural analyses for candidate coding variants.

Definition of loci and genes within each locus

Following the definition by Nalls et al,[1] all loci were defined based on the 90 independent risk variants (Supplementary Table 1). Variants within 250 kb were merged into a single locus which leaves us with 78 loci. All protein coding genes within 1 Mb of the risk variants were included in

the model. To exclude non-causal variants, echolocatoR was used as a comprehensive fine-mapping model.[40] This method leverages statistical and functional fine-mapping tools as well as epigenomic data to create posterior probability for each SNP in a locus.[40] We included SNPs nominated by echolocatoR in the credible sets and the 90 independent SNPs from the PD GWAS for downstream analysis.

Feature preprocessing

To leverage multi-omic data for the machine learning algorithm, we integrated a comprehensive list of datasets (Supplementary Table 2) which includes SNP functional annotation, expression and splicing quantitative trait loci (QTL), single-cell RNA sequencing (scRNA) and chromatin interaction. Since distance was previously shown to be the most predictive feature in about 60-70% of GWAS loci, the distance from each SNP to each gene in the locus and the distance to the transcription start site were included in the model.[41] To predict the severity of variant consequences, we used Variant Effect Predictor (VEP)[42] and Polyphen-2.[43] The SNP2GENE function on the FUMA platform was used to perform functional mapping of SNPs to expression QTLs (eQTLs).[44] In the FUMA settings, we chose the UKB release2b 10k European reference panel, a maximum distance of 1000kb from SNPs to gene, and included the MHC region. All other FUMA settings were kept as default. eQTL and 3D chromatin interaction mapping were performed using brain tissues, whole blood, FANTOM and GTEx datasets. Using scRNA datasets from Kamath et al,[9] we included gene expression from all ten subpopulations of dopaminergic neurons from postmortem brains of 10 PD and 10 control donors. A complete list of all datasets can be found in Supplementary Table 7.

Neighborhood scores

To integrate the concept of locus and LD in the model, we calculated the neighborhood scores for each feature by transforming the data relative to the best-scoring gene within each locus.[6] This allows, for example, the model to find the highest expressed genes across each locus. For example, if the feature is "maximum gene expression in blood", the gene with the highest expression in each locus would have a score of one while the score of the remaining genes in the locus would be calculated following expression of gene divided by the expression of highest expressed gene in the locus. To avoid having the closest gene as the smallest value, we used negative log transformation to keep the closest gene as the highest score.

Machine learning model to prioritize genes

We used XGBoost[45] to train the machine learning model. We selected well-established genes from Parkinson's disease loci for the training dataset (*GBA1*, *GCH1*, *LRRK2*, *MAPT*, *SNCA*, *TMEM175*, *VPS13C*). These genes were labeled as positive labels, and the remaining genes from these same loci were labeled as negative labels. In total, the training set was composed of 212 genes (7 positive labeled and 205 negative labeled). To address the imbalanced dataset, we set the scale_pos_weight parameter in XGBoost to the ratio of negative to positive labels. The model was trained once to remove redundant features and then to create the final training model. We performed hyperparameter tuning and five-fold cross-validation on both models. Mean average precision was used as an evaluation function to maximize the correct positive predictions made. Out of the total 284 features, 78 features passed feature selection for the final training model.

Functional enrichment analysis

To examine whether specific pathways may be involved in PD, based on the genes nominated in each locus, we performed an over-representation analysis using WebGestalt (WEB-based GEne

SeT AnaLysis Toolkit) on January 25, 2023.[46] We included the top candidate gene from each locus, and examined biological processes, cellular components and molecular functions from the Gene Ontology data. We set the reference gene list to "genome protein-coding", and pathways were considered to be associated with PD after FDR correction.

Single-cell and bulk RNAseq analyses

To examine whether genes nominated by the machine learning model may be differentially expressed in PD relevant models, we used publicly available single-cell and bulk RNAseq data from FOUNDIN-PD[10] and Kamath et al.[9] FOUNDIN-PD scRNA data includes 80 induced pluripotent stem cell (iPSC) lines collected after 65 days.[10] We then performed differential gene expression analyses between PD cases and controls. For scRNA, we used the MAST[47] package after adjusting for covariates such as age, sex and batch. For bulk RNAseq, we used DESeq2[48] while adjusting for the same covariates.

Pathway polygenic risk score analyses

Pathway-specific PRS analysis can further support a role for specific pathways in PD.[49] Using PRSet,[50] pathway-specific polygenic risk scores (PRS) were calculated for pathways nominated by gene set analysis on 14,828 PD cases and 13,283 controls from seven cohorts (McGill, Parkinson's Progression Markers Initiative (PPMI), Vance (dbGap phs000394), International Parkinson's Disease Genomics Consortium (IPDGC) NeuroX dataset (dbGap phs000918.v1.p1), National Institute of Neurological Disorders and Stroke (NINDS) Genome-Wide genotyping in Parkinson's Disease (dbGap phs000089.v4.p2), NeuroGenetics Research Consortium (NGRC) (dbGap phs000196.v3.p1) and UK Biobank). The number of cases and control for each cohort is described in Supplementary Table 8. Participants were unrelated individuals of European ancestry and were not gender mismatched. Rare SNPs (minor allele frequency < 0.01) with p-value < 0.05

were excluded from the analysis. LD clumping was performed using r2=0.1 and 250kb distance. Permutation test with 10000 label permutation was performed to generate empirical p-value for each gene set after adjusting for a prevalence of 0.005, age at onset for cases, age at enrollment for control, sex and top 10 principal components. Vance cohort was excluded from the meta-analysis due to significant heterogeneity.

Rare variant burden analyses

To examine whether there is association between rare variants in the genes nominated by the machine learning model and PD, we used MetaSKAT[51] to perform a meta-analyses of rare variants. We used whole exome sequencing (WES) available for 602 PD patients, 6,284 proxy patients and 140,207 controls from UK Biobank (n=147,093) and 2,600 PD patients, 3,677 controls from Accelerating Medicines Partnership Parkinson's Disease (AMP-PD)[52] datasets (n=6277). Additional selection criteria for UK Biobank and AMP PD were reported previously.[53,54] We performed the analysis on several groups of rare variants (allele frequency < 0.01): loss of function variants, nonsynonymous variants, potentially deleterious (CADD>20) variants and functional (including nonsynonymous, frame-shift, stop-gain, and splicing) variants. Pathway-specific rare variant analysis was performed by combining PD genes from the pathways nominated previously. All analyses were adjusted for age at onset for cases, age at sample for control and sex.

Structural analysis

We then set to examine the potential structural effects of candidate coding variants nominated by our analysis. The atomic coordinates of SPNS1 (Uniprot #H3BR82) were retrieved from the AlphaFold server (https://alphafold.ebi.ac.uk/). The structures of MLX-MAD1 and MLX-MLXIP were generated using AlphaFold-Multimer version 3, as implemented in ColabFold.[55,56] The ternary complex with a DNA duplex was generated by superposing the heterodimers on the crystal

structure of the MAD1-MAX-DNA complex (PDB 1NLW). The figures were generated using PyMol v.2.4.0.

Data availability

The data used for this study can be accessed on: FUMA https://fuma.ctglab.nl/; Cuomo et al. 2020; Bryois et al. 2021; SMR https://yanglab.westlake.edu.cn/software/smr/; and Kamath et al 2021.

Code availability

The scripts used for this study can be found on GitHub: github.com/gan-orlab/gene_prio

References

1 Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *The Lancet Neurology* **18**, 1091-1102 (2019).

2 Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190-1195, doi:doi:10.1126/science.1222794 (2012).

3 Li, Y. I., Wong, G., Humphrey, J. & Raj, T. Prioritizing Parkinson's disease genes using population-scale transcriptomic data. *Nature communications* **10**, 994 (2019).

4 Schilder, B. M. & Raj, T. Fine-mapping of Parkinson's disease susceptibility loci identifies putative causal variants. *Human Molecular Genetics* **31**, 888-900 (2022).

5 Kia, D. A. *et al.* Identification of candidate Parkinson disease genes by integrating genome-wide association study, expression, and epigenetic data sets. *JAMA neurology* **78**, 464-472 (2021).

6 Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nature Genetics* **53**, 1527-1533, doi:10.1038/s41588-021-00945-5 (2021).

7 Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2**, 56-67, doi:10.1038/s42256-019-0138-9 (2020).

8 Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017).

9 Kamath, T. *et al.* Single-cell genomic profiling of human dopamine neurons identifies a population that selectively degenerates in Parkinson's disease. *Nat Neurosci* **25**, 588-595, doi:10.1038/s41593-022-01061-1 (2022).

10 Bressan, E. *et al.* The Foundational Data Initiative for Parkinson Disease: Enabling efficient translation from genetic maps to mechanism. *Cell Genomics* **3**, 100261, doi:https://doi.org/10.1016/j.xgen.2023.100261 (2023).

11 He, M. *et al.* Spns1 is a lysophospholipid transporter mediating lysosomal phospholipid salvage. *Proceedings of the National Academy of Sciences* **119**, e2210353119, doi:doi:10.1073/pnas.2210353119 (2022).

12 Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590-596 (2021).

13 Billin, A. N. & Ayer, D. E. The Mlx network: evidence for a parallel Max-like transcriptional network that regulates energy metabolism. *Curr Top Microbiol Immunol* **302**, 255-278, doi:10.1007/3-540-32952-8_10 (2006).

14 Billin, A. N., Eilers, A. L., Queva, C. & Ayer, D. E. Mlx, a novel Max-like BHLHZip protein that interacts with the Max network of transcription factors. *J Biol Chem* **274**, 36344-36350, doi:10.1074/jbc.274.51.36344 (1999).

15 Nair, S. K. & Burley, S. K. X-ray structures of Myc-Max and Mad-Max recognizing DNA. Molecular bases of regulation by proto-oncogenic transcription factors. *Cell* **112**, 193-205, doi:10.1016/s0092-8674(02)01284-9 (2003).

16 Billin, A. N., Eilers, A. L., Coulter, K. L., Logan, J. S. & Ayer, D. E. MondoA, a novel basic helix-loop-helix-leucine zipper transcriptional activator that constitutes a positive branch of a max-like network. *Mol Cell Biol* **20**, 8845-8854, doi:10.1128/mcb.20.23.8845-8854.2000 (2000).

17 Chakraborty, A. The inositol pyrophosphate pathway in health and diseases. *Biol Rev Camb Philos Soc* **93**, 1203-1227, doi:10.1111/brv.12392 (2018).

18 Apicco, D. J. *et al.* The Parkinson's disease-associated gene ITPKB protects against α-synuclein aggregation by regulating ER-to-mitochondria calcium release. *Proceedings of the National Academy of Sciences* **118**, e2006476118 (2021).

19 Di Leva, F. *et al.* Increased Levels of the Parkinson's Disease-Associated Gene ITPKB Correlate with Higher Expression Levels of α-Synuclein, Independent of Mutation Status. *International Journal of Molecular Sciences* **24**, 1984 (2023).

20 Chakraborty, A., Kim, S. & Snyder, S. H. Inositol Pyrophosphates as Mammalian Cell Signals. *Science Signaling* **4**, re1-re1, doi:10.1126/scisignal.2001958 (2011).

21 Nagpal, L., Kornberg, M. D. & Snyder, S. H. Inositol hexakisphosphate kinase-2 non-catalytically regulates mitophagy by attenuating PINK1 signaling. *Proceedings of the National Academy of Sciences* **119**, e2121946119 (2022).

22 Cao, C.-H. *et al.* PPIP5K2 promotes colorectal carcinoma pathogenesis through facilitating DNA homologous recombination repair. *Oncogene* **40**, 6680-6691 (2021).

23 Yousaf, R. *et al.* Mutations in Diphosphoinositol-Pentakisphosphate Kinase PPIP5K2 are associated with hearing loss in human and mouse. *PLoS genetics* **14**, e1007297 (2018).

24 Nakatsu, F. *et al.* Sac2/INPP5F is an inositol 4-phosphatase that functions in the endocytic pathway. *Journal of Cell Biology* **209**, 85-95, doi:10.1083/jcb.201409064 (2015).

25 Chatree, S., Thongmaen, N., Tantivejkul, K., Sitticharoon, C. & Vucenik, I. Role of inositols and inositol phosphates in energy metabolism. *Molecules* **25**, 5079 (2020).

26 Kitamura, N., Hashimoto, T., Nishino, N. & Tanaka, C. Inositol 1, 4, 5-trisphosphate binding sites in the brain: regional distribution, characterization, and alterations in brains of patients with Parkinson's disease. *Journal of Molecular Neuroscience* **1**, 181-187 (1989).

27 Zhang, Z. *et al.* Neuroprotection of inositol hexaphosphate and changes of mitochondrion mediated apoptotic pathway and α-synuclein aggregation in 6-OHDA induced parkinson′ s disease cell model. *Brain research* **1633**, 87-95 (2016).

28 Abe, T. K. & Taniguchi, M. Identification of myo-inositol hexakisphosphate (IP6) as a β-secretase 1 (BACE1) inhibitory molecule in rice grain extract and digest. *FEBS open bio* **4**, 162-167 (2014).

29 Shukla, A. K. *et al.* Metabolomic analysis provides insights on paraquat-induced Parkinson-like symptoms in Drosophila melanogaster. *Molecular neurobiology* **53**, 254-269 (2016).

30 Ibrahim, T. & McLaurin, J. α-Synuclein aggregation, seeding and inhibition by scyllo-inositol. *Biochemical and biophysical research communications* **469**, 529-534 (2016).

31 Gröger, A., Kolb, R., Schäfer, R. & Klose, U. Dopamine reduction in the substantia nigra of Parkinson's disease patients confirmed by in vivo magnetic resonance spectroscopic imaging. *PloS one* **9**, e84081 (2014).

32 Shah, A., Han, P., Wong, M.-Y., Chang, R. C.-C. & Legido-Quigley, C. Palmitate and Stearate are Increased in the Plasma in a 6-OHDA Model of Parkinson's Disease. *Metabolites* **9**, 31 (2019).

33 Quadri, M. *et al.* Mutation in the SYNJ1 Gene Associated with Autosomal Recessive, Early-Onset Parkinsonism. *Human Mutation* **34**, 1208-1215, doi:https://doi.org/10.1002/humu.22373 (2013).

34 Choong, C.-J. *et al.* Phosphatidylinositol-3, 4, 5-trisphosphate interacts with alpha-synuclein and initiates its aggregation and formation of Parkinson's disease-related fibril polymorphism. *Acta Neuropathologica*, 1-23 (2023).

35 Mishori, A., Levine, J., Kahana, E. & Belmaker, R. H. Inositol is not therapeutic in Parkinson's Disease. *Human Psychopharmacology: Clinical and Experimental* **14**, 271-272, doi:https://doi.org/10.1002/(SICI)1099-1077(199906)14:4<271::AID-HUP86>3.0.CO;2-I (1999).

36 Senkevich, K. & Gan-Or, Z. Autophagy lysosomal pathway dysfunction in Parkinson's disease; evidence from human genetics. *Parkinsonism & related disorders* **73**, 60-71 (2020).

37 Senkevich, K. *et al.* GALC variants affect galactosylceramidase enzymatic activity and risk of Parkinson's disease. *Brain* **146**, 1859-1872, doi:10.1093/brain/awac413 (2023).

38 Tamura, N. *et al.* Single-Nucleotide Polymorphism of the MLX Gene Is Associated With Takayasu Arteritis. *Circ Genom Precis Med* **11**, e002296, doi:10.1161/circgen.118.002296 (2018).

39 Li, Y.-J. *et al.* Identification of novel genes for age-at-onset of Alzheimer's disease by combining quantitative and survival trait analyses. *Alzheimer's & Dementia* **n/a**, doi:https://doi.org/10.1002/alz.12927.

40 Schilder, B. M. & Raj, T. Fine-mapping of Parkinson's disease susceptibility loci identifies putative causal variants. *Human Molecular Genetics* **31**, 888-900, doi:10.1093/hmg/ddab294 (2021).

41 Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-838 (2010).

42 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biology* **17**, 122, doi:10.1186/s13059-016-0974-4 (2016).

43 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-249, doi:10.1038/nmeth0410-248 (2010).

44 Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nature Communications* **8**, 1826, doi:10.1038/s41467-017-01261-5 (2017).

45 Chen, T. & Guestrin, C. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* 785-794.

46 Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Research* **47**, W199-W205, doi:10.1093/nar/gkz401 (2019).

47 Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16**, 278, doi:10.1186/s13059-015-0844-5 (2015).

48 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

49 Bandres-Ciga, S. *et al.* Large-scale pathway specific polygenic risk and transcriptomic community network analysis identifies novel functional pathways in Parkinson disease. *Acta neuropathologica* **140**, 341-358 (2020).

50 Choi, S. W. *et al.* PRSet: Pathway-based polygenic risk score analyses and software. *Plos Genetics* **19**, e1010624 (2023).

51 Lee, S., Teslovich, T. M., Boehnke, M. & Lin, X. General framework for meta-analysis of rare variants in sequencing association studies. *Am J Hum Genet* **93**, 42-53, doi:10.1016/j.ajhg.2013.05.010 (2013).

52 Iwaki, H. *et al.* Accelerating Medicines Partnership: Parkinson's Disease. Genetic Resource. *Movement Disorders* **36**, 1795-1804, doi:https://doi.org/10.1002/mds.28549 (2021).

53 Yu, E. *et al.* Fine mapping of the HLA locus in Parkinson's disease in Europeans. *NPJ Parkinsons Dis* **7**, 84, doi:10.1038/s41531-021-00231-5 (2021).

54 Senkevich, K. *et al.* Association of rare variants in ARSA with Parkinson’s disease. *medRxiv*, 2023.2003.2008.23286773, doi:10.1101/2023.03.08.23286773 (2023).

55 Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *BioRxiv*, 2021.2010. 2004.463034 (2021).

56 Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nature methods* **19**, 679-682 (2022).

Chapter 6: General Discussion

Synucleinopathies, including RBD, pose significant challenges and burdens to the global population. Unfortunately, current clinical trials have not provided effective solutions for these disorders. Researchers are confronted with various obstacles, including the lack of early diagnosis methods, the need for more reliable biomarkers, and the importance of genetic screening.

In this thesis, the importance of genetic screening in clinical trials is extensively explored. Specifically, the focus is on investigating the role of heterozygous *PRKN* variants, which have been subject to debate and controversy in relation to PD. By conducting thorough investigations, including systematic sequencing of rare variants and CNVs, and performing meta-analyses, the study aims to shed light on the potential impact of these *PRKN* variants in PD and synucleinopathies.

Furthermore, the thesis delves into the exploration of novel biomarkers and genetic targets for future therapeutic interventions. This is achieved through fine-mapping of the *HLA* locus, a genetic region associated with neuroinflammation, and the prioritization of candidate genes within the PD GWAS loci. By integrating genetic, transcriptomic, and epigenetic data, as well as employing machine learning approaches, the study aims to identify promising biomarkers and potential genetic targets for the development of future therapeutics in PD.

In Chapter 2, I found no evidence of association between heterozygous *PRKN* variants and PD. Our results got validated by the findings from other studies, which included a larger meta-analysis with a substantial number of patients with whole genome sequencing and participants from the UK Biobank.[145,146]

Functional studies provide evidence that "cryptic" *PRKN* mutations, such as deep intronic variants and exon inversions, do not contribute significantly to missed second mutations.[71] Furthermore, it was observed that cryptic *PRKN* mutations account for a considerable proportion (44%) of early-onset PD patients with monoallelic *PRKN* mutations. These findings highlight the importance of characterizing pathogenic *PRKN* variants for accurate diagnosis and management of *PRKN*-PD patients.

Although clinical trials targeting *PRKN* have not been successful thus far, one study suggests that enhancing mitophagy could potentially be a therapeutic approach for *PRKN*-PD patients.[147] Natural occurring *PRKN* variants, such as p.V224A, as well as structure-guided designer variants (p.W403A, p.F146A), have shown the ability to rescue common pathogenic *PRKN* variants, indicating the potential for targeted interventions.

While this study focused on participants of European ancestry, *PRKN* CNVs were found in more than 0.5% of all UK Biobank patients, regardless of ancestry.[146] The observation of distinct lengths of deletions in *PRKN* suggests independent origins, highlighting the diverse range of CNVs in different populations. The identification of biallelic *PRKN* variants through screening in clinical trials is crucial for precise drug.

In Chapter 3, I identified *HLA-DRB1* 11V, 13H, and 33H as variants potentially driving the association with PD. A previous study by Hollenbach *et al* suggested an interaction between smoking history, the "shared epitope," and *HLA-DRB1* 11V in rheumatoid arthritis.[148] They found that a positive smoking history combined with the presence of one or two shared epitope alleles and *HLA-DRB1* 11V had a greater effect size in PD risk compared to considering *HLA-DRB1* alone. Smoking-induced citrullination of proteins, the conversion of arginine to citrulline, may

lead to the development of antibodies to citrullinated protein antigens, potentially affecting HLA binding affinity for alleles protective in PD.[149]

Hollenbach *et al* suggested an association between *HLA-DRB1* 01:01 and PD risk, although the association was borderline significant (p=0.02).[148] However, in this study with a larger cohort of 12,137 patients and 14,422 proxy patients, *HLA-DRB1* 01:01 was not significant after meta-analysis. The larger sample size in this study provides more robust evidence compared to the smaller cohort in the study by Hollenbach *et al*.

Another recent study, in which I am a co-author, proposed a shared mechanism between PD, AD, and amyotrophic lateral sclerosis involving *HLA-DRB1**04, which carries the 33H amino acid change.[150] This variant was associated with decreased neurofibrillary tangles in postmortem brains and exhibited binding to a K311 acetylated tau PHF6 sequence. While the involvement of smoking and other proteins cannot be excluded, the selective reactivity of CD4+ T cells toward K311 acetylated tau may play a role in facilitating early clearance of toxic tau aggregates. The study suggested that antibody therapy targeting K311 acetylated tau could be an interesting avenue to explore in future clinical trials.[150]

In Chapter 4, I nominated the potential implication of the HLA locus in iRBD. *HLA-DRB1**11:01 was suggested to be associated with risk for iRBD. No association was found in LBD. Larger studies will be necessary in iRBD and LBD to have more power to detect association with HLA. For instance, my study did not have the power to examine the association of HLA-DRB1 33H, which was associated in PD, in iRBD and LBD.

Large reference panels are important for accurate HLA imputation and fine-mapping studies due to the extreme polymorphism and complex LD structure of HLA genes and the MHC

locus.[151] The extensive diversity in *HLA* genes is driven by an evolutionary pressure to bind to a wide range of foreign antigens, allowing for effective detection of various foreign substances throughout our lifetime.[152]

In Chapter 5, I focused on gene prioritization of PD GWAS loci using machine learning. During the analysis, the most likely causal genes from the PD GWAS were nominated, but some loci had two or more genes with high probability scores. For example, *CLCN3* and *NEK1* had high scores of 0.98 and 0.92, respectively. Although only one causal gene per locus was selected in the training set, it is possible that some GWAS loci may contain multiple causal genes. This could be due to the presence of multiple independent variants within the same locus. However, it's important to interpret these findings with caution, as independent variants can also affect the same gene. Additionally, it's possible that the model is unable to accurately determine the causal gene, resulting in multiple genes with high probability scores.

Interestingly, the top genes in some loci, such as *HLA*, had low probability scores. This could be attributed to the complex LD structure, which leads to many weak eQTLS as the variants in LD are associated with multiple genes. The model may struggle to accurately predict the causal gene in such cases. Furthermore, the number of samples used in statistical testing of features, such as eQTLs and enhancer-promoter interactions, is relevant to the training of the model. Features generated from studies with smaller sample sizes may contain more missing data and are more likely to be excluded from the model. For example, although enhancer-promoter interaction data was part of the training features, it may not have been considered important for most variant-gene pairs.

While distance between variants and genes is a strong predictor in the model, it's important to note that not all top genes can be predicted based on distance alone. 13 out of the 78 genes were

not the closest genes based on distance from the gene to the top GWAS SNPs, and 25 based on distance to the transcription start site. The model only includes features from tissues and cell types that are relevant to the training data. Therefore, if the majority of causal genes in the training data are not associated with microglia, for example, microglia data may not be considered important in the model. Although many PD-relevant tissues and cell types were included, some causal genes may have been missed because the model prioritized tissues and cell types associated with genes from the training data. As more well-established PD genes are identified in the future, leveraging additional data types may improve the model's performance.

Additional studies can be performed on lesser-studied genes associated with PD. For example, we can generate knockout and overexpression iPSC lines for the nominated genes and grow them into midbrain organoids. The midbrain organoids will be profiled using targeted phenotypic assays related to known PD mechanisms, including α-synuclein accumulation, mitochondrial dysfunction, and lysosomal/*GBA1* function.

Parkinson's disease genes nominated by machine learning

Several genes nominated by the model are associated with known PD pathways, such as the ALP. One of the top nominated genes, *CLCN3*, which scored 0.98 in its respective locus, encodes a chloride voltage-gated channel that is present in all cell types.[153] *CLCN3* has been identified in endosomes and on the vesicles of the lysosome. Loss-of-function variants in *CLCN3* have been shown to be associated with neurodevelopmental disorders.[153]

Another interesting finding is the nomination of *NEK1* as the second gene (score = 0.92) in the *CLCN3* locus. *NEK1* encodes a neuronal kinase that is involved in cell cycle regulation. *NEK1* has previously been implicated in amyotrophic lateral sclerosis.[154] *NEK1* deficiency has been associated with lysosomal dysfunction due to reduced glucose uptake and mitochondrial

dysfunction. Furthermore, mice neurons with *NEK1* deficiency have shown the presence of alpha-synuclein, a protein associated with PD pathology.[155]

Another interesting gene nominated by the model is *TMEM163*. *TMEM163* encodes a zinc ion transporter located in the synaptic vesicle membrane.[156] Zinc plays a vital role in various cellular processes, including immune response activation. However, excessive zinc accumulation can lead to the generation of ROS and subsequent mitochondrial dysfunction.[156]

The association of zinc transporters with several diseases, such as AD, cancer, and diabetes, suggests their potential involvement in disease mechanisms.[156] Dysregulation of zinc homeostasis has been implicated in the pathology of these diseases. *TMEM163's* role in zinc transport may contribute to its association with intracranial injury and possibly other neurological conditions.

The *SCARB2* locus has been identified in GWAS as being associated with both PD and RBD. Interestingly, the variants within the *SCARB2* locus that are associated with PD and RBD were independent. Furthermore, these *SCARB2* variants have the potential to affect gene expression differently in various brain regions.[126] *SCARB2*, also known as lysosomal integral membrane protein-2 (LIMP-2), plays an essential role in the transport of GCase from the Golgi apparatus to the lysosome. Studies have shown that knockout of LIMP-2 in mice leads to a reduction in GCase levels across various tissues.[157] The association of the *SCARB2* locus with PD and RBD suggests that variations in *SCARB2* may influence GCase transport and lysosomal function.

Progranulin, encoded by the *GRN* gene, is a gene nominated by the model as a potential candidate gene for PD. Progranulin is a protein that undergoes cleavage by serine proteases to

generate smaller peptides called granulins.[158] It has been implicated in various neurodegenerative disorders, including PD, AD, and amyotrophic lateral sclerosis.[159]

Heterozygous loss-of-function variants in the *GRN* gene are known to be associated with a specific form of frontotemporal dementia, a neurodegenerative disorder characterized by progressive cognitive and behavioral changes.[160] This suggests that *GRN* plays a crucial role in neuronal function and maintenance.

Progranulin is a key neuronal gene involved in multiple cellular processes. It is involved in the development, survival, and maintenance of neurons and microglia, the immune cells of the central nervous system.[158] Progranulin is also implicated in regulating lysosomal biogenesis, inflammation, tissue repair, stress response, and aging processes within the brain.

Decreased expression of progranulin has been associated with neuroinflammation and an increased risk for various neurological diseases.[159] Importantly, progranulin exhibits anti-inflammatory properties, while the granulin peptides derived from its cleavage process can have pro-inflammatory effects.[161] This suggests that the balance between progranulin and granulin levels may be critical for maintaining proper immune responses in the brain.

In mice, progranulin has been found to bind to GCase and recruit heat shock protein 70 under stress conditions.[162] The interaction between progranulin and GCase, along with the recruitment of heat shock protein 70, suggests a potential role for progranulin in modulating cellular stress responses and proteostasis.

Progranulin can be a potential therapeutic target. By knocking out sortilin, a receptor that regulates the trafficking of progranulin to the lysosome, increased progranulin levels can be

achieved.[163] This approach holds potential for augmenting progranulin levels and potentially modulating its anti-inflammatory and neuroprotective effects in neurodegenerative diseases.

GALC, which encodes galactosylceramidase, is a gene involved in the breakdown of galactosylceramide and galactosylsphingosine in the lysosome.[164] While *GALC* mutations have been linked to an increased risk for PD, the specific role of *GALC* in PD pathogenesis is not yet well understood.[74] Interestingly, one study have shown that *GALC* knockout does not significantly affect the activity of GCase. GCase activity and α-synuclein accumulation in induced pluripotent stem cell (iPSC)-derived neurons were not altered by *GALC* knockout.[165]

These findings suggest that *GALC* may have a distinct role in PD, possibly independent of GCase activity and α-synuclein accumulation. However, further studies are needed to uncover the specific mechanisms by which *GALC* may contribute to PD pathology.

CTSB, which encodes for Cathepsin B, is a gene that belongs to the family of lysosomal cysteine proteases.[166] Cathepsin B and other cysteine cathepsins play crucial roles in various physiological processes, including protein processing, MHC class II-mediated antigen presentation, and apoptosis.[167] Additionally, the cysteine cathepsin gene family has been implicated in cancer, inflammation, and neurodegenerative diseases.

CTSB has been found to be highly expressed in neocortical and hippocampal neurons. Microglia-induced neuronal apoptosis has been associated with *CTSB*, suggesting its involvement in the neuroinflammatory response.[167] Studies in transgenic mice carrying the *SNCA* A53T mutation, which is associated with familial forms of PD, have shown that *CTSB* can cleave α-synuclein.[168] This cleavage of α-synuclein by *CTSB* may have implications for the aggregation and accumulation of α-synuclein in Parkinson's disease.

Furthermore, a specific intronic variant of *CTSB*, rs1293298, has been found to be protective for individuals who carry mutations in the *GBA1* gene and have PD or DLB. This variant is also associated with a lower age at disease onset.[169] Interestingly, this protective effect may be attributed to increased *CTSB* expression in the brain, potentially due to the activation of GCase, an enzyme involved in the breakdown of glucosylceramide, through the cleavage of prosaposin into saposin C by *CTSB*.

While its probability score may not be high, *P2RY12*, which encodes for a G protein-coupled receptor, has been associated with PD in previous studies.[170] *P2RY12* is known to be a marker of nonactivated microglia, which are immune cells in the brain.[171] Upon activation, microglia downregulate *P2RY12* expression. Some evidence suggests that *P2RY12* is involved in facilitating the migration of microglia towards sites of injury, where they can contribute to tissue repair and maintenance, such as maintaining the integrity of the blood-brain barrier.[172]

Clopidogrel, a drug commonly used as an antiplatelet agent, has been shown to inhibit *P2RY12*. However, it is important to note that while clopidogrel may affect *P2RY12* signaling and function in platelets, there is currently no evidence supporting its direct effect on Parkinson's disease.[173] The potential role of *P2RY12* in PD pathogenesis and its modulation by clopidogrel or other agents in the context of the disease requires further investigation.

TOX3, which encodes for TOX high-mobility group box family member 3, has been nominated in both PD and restless leg syndrome (RLS) in genetic studies with opposite direction of effect.[174] TOX3 is a nuclear transcription regulation factor that is expressed in the brain and is also involved in breast cancer.[175]

TOX3 has been implicated in the regulation of estrogen receptor-mediated gene expression.[175] Estrogen receptors play important roles in various physiological processes, including neuronal function. The exact mechanisms by which *TOX3* may contribute to PD or RLS are not yet fully understood.

RLS is a common motor disorder characterized by an uncontrollable urge to move the legs, typically occurring during periods of rest or inactivity, particularly in the evening or at night.[176] Although RLS can occur independently, it has been observed that a higher percentage of individuals with PD also experience RLS symptoms compared to the general population.[176] However, the association between RLS and PD is complex and not fully elucidated.

Chapter 7: Conclusions and Future Directions

Synucleinopathies are a group of neurodegenerative diseases without disease-modifying treatments. Using precision medicine, we can develop treatment for targeted patient. This thesis focuses on methods to isolate specific populations for clinical trials and identify novel genetic drug targets. Genetic imbalance between randomized arm can confound true therapeutic effects. For example, *PRKN*-PD patients are known to have different disease progression and brain pathology compared to idiopathic PD. This study addressed the lack of evidence supporting the association between rare heterozygous *PRKN* variant carriers and PD. By highlighting this gap, this study contributes to enhancing inclusion criteria in therapeutic trials.

To discover potential genetic risk factors, I nominated *HLA-DRB1* variants and alleles within the *HLA* locus for PD and RBD. Further studies are needed, particularly for RBD, as *HLA* allele frequencies can significantly vary across populations and subpopulations. Exploring non-European populations is crucial for understanding *HLA* in RBD and LBD.

Given the relative obscurity of RBD, many cases likely go unreported. This is further compounded by the limited availability of vPSG in many countries. Consequently, alternative diagnostic methods, such as accurate biomarkers, play a vital role in patient recruitment. Identifying biomarkers for disease conversion is also valuable for diagnosing PD, DLB, and MSA.

In addition to exploring the *HLA* locus, an important aspect of my research involved performing gene prioritization of genes identified in the PD GWAS. This approach aimed to promote and stimulate further investigations into lesser-studied PD genes, which have not received as much attention in previous research.

By nominating these lesser-known PD genes for prioritized study, we can open up new avenues of research and expand our understanding of the complex mechanisms underlying PD. These genes can shed light on the disease's etiology, disease onset, progression, and even disease severity.

Moreover, investigating the lesser-known PD genes may reveal previously unrecognized subtypes or phenotypes of PD. This knowledge can contribute to the development of personalized medicine approaches, where treatment strategies can be tailored to specific genetic profiles or disease subgroups. By unraveling the unique characteristics associated with these genes, we may gain a more comprehensive understanding of the heterogeneity observed within PD.

As more genes are discovered, additional PD subtypes can be identified, leading to the development of targeted therapeutics. Currently, clinical trials are targeting *GBA1* and *LRRK2* patients. It is advisable to examine patients who are genetically or clinically similar to *GBA1* and *LRRK2* carriers, as they are most likely to benefit from treatments targeting these specific targets.

Reference

1. Coon, E. A. & Singer, W. Synucleinopathies. *Continuum (Minneap Minn)* **26**, 72–92 (2020).
2. Parkinson, J. An Essay on the Shaking Palsy. *JNP* **14**, 223–236 (2002).
3. Dorsey, E. R. *et al.* Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. *Neurology* **68**, 384–386 (2007).
4. Dorsey, E. R. *et al.* Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology* **17**, 939–953 (2018).
5. Poewe, W. *et al.* Parkinson disease. *Nat Rev Dis Primers* **3**, 1–21 (2017).
6. Payami, H., Zareparsi, S., James, D. & Nutt, J. Familial Aggregation of Parkinson Disease: A Comparative Study of Early-Onset and Late-Onset Disease. *Archives of Neurology* **59**, 848–850 (2002).
7. Postuma, R. B. *et al.* MDS clinical diagnostic criteria for Parkinson's disease. *Movement Disorders* **30**, 1591–1601 (2015).
8. Clarke, C. E. *et al. Clinical effectiveness and cost-effectiveness of physiotherapy and occupational therapy versus no therapy in mild to moderate Parkinson's disease: a large pragmatic randomised controlled trial (PD REHAB)*. (NIHR Journals Library, 2016).
9. Levin, J., Kurz, A., Arzberger, T., Giese, A. & Höglinger, G. U. The Differential Diagnosis and Treatment of Atypical Parkinsonism. *Dtsch Arztebl Int* **113**, 61–69 (2016).
10. Postuma, R. B. *et al.* Identifying prodromal Parkinson's disease: Pre-Motor disorders in Parkinson's disease. *Movement Disorders* **27**, 617–626 (2012).
11. Poewe, W. The natural history of Parkinson's disease. *J Neurol* **253**, vii2–vii6 (2006).

12. Poewe, W. & Mahlknecht, P. The clinical progression of Parkinson's disease. *Parkinsonism & Related Disorders* **15**, S28–S32 (2009).

13. Aarsland, D., Zaccai, J. & Brayne, C. A systematic review of prevalence studies of dementia in Parkinson's disease. *Movement Disorders* **20**, 1255–1263 (2005).

14. Damier, P., Hirsch, E. C., Agid, Y. & Graybiel, A. M. The substantia nigra of the human brain: II. Patterns of loss of dopamine-containing neurons in Parkinson's disease. *Brain* **122**, 1437–1448 (1999).

15. Recasens, A. & Dehay, B. Alpha-synuclein spreading in Parkinson's disease. *Frontiers in Neuroanatomy* **8**, (2014).

16. Braak, H. *et al.* Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiology of Aging* **24**, 197–211 (2003).

17. Burke, R. E., Dauer, W. T. & Vonsattel, J. P. G. A critical evaluation of the Braak staging scheme for Parkinson's disease. *Annals of Neurology* **64**, 485–491 (2008).

18. Kalia, L. V. *et al.* Clinical Correlations With Lewy Body Pathology in LRRK2-Related Parkinson Disease. *JAMA Neurology* **72**, 100–105 (2015).

19. Collier, T. J., Kanaan, N. M. & Kordower, J. H. Ageing as a primary risk factor for Parkinson's disease: evidence from studies of non-human primates. *Nat Rev Neurosci* **12**, 359–366 (2011).

20. Pringsheim, T., Jette, N., Frolkis, A. & Steeves, T. D. L. The prevalence of Parkinson's disease: A systematic review and meta-analysis. *Movement Disorders* **29**, 1583–1590 (2014).

21. Collier, T. J., Kanaan, N. M. & Kordower, J. H. Aging and Parkinson's disease: Different sides of the same coin? *Movement Disorders* **32**, 983–990 (2017).

22. Wooten, G. F., Currie, L. J., Bovbjerg, V. E., Lee, J. K. & Patrie, J. Are men at greater risk for Parkinson's disease than women? *Journal of Neurology, Neurosurgery & Psychiatry* **75**, 637–639 (2004).

23. Rocca, W. A. *et al.* Increased risk of parkinsonism in women who underwent oophorectomy before menopause. *Neurology* **70**, 200–209 (2008).

24. Noyce, A. J. *et al.* Meta-analysis of early nonmotor features and risk factors for Parkinson disease. *Annals of Neurology* **72**, 893–901 (2012).

25. Cotzias, G. C., Van Woert, M. H. & Schiffer, L. M. Aromatic Amino Acids and Modification of Parkinsonism. *N Engl J Med* **276**, 374–379 (1967).

26. Fox, S. H. *et al.* The Movement Disorder Society Evidence-Based Medicine Review Update: Treatments for the motor symptoms of Parkinson's disease. *Movement Disorders* **26**, S2–S41 (2011).

27. Connolly, B. S. & Lang, A. E. Pharmacological Treatment of Parkinson Disease: A Review. *JAMA* **311**, 1670–1683 (2014).

28. Jenner, P. Molecular mechanisms of L-DOPA-induced dyskinesia. *Nat Rev Neurosci* **9**, 665–677 (2008).

29. Borovac, J. A. Side effects of a dopamine agonist therapy for Parkinson's disease: a mini-review of clinical pharmacology. *Yale J Biol Med* **89**, 37–47 (2016).

30. Postuma, R. B. *et al.* Risk and predictors of dementia and parkinsonism in idiopathic REM sleep behaviour disorder: a multicentre study. *Brain* **142**, 744–759 (2019).

31. Dauvilliers, Y. *et al.* REM sleep behaviour disorder. *Nat Rev Dis Primers* **4**, 1–16 (2018).

32. Hu, M. T. REM sleep behavior disorder (RBD). *Neurobiol Dis* **143**, 104996 (2020).

33. Kang, S.-H. *et al.* REM sleep behavior disorder in the Korean elderly population: prevalence and clinical characteristics. *Sleep* **36**, 1147–1152 (2013).

34. Haba-Rubio, J. *et al.* Prevalence and determinants of rapid eye movement sleep behavior disorder in the general population. *Sleep* **41**, zsx197 (2018).

35. Pujol, M. *et al.* Idiopathic REM sleep behavior disorder in the elderly Spanish community: a primary care center study with a two-stage design using video-polysomnography. *Sleep Medicine* **40**, 116–121 (2017).

36. Thomas, A. J. *et al.* Improving the identification of dementia with Lewy bodies in the context of an Alzheimer's-type dementia. *Alzheimer's Research & Therapy* **10**, 27 (2018).

37. Lippa, C. F. *et al.* DLB and PDD boundary issues: Diagnosis, treatment, molecular pathology, and biomarkers. *Neurology* **68**, 812–819 (2007).

38. McKeith, I. G. *et al.* Diagnosis and management of dementia with Lewy bodies: third report of the DLB Consortium. *Neurology* **65**, 1863–1872 (2005).

39. Emre, M. *et al.* Clinical diagnostic criteria for dementia associated with Parkinson's disease. *Mov Disord* **22**, 1689–1707; quiz 1837 (2007).

40. Polymeropoulos, M. H. *et al.* Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science* **276**, 2045–2047 (1997).

41. Schneider, S. A. & Alcalay, R. N. Neuropathology of genetic synucleinopathies with parkinsonism: Review of the literature. *Movement Disorders* **32**, 1504–1523 (2017).

42. Singleton, A. B. *et al.* alpha-Synuclein locus triplication causes Parkinson's disease. *Science* **302**, 841 (2003).

43. Devine, M. J., Gwinn, K., Singleton, A. & Hardy, J. Parkinson's disease and α-synuclein expression. *Mov Disord* **26**, 2160–2168 (2011).

44. Farrer, M. *et al.* Comparison of kindreds with parkinsonism and α-synuclein genomic multiplications. *Annals of Neurology* **55**, 174–179 (2004).

45. Chen, V. *et al.* The mechanistic role of alpha-synuclein in the nucleus: impaired nuclear function caused by familial Parkinson's disease SNCA mutations. *Human Molecular Genetics* **29**, 3107–3121 (2020).

46. Espay, A. J. *et al.* Revisiting protein aggregation as pathogenic in sporadic Parkinson and Alzheimer diseases. *Neurology* **92**, 329–337 (2019).

47. Baba, M. *et al.* Aggregation of alpha-synuclein in Lewy bodies of sporadic Parkinson's disease and dementia with Lewy bodies. *Am J Pathol* **152**, 879–884 (1998).

48. Whone, A. Monoclonal Antibody Therapy in Parkinson's Disease — The End? *New England Journal of Medicine* **387**, 466–467 (2022).

49. Lücking, C. B. *et al.* Association between Early-Onset Parkinson's Disease and Mutations in the Parkin Gene. *New England Journal of Medicine* **342**, 1560–1567 (2000).

50. Alcalay, R. N. *et al.* Frequency of known mutations in early-onset Parkinson disease: implication for genetic counseling: the consortium on risk for early onset Parkinson disease study. *Archives of neurology* **67**, 1116–1122 (2010).

51. Kilarski, L. L. *et al.* Systematic review and UK-based study of PARK2 (parkin), PINK1, PARK7 (DJ-1) and LRRK2 in early-onset Parkinson's disease. *Movement Disorders* **27**, 1522–1529 (2012).

52. Bandrés-Ciga, S. *et al.* Genome-wide assessment of Parkinson's disease in a Southern Spanish population. *Neurobiology of Aging* **45**, 213.e3-213.e9 (2016).

53. Benitez, B. A. *et al.* Resequencing analysis of five Mendelian genes and the top genes from genome-wide association studies in Parkinson's Disease. *Mol Neurodegeneration* **11**, 29 (2016).

54. Bras, J. *et al.* Analysis of Parkinson disease patients from Portugal for mutations in SNCA, PRKN, PINK1 and LRRK2. *BMC Neurology* **8**, 1 (2008).

55. Brooks, J. *et al.* Parkin and PINK1 mutations in early-onset Parkinson's disease: comprehensive screening in publicly available cases and control. *Journal of Medical Genetics* **46**, 375–381 (2009).

56. Camacho, J. L. G. *et al.* High frequency of Parkin exon rearrangements in Mexican-mestizo patients with early-onset Parkinson's disease. *Movement Disorders* **27**, 1047–1051 (2012).

57. Camargos, S. T. *et al.* Familial Parkinsonism and early onset Parkinson's disease in a Brazilian movement disorders clinic: Phenotypic characterization and frequency of SNCA, PRKN, PINK1, and LRRK2 mutations. *Movement Disorders* **24**, 662–666 (2009).

58. Lesage, S. *et al.* Rare heterozygous parkin variants in French early-onset Parkinson disease patients and controls. *Journal of Medical Genetics* **45**, 43–46 (2008).

59. Huttenlocher, J. *et al.* Heterozygote carriers for CNVs in PARK2 are at increased risk of Parkinson's disease. *Human Molecular Genetics* **24**, 5637–5643 (2015).

60. Hopfner, F. *et al.* Private variants in PRKN are associated with late-onset Parkinson's disease. *Parkinsonism & Related Disorders* **75**, 24–26 (2020).

61. Kay, D. M. *et al.* A comprehensive analysis of deletions, multiplications, and copy number variations in PARK2. *Neurology* **75**, 1189–1194 (2010).

62. Voutsinos, V., Munk, S. H. N. & Oestergaard, V. H. Common Chromosomal Fragile Sites—Conserved Failure Stories. *Genes* **9**, 580 (2018).

63. Helmrich, A., Ballarino, M. & Tora, L. Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Mol Cell* **44**, 966–977 (2011).

64. Chan, N. C. *et al.* Broad activation of the ubiquitin-proteasome system by Parkin is critical for mitophagy. *Hum Mol Genet* **20**, 1726–1737 (2011).

65. Valente, E. M. *et al.* Hereditary Early-Onset Parkinson's Disease Caused by Mutations in PINK1. *Science* **304**, 1158–1160 (2004).

66. Schofield, J. H. & Schafer, Z. T. Mitochondrial Reactive Oxygen Species and Mitophagy: A Complex and Nuanced Relationship. *Antioxidants & Redox Signaling* **34**, 517–530 (2021).

67. Xin, D., Gu, H., Liu, E. & Sun, Q. Parkin negatively regulates the antiviral signaling pathway by targeting TRAF3 for degradation. *Journal of Biological Chemistry* **293**, 11996–12010 (2018).

68. Farrer, M. *et al.* Lewy bodies and parkinsonism in families with parkin mutations. *Ann Neurol* **50**, 293–300 (2001).

69. Dawson, T. M. & Dawson, V. L. The role of parkin in familial and sporadic Parkinson's disease. *Movement Disorders* **25**, S32–S39 (2010).

70. Alcalay, R. N. *et al.* Cognitive and Motor Function in Long-Duration PARKIN-Associated Parkinson Disease. *JAMA Neurology* **71**, 62–67 (2014).

71. Lubbe, S. J. *et al.* Assessing the relationship between monoallelic PRKN mutations and Parkinson's risk. *Human Molecular Genetics* **30**, 78–86 (2021).

72. Schouten, J. P. *et al.* Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res* **30**, e57 (2002).

73. Paisán-Ruíz, C. *et al.* Cloning of the Gene Containing Mutations that Cause PARK8-Linked Parkinson's Disease. *Neuron* **44**, 595–600 (2004).

74. Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol* **18**, 1091–1102 (2019).

75. Lesage, S. *et al.* LRRK2emph Exon 41 Mutations in Sporadic Parkinson Disease in Europeans. *Archives of Neurology* **64**, 425–430 (2007).

76. Benamer, H. T. S. & Silva, R. de. LRRK2 G2019S in the North African Population: A Review. *ENE* **63**, 321–325 (2010).

77. Hassin-Baer, S. *et al.* The leucine rich repeat kinase 2 (LRRK2) G2019S substitution mutation. *J Neurol* **256**, 483–487 (2009).

78. Lee, A. J. *et al.* Penetrance estimate of LRRK2 p.G2019S mutation in individuals of non-Ashkenazi Jewish ancestry. *Movement Disorders* **32**, 1432–1438 (2017).

79. Cookson, M. R. The role of leucine-rich repeat kinase 2 (LRRK2) in Parkinson's disease. *Nat Rev Neurosci* **11**, 791–797 (2010).

80. Mata, I. F. *et al.* The discovery of LRRK2 p.R1441S, a novel mutation for Parkinson's disease, adds to the complexity of a mutational hotspot. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **171**, 925–930 (2016).

81. Ross, O. A. *et al.* Association of LRRK2 exonic variants with susceptibility to Parkinson's disease: a case–control study. *The Lancet Neurology* **10**, 898–908 (2011).

82. Wauters, F. *et al.* LRRK2 mutations impair depolarization-induced mitophagy through inhibition of mitochondrial accumulation of RAB10. *Autophagy* **16**, 203–222 (2020).

83. Jeong, G. R. & Lee, B. D. Pathological Functions of LRRK2 in Parkinson's Disease. *Cells* **9**, 2565 (2020).

84. Sosero, Y. L. & Gan-Or, Z. LRRK2 and Parkinson's disease: from genetics to targeted therapy. *Annals of Clinical and Translational Neurology* **n/a**,.

85. Uffelmann, E. *et al.* Genome-wide association studies. *Nat Rev Methods Primers* **1**, 1–21 (2021).

86. Korte, A. & Farlow, A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* **9**, 29 (2013).

87. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat Rev Genet* **20**, 467–484 (2019).

88. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).

89. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *The American Journal of Human Genetics* **99**, 1245–1260 (2016).

90. Sidransky, E. Gaucher disease: complexity in a "simple" disorder. *Molecular Genetics and Metabolism* **83**, 6–15 (2004).

91. Parlar, S. C., Grenn, F. P., Kim, J. J., Baluwendraat, C. & Gan-Or, Z. Classification of GBA1 Variants in Parkinson's Disease: The GBA1-PD Browser. *Movement Disorders* **38**, 489–495 (2023).

92. Balestrino, R. *et al.* Penetrance of Glucocerebrosidase (GBA) Mutations in Parkinson's Disease: A Kin Cohort Study. *Movement Disorders* **35**, 2111–2114 (2020).

93. Cilia, R. *et al.* Survival and dementia in GBA-associated Parkinson's disease: The mutation matters. *Annals of Neurology* **80**, 662–673 (2016).

94. Maple-Grødem, J. *et al.* Association of GBA Genotype With Motor and Functional Decline in Patients With Newly Diagnosed Parkinson Disease. *Neurology* **96**, e1036–e1044 (2021).

95. Alcalay, R. N. *et al.* Cognitive performance of GBA mutation carriers with early-onset PD: The CORE-PD study. *Neurology* **78**, 1434–1440 (2012).

96. Alcalay, R. N. *et al.* Glucocerebrosidase activity in Parkinson's disease with and without GBA mutations. *Brain* **138**, 2648–2658 (2015).

97. Keatinge, M. *et al.* Glucocerebrosidase 1 deficient Danio rerio mirror key pathological aspects of human Gaucher disease and provide evidence of early microglial activation preceding alpha-synuclein-independent neuronal cell death. *Human Molecular Genetics* **24**, 6640–6652 (2015).

98. Mazzulli, J. R. *et al.* Gaucher Disease Glucocerebrosidase and α-Synuclein Form a Bidirectional Pathogenic Loop in Synucleinopathies. *Cell* **146**, 37–52 (2011).

99. Lesage, S. *et al.* Loss of VPS13C Function in Autosomal-Recessive Parkinsonism Causes Mitochondrial Dysfunction and Increases PINK1/Parkin-Dependent Mitophagy. *The American Journal of Human Genetics* **98**, 500–513 (2016).

100. Schormair, B. *et al.* Diagnostic exome sequencing in early-onset Parkinson's disease confirms VPS13C as a rare cause of autosomal-recessive Parkinson's disease. *Clinical Genetics* **93**, 603–612 (2018).

101. Rademakers, R., Cruts, M. & van Broeckhoven, C. The role of tau (MAPT) in frontotemporal dementia and related tauopathies. *Human Mutation* **24**, 277–295 (2004).

102. Irwin, D. J. *et al.* Neuropathological and genetic correlates of survival and dementia onset in synucleinopathies: a retrospective analysis. *The Lancet Neurology* **16**, 55–65 (2017).

103. Hamilton, R. L. Lewy Bodies in Alzheimer's Disease: A Neuropathological Review of 145 Cases Using α-Synuclein Immunohistochemistry. *Brain Pathol* **10**, 378–384 (2006).

104. Zody, M. C. *et al.* Evolutionary Toggling of the MAPT 17q21.31 Inversion Region. *Nat Genet* **40**, 1076–1083 (2008).

105. Pittman, A. *et al.* Linkage disequilibrium fine mapping and haplotype association analysis of the tau gene in progressive supranuclear palsy and corticobasal degeneration. *J Med Genet* **42**, 837–846 (2005).

106. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nat Genet* **37**, 129–137 (2005).

107. Li, J. *et al.* Full sequencing and haplotype analysis of MAPT in Parkinson's disease and rapid eye movement sleep behavior disorder. *Mov Disord* **33**, 1016–1020 (2018).

108. Leveille, E., Ross, O. A. & Gan-Or, Z. Tau and MAPT genetics in tauopathies and synucleinopathies. *Parkinsonism Relat Disord* **90**, 142–154 (2021).

109. Moussaud, S. *et al.* Alpha-synuclein and tau: teammates in neurodegeneration? *Mol Neurodegener* **9**, 43 (2014).

110. Lei, P. *et al.* Tau protein: relevance to Parkinson's disease. *Int J Biochem Cell Biol* **42**, 1775–1778 (2010).

111. Strang, K. H., Golde, T. E. & Giasson, B. I. MAPT mutations, tauopathy, and mechanisms of neurodegeneration. *Lab Invest* **99**, 912–928 (2019).

112. Arendt, T., Stieler, J. T. & Holzer, M. Tau and tauopathies. *Brain Res Bull* **126**, 238–292 (2016).

113. Spires-Jones, T. L., Kopeikina, K. J., Koffie, R. M., de Calignon, A. & Hyman, B. T. Are Tangles as Toxic as They Look? *J Mol Neurosci* **45**, 438–444 (2011).

114. Hu, M. *et al.* Parkinson's disease-risk protein TMEM175 is a proton-activated proton channel in lysosomes. *Cell* **185**, 2292-2308.e20 (2022).

115. Krohn, L. *et al.* Genetic, Structural, and Functional Evidence Link TMEM175 to Synucleinopathies. *Annals of Neurology* **87**, 139–153 (2020).

116. Kurian, M. A., Gissen, P., Smith, M., Heales, S. J. & Clayton, P. T. The monoamine neurotransmitter disorders: an expanding range of neurological syndromes. *The Lancet Neurology* **10**, 721–733 (2011).

117. Rudakou, U. *et al.* Common and rare GCH1 variants are associated with Parkinson's disease. *Neurobiology of Aging* **73**, 231.e1-231.e6 (2019).

118. Mosaad, Y. M. Clinical Role of Human Leukocyte Antigen in Health and Disease. *Scandinavian Journal of Immunology* **82**, 283–306 (2015).

119. Klein, J. & Sato, A. The HLA System. *N Engl J Med* **343**, 782–786 (2000).

120. Wissemann, W. T. *et al.* Association of Parkinson Disease with Structural and Regulatory Variants in the HLA Region. *The American Journal of Human Genetics* **93**, 984–993 (2013).

121. Hamza, T. H. *et al.* Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. *Nat Genet* **42**, 781–785 (2010).

122. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat Genet* **51**, 414–430 (2019).

123. Nona, R. J., Greer, J. M., Henderson, R. D. & McCombe, P. A. HLA and amyotrophic lateral sclerosis: a systematic review and meta-analysis. *Amyotroph Lateral Scler Frontotemporal Degener* **24**, 24–32 (2023).

124. Chia, R. *et al.* Genome sequencing analysis identifies new loci associated with Lewy body dementia and provides insights into its genetic architecture. *Nat Genet* **53**, 294–303 (2021).

125. Bras, J. *et al.* Genetic analysis implicates APOE, SNCA and suggests lysosomal dysfunction in the etiology of dementia with Lewy bodies. *Hum Mol Genet* **23**, 6139–6146 (2014).

126. Krohn, L. *et al.* Genome-wide association study of REM sleep behavior disorder identifies polygenic risk and brain expression effects. *Nat Commun* **13**, 7496 (2022).

127. Ouled Amar Bencheikh, B. *et al.* LRRK2 protective haplotype and full sequencing study in REM sleep behavior disorder. *Parkinsonism Relat Disord* **52**, 98–101 (2018).

128. Gan-Or, Z. *et al.* The dementia-associated APOE ε4 allele is not associated with rapid eye movement sleep behavior disorder. *Neurobiol Aging* **49**, 218.e13-218.e15 (2017).

129. Ballabio, A. The awesome lysosome. *EMBO Mol Med* **8**, 73–76 (2016).

130. Webb, J. L., Ravikumar, B., Atkins, J., Skepper, J. N. & Rubinsztein, D. C. Alpha-Synuclein is degraded by both autophagy and the proteasome. *J Biol Chem* **278**, 25009–25013 (2003).

131. Mazzulli, J. R. *et al.* Activation of β-Glucocerebrosidase Reduces Pathological α-Synuclein and Restores Lysosomal Function in Parkinson's Patient Midbrain Neurons. *J Neurosci* **36**, 7693–7706 (2016).

132. Lara Ordóñez, A. J., Fasiczka, R., Naaldijk, Y. & Hilfiker, S. Rab GTPases in Parkinson's disease: a primer. *Essays in Biochemistry* **65**, 961–974 (2021).

133. Boecker, C. A., Goldsmith, J., Dou, D., Cajka, G. G. & Holzbaur, E. L. F. Increased LRRK2 kinase activity alters neuronal autophagy by disrupting the axonal transport of autophagosomes. *Current Biology* **31**, 2140-2154.e6 (2021).

134. Fivenson, E. M. *et al.* Mitophagy in neurodegeneration and aging. *Neurochem Int* **109**, 202–209 (2017).

135. Malpartida, A. B., Williamson, M., Narendra, D. P., Wade-Martins, R. & Ryan, B. J. Mitochondrial Dysfunction and Mitophagy in Parkinson's Disease: From Mechanism to Therapy. *Trends Biochem Sci* **46**, 329–343 (2021).

136. Lazarou, M. *et al.* The ubiquitin kinase PINK1 recruits autophagy receptors to induce mitophagy. *Nature* **524**, 309–314 (2015).

137. Rocha, E. M., De Miranda, B. & Sanders, L. H. Alpha-synuclein: Pathology, mitochondrial dysfunction and neuroinflammation in Parkinson's disease. *Neurobiology of Disease* **109**, 249–257 (2018).

138. Singh, A., Zhi, L. & Zhang, H. LRRK2 and mitochondria: Recent advances and current views. *Brain Res* **1702**, 96–104 (2019).

139. Li, H. *et al.* Mitochondrial dysfunction and mitophagy defect triggered by heterozygous GBA mutations. *Autophagy* **15**, 113–130 (2019).

140. McGeer, P. L., Itagaki, S., Boyes, B. E. & McGeer, E. G. Reactive microglia are positive for HLA-DR in the substantia nigra of Parkinson's and Alzheimer's disease brains. *Neurology* **38**, 1285–1285 (1988).

141. Hirsch, E. C. & Hunot, S. Neuroinflammation in Parkinson's disease: a target for neuroprotection? *Lancet Neurol* **8**, 382–397 (2009).

142. Pajares, M., I Rojo, A., Manda, G., Boscá, L. & Cuadrado, A. Inflammation in Parkinson's Disease: Mechanisms and Therapeutic Implications. *Cells* **9**, 1687 (2020).

143. Tan, E.-K. *et al.* Parkinson disease and the immune system - associations, mechanisms and therapeutics. *Nat Rev Neurol* **16**, 303–318 (2020).

144. Senkevich, K., Rudakou, U. & Gan-Or, Z. New therapeutic approaches to Parkinson's disease targeting GBA, LRRK2 and Parkin. *Neuropharmacology* **202**, 108822 (2022).

145. Hu, J. *et al.* Gene-based burden analysis of damaging private variants in PRKN, PARK7 and PINK1 in Parkinson's disease cohorts of European descent. *Neurobiology of Aging* **119**, 136–138 (2022).

146. Zhu, W. *et al.* Heterozygous PRKN mutations are common but do not increase the risk of Parkinson's disease. *Brain* **145**, 2077–2091 (2022).

147. Yi, W. *et al.* The landscape of Parkin variants reveals pathogenic mechanisms and therapeutic targets in Parkinson's disease. *Hum Mol Genet* **28**, 2811–2825 (2019).

148. Hollenbach, J. A. *et al.* A specific amino acid motif of HLA-DRB1 mediates risk and interacts with smoking history in Parkinson's disease. *Proc Natl Acad Sci U S A* **116**, 7419–7424 (2019).

149. Baka, Z. *et al.* Citrullination under physiological and pathological conditions. *Joint Bone Spine* **79**, 431–436 (2012).

150. Guen, Y. L. *et al.* Protective association of HLA-DRB1*04 subtypes in neurodegenerative diseases implicates acetylated Tau PHF6 sequences. 2021.12.26.21268354 Preprint at https://doi.org/10.1101/2021.12.26.21268354 (2021).

151. Luo, Y. *et al.* A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. *Nat Genet* **53**, 1504–1516 (2021).

152. Bodmer, W. & Bodmer, J. Evolution and function of the HLA system. *British medical bulletin* **34**, (1978).

153. Nakashima, M. *et al.* De novo CLCN3 variants affecting Gly327 cause severe neurodevelopmental syndrome with brain structural abnormalities. *J Hum Genet* **68**, 291–298 (2023).

154. Kenna, K. P. *et al.* NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. *Nat Genet* **48**, 1037–1042 (2016).

155. Wang, H. *et al.* NEK1-mediated retromer trafficking promotes blood–brain barrier integrity by regulating glucose metabolism and RIPK1 activation. *Nat Commun* **12**, 4826 (2021).

156. Styrpejko, D. J. & Cuajungco, M. P. Transmembrane 163 (TMEM163) Protein: A New Member of the Zinc Efflux Transporter Family. *Biomedicines* **9**, 220 (2021).

157. Rothaug, M. *et al.* LIMP-2 expression is critical for β-glucocerebrosidase activity and α-synuclein clearance. *Proceedings of the National Academy of Sciences* **111**, 15573–15578 (2014).

158. Eriksen, J. L. & Mackenzie, I. R. A. Progranulin: normal function and role in neurodegeneration. *J Neurochem* **104**, 287–297 (2008).

159. Nalls, M. A. *et al.* Evidence for GRN connecting multiple neurodegenerative diseases. *Brain Commun* **3**, fcab095 (2021).

160. Cruts, M. & Van Broeckhoven, C. Loss of progranulin function in frontotemporal lobar degeneration. *Trends Genet* **24**, 186–194 (2008).

161. Kleinberger, G., Capell, A., Haass, C. & Van Broeckhoven, C. Mechanisms of Granulin Deficiency: Lessons from Cellular and Animal Models. *Mol Neurobiol* **47**, 337–360 (2013).

162. Jian, J. *et al.* Progranulin Recruits HSP70 to β-Glucocerebrosidase and Is Therapeutic Against Gaucher Disease. *EBioMedicine* **13**, 212–224 (2016).

163. Kashyap, S. N., Boyle, N. R. & Roberson, E. D. Preclinical Interventions in Mouse Models of Frontotemporal Dementia Due to Progranulin Mutations. *Neurotherapeutics* **20**, 140–153 (2023).

164. Paciotti, S., Albi, E., Parnetti, L. & Beccari, T. Lysosomal Ceramide Metabolism Disorders: Implications in Parkinson's Disease. *J Clin Med* **9**, 594 (2020).

165. Senkevich, K. *et al.* GALC variants affect galactosylceramidase enzymatic activity and risk of Parkinson's disease. *Brain* **146**, 1859–1872 (2023).

166. Drobny, A. *et al.* The role of lysosomal cathepsins in neurodegeneration: Mechanistic insights, diagnostic potential and therapeutic approaches. *Biochim Biophys Acta Mol Cell Res* **1869**, 119243 (2022).

167. Pišlar, A. & Kos, J. Cysteine cathepsins in neurological disorders. *Mol Neurobiol* **49**, 1017–1030 (2014).

168. McGlinchey, R. P. *et al.* C-terminal α-synuclein truncations are linked to cysteine cathepsin activity in Parkinson's disease. *J Biol Chem* **294**, 9973–9984 (2019).

169. Blauwendraat, C. *et al.* Genetic modifiers of risk and age at onset in GBA associated Parkinson's disease and Lewy body dementia. *Brain* **143**, 234–248 (2020).

170. Andersen, M. S. *et al.* Heritability Enrichment Implicates Microglia in Parkinson's Disease Pathogenesis. *Annals of Neurology* **89**, 942–951 (2021).

171. Galatro, T. F. *et al.* Transcriptomic analysis of purified human cortical microglia reveals age-associated changes. *Nat Neurosci* **20**, 1162–1171 (2017).

172. Ransohoff, R. M. & Cardona, A. E. The myeloid cells of the central nervous system parenchyma. *Nature* **468**, 253–262 (2010).

173. Lou, N. *et al.* Purinergic receptor P2RY12-dependent microglial closure of the injured blood-brain barrier. *Proc Natl Acad Sci U S A* **113**, 1074–1079 (2016).

174. Mohtashami, S. *et al.* TOX3 Variants Are Involved in Restless Legs Syndrome and Parkinson's Disease with Opposite Effects. *J Mol Neurosci* **64**, 341–345 (2018).

175. Seksenyan, A. *et al.* TOX3 is expressed in mammary ER(+) epithelial cells and regulates ER target genes in luminal breast cancer. *BMC Cancer* **15**, 22 (2015).

176. Yeh, P., Walters, A. S. & Tsuang, J. W. Restless legs syndrome: a comprehensive overview on its epidemiology, risk factors, and treatment. *Sleep Breath* **16**, 987–1007 (2012).

Appendices

ELSEVIER LICENSE
TERMS AND CONDITIONS
Apr 20, 2023

This Agreement between Mr. Eric Yu ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	5533390509452
License date	Apr 20, 2023
Licensed Content Publisher	Elsevier
Licensed Content Publication	The Lancet
Licensed Content Title	Parkinson's disease
Licensed Content Author	Lorraine V Kalia,Anthony E Lang
Licensed Content Date	29 August–4 September 2015
Licensed Content Volume	386
Licensed Content Issue	9996
Licensed Content Pages	17
Start Page	896
End Page	912
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Title	Identification of variants, genes and pathways in synucleinopathies using bioinformatics and machine learning
Institution name	McGill University
Expected presentation date	Jul 2023

Portions	Figure 1
	Mr. Eric Yu 1033 Pine Ave W, Room 308
Requestor Location	
	Montreal, QC H3A 1A1 Canada Attn: McGill University
Publisher Tax ID	GB 494 6272 12
Total	0.00 USD

Terms and Conditions

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."

4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.

5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier

Ltd. (Please contact Elsevier's permissions helpdesk here). No modifications can be made to any Lancet figures/tables and they must be reproduced in full.

6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.

7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.

10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).

13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and

conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. **Translation**: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.

16. **Posting licensed content on any Website**: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.

17. **For journal authors:** the following clauses are applicable in addition to the above:

Preprints:

A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).

Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).

If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.

Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.

Authors can share their accepted author manuscript:

- immediately
    - via their non-commercial person homepage or blog
    - by updating a preprint in arXiv or RePEc with the accepted manuscript
    - via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
    - directly by providing copies to their students or to research collaborators for their personal use
    - for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- After the embargo period
    - via non-commercial hosting platforms such as their institutional repository
    - via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all

value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.

Policies for sharing publishing journal articles differ for subscription and gold open access articles:

Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version.

Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.

Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.

Please refer to Elsevier's posting policy for further information.

18. **For book authors** the following clauses are applicable in addition to the above: Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.

19. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions

You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our open access license policy for more information.

Terms & Conditions applicable to all Open Access articles published with Elsevier:

Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.

The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.

If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.

Additional Terms & Conditions applicable to each Creative Commons user license:

CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by/4.0.

CC BY NC SA: The CC BY-NC-SA license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article, provided this is not done for commercial purposes, and that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. Further, any new works must be made available on the same conditions. The full details of the license are available at http://creativecommons.org/licenses/by-nc-sa/4.0.

CC BY NC ND: The CC BY-NC-ND license allows users to copy and distribute the Article, provided this is not done for commercial purposes and further does not permit distribution of the Article if it is changed or edited in any way, and provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by-nc-nd/4.0. Any commercial reuse of Open

Access articles published with a CC BY NC SA or CC BY NC ND license requires permission from Elsevier and will be subject to a fee.

Commercial reuse includes:

- Associating advertising with the full text of the Article
- Charging fees for document delivery or access
- Article aggregation
- Systematic distribution via e-mail lists or share buttons

Posting or linking by commercial companies for use by customers of those companies.

20. **Other Conditions**:

v1.10
Questions? customercare@copyright.com.

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS
Apr 07, 2023

This Agreement between Mr. Eric Yu ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	5523950648677
License date	Apr 07, 2023
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Movement Disorders

Licensed Content Title	Analysis of Heterozygous PRKN Variants and Copy-Number Variations in Parkinson's Disease
Licensed Content Author	Eric Yu, Uladzislau Rudakou, Lynne Krohn, et al
Licensed Content Date	Sep 24, 2020
Licensed Content Volume	36
Licensed Content Issue	1
Licensed Content Pages	10
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Electronic
Portion	Full article
Will you be translating?	No
Title	Identification of variants, genes and pathways in synucleinopathies using bioinformatics and machine learning
Institution name	McGill University
Expected presentation date	Jul 2023
Requestor Location	Mr. Eric Yu 1033 Pine Ave W, Room 308 Montreal, QC H3A 1A1 Canada Attn: McGill University
Publisher Tax ID	EU826007151
Total	0.00 CAD
Terms and Conditions	

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.

- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, **and any CONTENT (PDF or image file) purchased as part of your order,** is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.

- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.**For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts,** You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.

- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is

granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed

waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.

- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY)

License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-

Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.

Further details can be found on Wiley Online Library http://olabout.wiley.com/WileyCDA/Section/id-410895.html

Other Terms and Conditions:

v1.10 Last updated September 2015

Questions? customercare@copyright.com.