Multi-omics and clinical study of Parkinson's disease and progression

Alejandro Martinez Carrasco

UCL Queen Square Institute of Neurology

For the award of Doctor of Philosophy

Declaration

I, Alejandro Martinez Carrasco, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Parkinson's disease (PD) is a progressive neurological condition that can be measured using clinical scales. Features of PD progression include motor and cognitive decline, as well as the emergence of motor fluctuations such as levodopa-induced dyskinesias. Although some patients follow a common progression trend, there is significant heterogeneity, with some patients progressing more quickly and exhibiting distinct clinical features. This heterogeneity is also notable at disease onset. I hypothesised that PD progression might be explained by common genetic variability.

During my PhD, I performed longitudinal genome-wide association studies (GWAS) to understand the genetic basis of motor progression and the time to develop dyskinesias, a motor fluctuation influenced by PD onset and chronic levodopa treatment. Additionally, I conducted a large-scale disease severity analysis using 10 different clinical instruments. For this large-scale analysis I used long-gwas, an end-to-end Nextflow pipeline to conduct cross-sectional and longitudinal GWAS.

Based on these GWAS approaches, I identified several loci significantly associated with prognosis, severity and survival. Applying functional annotation analyses to decode GWAS, I successfully nominated genes to be associated with the outcomes at most GWAS significant loci. I nominated the *ACP6* gene to be associated with the progression of axial PD motor features, and *MAD1L1* and *SOX9* genes to be associated with the severity of axial PD motor features. In addition, I nominated the *LRP8, XYLT1, and DNAJB4* genes as associated with the time to develop dyskinesias in PD. Notably, I validated three novel loci (*SERGEF*, *OTUD7A*, *SCN1A*) associated with the severity of hyposmia, alongside previously reported *LRRK2* and *GBA1* genes, involved in the autophagy-lysosomal pathway which may serve as surrogates for α-synuclein pathology.

Finally, I conducted a cell-type enrichment analysis of PD progression and susceptibility using publicly available longitudinal GWAS cell type expression data. We found a significant association between genes implicated in PD motor progression and microglia. Furthermore, we proposed a new framework for cell type enrichment that efficiently incorporates information about cis-regulation of gene expression.

Impact Statement

During my PhD, I have contributed to understanding the common genetic variability that influences clinical progression and presentation in Parkinson's disease (PD). Additionally, I have been involved in developing a pipeline that democratises this type of analysis. I have explored multiple methods to interpret results from genetic association studies and proposed novel ways to perform cell type enrichment analyses.

I hope that the progress achieved during my PhD contributes to the long journey of developing novel disease-modifying therapeutic avenues. We have identified several genes associated with motor progression and severity, as well as the survival time of levodopa-induced dyskinesias. We anticipate that these associations will be further tested in mice and cell models by academic collaborators to explore their potential impact on PD prognosis and to further characterise the involved pathways.

Some of my research outcomes have been used for the Aligning Science Across Parkinson's grant renewal. I have also presented my research findings at several conferences through poster presentations

Based on a large-scale multi-ancestry disease severity genetic study, I have shed new light on the genetic drivers of hyposmia, a feature that might be a surrogate for α-synuclein pathology. I hope these novel markers will be further tested and understood in relation to the LRRK2-GBA1 autophagy-lysosomal pathway. These new genes could be widely used in clinical practice, either by improving current diagnostic tools or by targeting the novel putative genes that might relate to α-synuclein pathology and spread.

I hope that long-gwas, a freely available end-to-end open-access pipeline for genetic association studies of severity and prognosis, will lead to a significant increase in findings in population genetics and provides a resource for investigators in multiple fields. With the automation of the most up-to-date approaches to account for confounding sources and efficient quality control, I envision an increase in novel loci proposed for further investigation in relation to disease modification strategies.

Finally, understanding which cell types are affected by genetic variants linked to traits brings benefits both inside and outside academia. When testing new disease-modifying therapeutics, it is crucial that the drug is delivered to the affected cell type in a disease state. In my research, I have developed an analysis on cell type enrichment of PD progression using relevant information. I have made my code available, and I hope further research in my lab will focus on enhancing and utilising these methods. This work will be relevant in building a cellular map to link cells and traits in PD.

Research Paper Declaration Form

UCL Research Paper Declaration Form 1

referencing the doctoral candidate's own published work(s)

1. **For a research manuscript that has already been published**

   a) **What is the title of the manuscript?** Genome-wide Analysis of Motor Progression in Parkinson Disease

   b) **Please include a link to or doi for the work** doi: 10.1212/NXG.0000000000200092

   c) **Where was the work published?** Neurology Genetics

   d) **Who published the work?** Wolters Kluwer

   e) **When was the work published?** August 8, 2023

   f) **List the manuscript's authors in the order they appear on the publication**
   Alejandro Martínez Carrasco, Raquel Real, Michael Lawton, Regina Hertfelder Reynolds, Manuela Tan, Lesley Wu, Nigel Williams, Camille Carroll, Jean-Christophe Corvol, Michele Hu, Donald Grosset, John Hardy, Mina Ryten, Yoav Ben-Shlomo, Maryam Shoai, Huw R. Morris.

   g) **Was the work peer reviewed?** Yes

   h) **Have you retained the copyright?** Yes - Under CC BY public copyright licence

   i) **Was an earlier form of the manuscript uploaded to a preprint server?**
   Yes - MedRxiv - https://doi.org/10.1101/2022.10.28.22281645

2. **For multi-authored work, please give a statement of contribution covering all authors**
   Huw R. Morris and Alejandro Martinez-Carrasco designed the study. Huw R Morris supervised the study. Donald Grosset, Michelle Hu, Yoav Ben-Shlomo, Michael Lawton, John Hardy and Huw R. Morris conceived and led the TPD and OPDC clinical cohorts, as well as performed data management and curation. Jean-Christophe Corvol conceived and led the DIGPD clinical cohort, as well as performed data management and curation. Alejandro Martinez-Carrasco performed all the quality control and analyses in the cohorts included in the present manuscript. Raquel Real supervised the trajectory of the research giving valuable suggestions on ways to move forward as the research progressed. Camille Carroll led the PD STAT study, as well as performed data management and curation. Manuela Tan, helped with the design of the quality control strategy. Lesley Wu provided access to an harmonised version of the AMP-PD genetic data. Regina Hertfelder Reynolds and Mina Rayten, helped with the colocalization analysis and the development of regional plots. Michael Lawton provided the equations to adjust the motor outcomes based on levodopa usage. Maryam Shoal reviewed many steps performed in the research process. Maryam Shoai provided the code to perform power calculation. Alejandro Martinez-Carrasco wrote the initial manuscript. All authors critically reviewed the manuscript.

3. **In which chapter(s) of your thesis can this material be found?** Chapters 2 and 3

UCL Research Paper Declaration Form 2

referencing the doctoral candidate's own published work(s)

1. **For a research manuscript that has already been published**

   a) **What is the title of the manuscript?** Genetic meta-analysis of levodopa induced dyskinesia in Parkinson's disease

   b) **Please include a link to or doi for the work** https://doi.org/10.1038/s41531-023-00573-2

   c) **Where was the work published?** npj parkinson's disease

   d) **Who published the work?** Springer Nature in partnership with The Parkinson's Foundation

   e) **When was the work published?** August 8, 2023

   f) **List the manuscript's authors in the order they appear on the publication**
   Alejandro Martinez-Carrasco, Raquel Real, Michael Lawton, Hirotaka Iwaki, Manuela M. X. Tan, Lesley Wu, Nigel M. Williams, Camille Carroll, Michele T. M. Hu, Donald G. Grosset, John Hardy, Mina Ryten, Tom Foltynie, Yoav Ben-Shlomo, Maryam Shoai, Huw R. Morris

   g) **Was the work peer reviewed?** Yes

   h) **Have you retained the copyright?** Yes - Under CC BY public copyright licence

   i) **Was an earlier form of the manuscript uploaded to a preprint server?** Yes - MedRxiv - doi: 10.1038/s41531-023-00573-2

2. **For multi-authored work, please give a statement of contribution covering all authors**
   Huw R. Morris and Alejandro Martinez-Carrasco designed the study. Huw R Morris supervised the study. Donald Grosset, Michelle Hu, Yoav Ben-Shlomo, Michael Lawton, John Hardy and Huw R. Morris conceived and led the TPD and OPDC clinical cohorts, as well as performed data management and curation. Alejandro Martinez-Carrasco performed all the quality control and analyses in the cohorts included in the present manuscript. Raquel Real supervised the trajectory of the research giving valuable suggestions on ways to move forward as the research progressed. Camille Carroll led the PD STAT study, as well as performed data management and curation. Manuela Tan helped with the design of the quality control strategy. Lesley Wu provided access to an harmonised version of the AMP-PD genetic data. Michael Lawton provided the equations to adjust the motor outcomes based on levodopa usage. Maryam Shoai reviewed many steps performed in the research process. Maryam Shoai provided the code to perform power calculation. Alejandro Martinez-Carrasco wrote the initial manuscript. All authors critically reviewed the manuscript. Tom Foltynie provided feedback on manuscript design. Hirotaka Iwaki helped with conceptualisation.

3. **In which chapter(s) of your thesis can this material be found?** Chapters 2 and 4

UCL Research Paper Declaration Form 3

referencing the doctoral candidate's own published work(s)

1. **For a research manuscript prepared for publication but that has not yet been published**

   a) **What is the current title of the manuscript?**

   Global large-scale analysis in Parkinson's disease using *long-gwas* provides new insights into the genetic determinants of Parkinson's disease phenotypes

   b) **Has the manuscript been uploaded to a preprint server?**

   No

   c) **Where is the work intended to be published?**

   Nature genetics or Nature communications

   d) **List the manuscript's authors in the intended authorship order**

   Alejandro Martinez Carrasco, Michael Ta, Oiher Serrano Asensio, Dan Vitale, Raquel Real, Alberto Imarisio, Hirotaka Iwaki, Huw R. Morris

   e) **Stage of publication**

   Circulating around collaborators and meeting open access policies before journal submission

2. **For multi-authored work, please give a statement of contribution covering all authors**
   Alejandro Martinez Carrasco led all the analyses and participated in the development of long-gwas, Michael Ta was the main long-gwas developer, Oiher Serrano Asensio helped building the web based long-was documentation, Dan Vitale provided ancestry inferences for AMP-PD, Raquel Real helped with the manuscript reformatting and writing as well as provided feedback to improve analyses, Alberto Imarisio helped with the manuscript reformatting and writing, Hirotaka Iwaki provided feedback to add corrections to the manuscript , Huw R. Morris and Alejandro Martinez Carrasco designed the study

3. **In which chapter(s) of your thesis can this material be found?** Chapter 2 and 5

UCL Research Paper Declaration Form 4

referencing the doctoral candidate's own published work(s)

1. **For a research manuscript that has already been published**

   a) **What is the title of the manuscript?** Mapping the Diverse and Inclusive Future of Parkinson's Disease Genetics and Its Widespread Impact

   b) **Please include a link to or doi for the work** https://doi.org/10.3390/genes12111681

   c) **Where was the work published?** Genes

   d) **Who published the work?** MDPI

   e) **When was the work published?** September 23, 2021

   f) **List the manuscript's authors in the order they appear on the publication**
   Alejandro Martinez-Carrasco, Inas Elsayed, Mario Cornejo-Olivas, Sara Bandres-Ciga

   g) **Was the work peer reviewed?** Yes

   h) **Have you retained the copyright?** Yes - Under CC BY public copyright licence

   i) **Was an earlier form of the manuscript uploaded to a preprint server?**
   No

2. **For multi-authored work, please give a statement of contribution covering all authors**
   Authors have contributed equally to writing and revising the manuscript. All authors have read and agreed to the published version of the manuscript.

3. **In which chapter(s) of your thesis can this material be found?** Chapter 7

Acknowledgements

I would like to thank the Global Parkinson's Genetic Program for funding this project. I also thank the entire Parkinson's community for their participation in research studies and the sites making efforts to collect their samples, which altogether make Parkinson's disease basic science possible.

I am deeply grateful to my primary supervisor, Prof. Huw Morris, for his tireless support in completing my PhD. Without his guidance and support, this would not have been possible. His dedication has shaped me as a researcher.

I also extend my gratitude to my secondary supervisors, Dr. Maryam Shoai, Manuela MX Tan, and Sir John Hardy, for their support and expert input throughout this project. In particular, I thank Maryam for her expertise in bioinformatics and statistics, which were crucial to completing this PhD. I also thank Manuela for providing materials and methods developed during her PhD, as well as for her valuable feedback.

A huge thank you to the Morris lab group I have worked with during my PhD, especially Raquel Real, Mary Makarious, Lesley Wu, Ana Luisa Gil Martinez, Oiher Serrano Asensio, Simona Jasaytite, Ellie Stafford, and Tatiana Georgiades.

I would like to thank Dr. Donald Grosset, Dr. Michele Hu, Prof. Camille Carroll, and Prof. Jean Christophe-Corvol for allowing me to access and analyse data from the Tracking Parkinson's and Oxford Discovery cohorts, PD-STAT, and DIGPD. I am also grateful for their active feedback on the analyses, results, presentations, and publications derived from my work.

I would like to thank Prof. Nigel Williams for his expert genetic advice on my project. Additionally, I thank Prof. Yoav Ben-Shlomo and Dr. Michael Lawton for their valuable statistical support.

I also thank colleagues at UCL who have assisted with various aspects of the project: Prof. Mina Ryten, Regina Reynolds, and Aine Fairbrother-Browne.

Thank you to the entire GP2 Network for supporting my work and allowing me to present at GP2-organised conferences and events.

A huge thank you to the GP2 Trainee Network, for providing valuable resources, opportunities, and training materials that were instrumental during the initial learning phase. Additionally, thank you to Dr Sara Bandres-Ciga, Sumit Dey, and Dr Teresa Perinan, with whom I have been delighted to lecture in person GP2 Bioinformatic workshops.

Thank you to the Cohort Integration Working Group in GP2 for collaborating to collect and harmonise data for PD bioinformatics research. Special thanks to Dr. Hirotaka Iwaki for trusting my work developing apps and strategies to improve data harmonisation and standardisation.

I am grateful to Dr. Andrew Singleton for the opportunity to visit the National Institutes of Health in 2023, which was an invaluable learning experience. I would especially like to thank Dr. Cornelis Blauwendraat, Dr. Hirotaka Iwaki, Dr. Mike Nalls, and Dr. Sara Bandres-Ciga for their support during and beyond my visit to NIH.

Finally, a huge thanks to my friends, my girlfriend, and my family for their encouragement and support throughout this journey.

List of tables

List of figures

Abbreviations

Age at baseline (AAB)

Age at diagnosis (AAD)

Age at Onset (AAO)

Allele frequency (AF)

Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq)

Cambridgeshire Parkinson's Incidence from GP to Neurologist (CamPaIGN)

Case-control (cc)

Chromatin immunoprecipitation sequencing (ChIP-seq)

Combined Annotation Dependent Depletion (CADD)

Cox proportional hazard models (CPH)

Credible Set (CS)

Deoxyribonucleic acid (DNA)

Early motor Parkinson's disease (EMPD)

Electronic health records (EHR)

Expression quantitative trait locus (eQTL)

Epworth Sleepiness Scale (ESS)

Gaucher's disease (GD)

Generalised linear model (GLM)

Genome wide association study (GWAS)

Global Parkinson's Genetics Program (GP2)

Genome-wide Analysis of Large-scale Longitudinal Outcomes using Penalization (GALLOP)

Han Chinese in Beijing, China (CHB)

Hardy-Weinberg equilibrium (HWE)

Hazard Ratio (HR)

Instrumental variable (IV)

Kaplan-Meier (KM)

Levodopa-induced dyskinesias (LiD)

Linear mixed effect model (LMM)

Mendelian randomization (MR)

Michigan Imputation Server (MIS)

Minor allele frequency (MAF)

Mini-Mental State Examination (MMSE)

Movement Disorders Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS)

Montreal Cognitive Assessment (MoCA)

NeuroBooster Array (NBA)

Northern and Western European ancestry (CEU)
Ordinary least-squares (OLS)
Parkinson’s disease (PD)
Parkinson's Disease Biomarkers Program (PDBP)
Parkinson's disease with dementia (PDD)
Simvastatin as a neuroprotective treatment for PD (PD-STAT)
Parkinson’s disease questionnaire (PDQ8)
Polygenic Hazard Score (PHS)
Postural Instability Gait DisordeR (PIGD)
Posterior Inclusion Probability (PIP)
Parkinson's Progression Markers Initiative (PPMI)
Proximity Ligation-Assisted ChIP-Seq (PLAC-Seq)
Posterior Probability (PP)
Principal component analysis (PCA)
Quantile-Quantile (QQ)
quantitative trait locus (QTL)
Rapid eye movement sleep behaviour disorder (RBD)
Receiver operating characteristic (ROC)
Schwab and England Activities of Daily Living (SEADL)
Standard Deviations (SD)
Standard Error (SE)
Substantia Nigra Pars Compacta (SNpc)
Oxford Parkinson's Disease Centre Discovery Cohort (OPDC
single-cell (sc)
Single nucleotide polymorphisms (SNPs)
Single Nucleotide Variant (SNP)
Single-nuclei (sn)
Simultaneous correction for empirical Bayesian estimates (SCEBE)
Sum of Single Effects (SuSiE)
The Adult Genotype Tissue Expression (GTEx)
Tracking Parkinson’s Cohort (PROBAND; TP)
Trans-Omics for Precision Medicine (TOPMed)
Transcripts Per Million (TPM)
Unique Molecular Identifiers (UMI)
University of Pennsylvania Smell Identification Test (UPSIT)
Utah residents with Northern and Western European ancestry (CEU)
Whole Genome Sequencing (WGS)

Table of contents

1) Introduction

a) Parkinson's disease

Parkinson's disease (PD) is one of the conditions included under the term "parkinsonism". Although it is the primary form of parkinsonism, Drug Induced Parkinsonism, Progressive Supranuclear Palsy, Dementia with Lewy Bodies, Multiple System Atrophy, Corticobasal degeneration, and Vascular Parkinsonism are other atypical forms of parkinsonism accounting for up to 15% of cases, and presenting with variable patterns of progression, treatment, signs, and symptoms. PD is a progressive neurodegenerative disorder that was first described by James Parkinson in 1817 [1]. It is the second most common neurodegenerative condition, thought to develop from an interaction of environmental and genetic factors [2]. However, PD is highly heterogeneous, with its symptoms and rate of progression varying between individuals [3]. Growing evidence suggests that genetics can explain some of the variability in progression [4]. Uncovering such genetic determinants would lead to further understanding of the molecular mechanisms of the condition and would help in identifying new targets for disease-modifying treatments.

i) Epidemiology

PD is a common condition affecting 6.1 million people worldwide (2016) [5]. The disorder impacts our society and health system. The prevalence and incidence of disease have increased in the past two decades [6].

Ageing is one of the strongest associations with PD risk [7,8]. Nevertheless, it is unclear if age-related cell death is the result of the chronic exposure to environmental toxins or instead a result of biological ageing [9]. It is possible that the complex interaction between environmental exposures, genetic changes and ageing lead to the underlying neurodegenerative condition [10]. Among PD patients, men have a higher incidence, prevalence and risk of mortality than women by a ratio of 1.4:1 [11], and this lower risk seems to happen at all ages [9]. Some potential explanations include men being more exposed to environmental risk factors or the protective role of female hormones [9]. Socioeconomic status is also a determinant factor of disease risk, with lower status associated with higher disease risk, reflecting a higher exposure to

adverse experiences during the group's lifespan [9]. There are also disease risk differences with respect to ethnicity, race, and geography, although according to Ben-Shlomo, they are difficult to explore due to existing inequities across groups as well as high quality data being limited to high-income regions [9]. The prevalence of PD is estimated to be lower in Africa, similar or lower in Asia, and similar in Latin America compared to Europe and North America [9]. The plausible reasons for differences in prevalence across regional groups are diverse including differences in genetic background or environmental exposures. A study on migrants has shown that PD prevalence estimates are higher in African and Japanese migrants living in the USA, than those living in their ancestral countries [12,13]. Multi-ethnic genetic studies have reported that the frequency and penetrance of genetic risk variants for PD differs across ethnic groups and geographical regions [9]. Nevertheless, differences in prevalence might also reflect difficulties in some geographical regions to access health care, complicating more extended diagnosis [9].

ii) Clinical and pathological aspects

From a clinical perspective, there are four cardinal features of Parkinson's: Bradykinesia, postural instability, rest tremor, and rigidity. Postural disturbance and gait freezing are recognised as motor parkinsonian symptoms related to advanced Parkinson's [14]. PD is not only a motor disorder. Before the syndrome is first diagnosed, rapid eye movement sleep behaviour disorder (RBD), olfactory problems, constipation, and depression can occur, and the correlation between PD and some of these early non-motor PD symptoms has been widely reported [15]. About a third of patients with RBD will develop PD within a decade [16,17]. In addition, as the course of the disease progresses, cognitive impairment becomes a prominent feature. Aarsland and colleagues performed a systematic review of studies that focused on the prevalence of dementia in PD [18]. They concluded that dementia affected 24 to 31% of all PD patients. Longitudinal studies support the increasing prevalence of dementia over time, and most Parkinson's patients with long disease duration will develop dementia [18]. These non-motor features, which are determinants of morbidity and poor quality of life, have been included by the Movement Disorder Society in the clinical diagnostic criteria for PD [19,20].

One pathological feature of PD is the loss of dopaminergic neurons in the substantia nigra pars compacta (SNpc) [21]. Clinico-pathological studies have shown a correlation between the loss of dopaminergic neurons at the SNpc and motor features such as bradykinesia and rigidity in the advanced stage of the disease [22]. Another prominent pathological feature is Lewy pathology, that is the aggregation of the abnormally folded α-synuclein protein. When α-synuclein misfolds, it becomes insoluble and accumulates to create intracellular inclusions known as Lewy bodies within the cell body and Lewy neurites within neuronal processes [23]. Another feature of PD pathology is neuroinflammation through the development of an active inflammatory response mediated by astrocytes and microglia [24].

iii) Molecular and immune mechanisms

PD develops as a result of synaptic dysfunction and neurodegeneration, with α-synuclein as the main protein accumulating and leading to the deposition of oligomers and fibrils, and the formation of Lewy bodies and Lewy neurites [25]. From a molecular perspective, this α-synuclein accumulation in different forms and shapes may relate to impaired mitochondrial and lysosomal function [19]. Mitochondria are organelles implicated in survival cell signalling and energy production. Mitochondrial dysfunction happens in the early stages of PD. Alterations of mitochondrial structure and dynamics lead to abnormal intracellular calcium levels, reduced ATP production and an increase in reactive oxygen species [19]. Both genetic and environmental factors which have a direct impact on mitochondrial homeostasis have been linked to PD and Parkinsonian disorders [26,27]. Moreover, there are processes involved in mitochondrial homeostasis that may have a key role in PD pathogenesis. Mitophagy is the process of selectively removing damaged or redundant mitochondria through their signalling for lysosomal degradation. For instance *PRKN* and *PINK1* are PD causing genes involved in mitochondrial quality control which regulate mitophagy mechanisms [28]. These mitochondrial health processes are central in neurons, cells with high energy requirements [19]. Lysosomes, another type of organelles, are involved in mitophagy In addition, they are important in processing protein aggregates. Therefore, a malfunctioning lysosome can lead to an increased α-synuclein oligomer accumulation as well as impaired mitochondrial activity. The genetic evidence that PD risk is related

to lysosomal activity is mainly centred on the *GBA1* gene, which encodes the lysosomal enzyme acid β-glucocerebrosidase [19].

iv) Clinical progression

PD progression can be described on the basis of Kalia and Lang's description [21]. PD has a premotor or prodromal phase with prominent non-motor features such as RBD and olfactory dysfunction. PD diagnosis coincides with the onset of the classical motor symptoms. Then, progression follows the trend of worsening motor features over time and the prescription of symptomatic treatment, which as a side effect will end up leading to a characteristic trend of complications in the long term, including non-motor fluctuations, dyskinesias, and psychosis. The late stage of the disorder is characterised by motor and non-motor features resistant to treatment such as freezing of gait, falls, dysphagia and speech dysfunction. Autonomic symptoms are also prominent during the late disease stages. In addition, dementia is also characteristic of the late disease stage, affecting 83% of patients with PD who have had 20 years of disease duration [21].

Our ability to understand disease progression comes from long term outcomes reported from observational studies such as the Sydney Multicenter Study of levodopa naive, short disease duration idiopathic PD patients [29]. 15 years from study initiation, 48% of patients experienced dementia, 36% had mild cognitive impairment, 94% had experienced dyskinesias, 56% had developed dystonia. In addition, 81% had experienced falls, hallucinations and depression were experienced by 50%, choking was experienced in 50%, symptomatic postural hypotension in 35%, urinary incontinence in 41%, and 65% of the cohorts had died [30]. After 20 years, the main problems were related to non-levodopa responsive features characteristic of some PD patients. Dementia was present in 83% of patients, and 74% of patients had died [31]. This suggests that most patients follow this progressive decline in the clinical motor and non-motor hallmarks of the condition.

However, disease progression is heterogeneous and not all patients develop all the features in a uniform way. Presentation at disease onset and progression of motor features vary widely between patients. PD empirical subtypes have been proposed according to the motor symptoms. The two main subtypes are tremor-dominant PD

(prominent tremor related impairment) and non-tremor-dominant PD (phenotypes such as akinetic-rigid syndrome and postural instability gait disorder - PIGD) [32]. There is also an intermediate phenotype with prominent tremor and non-tremor related motor symptoms. The prognosis and the progression varies across patients divided in these groups. Tremor-dominant PD patients display on average a slower rate of progression and the functional disability is not as severe as compared with non-tremor PD (PIGD) [21]. But these are not the only proposed PD subtypes with distinct progression patterns. In the past two decades, PD cluster analyses have aimed to subgroup patients using data-driven approaches [33–39]. Marras and Lang commented that PD cluster analyses should take into account the effect of disease duration on defined subtypes. If disease duration is not taken into account, an identified PD subtype could simply be a variable stage over the PD course [40]. For instance Graham and Sagar, who identified several clusters based on cognition and motor performance, realised that there were differences in the average disease duration, proposing that their three short disease duration clusters would evolve into two motor only impairment and cognitive and motor impairment long disease duration clusters [33]. PD subtypes may reflect differences in underlying biological mechanisms, which could further refine our understanding of disease heterogeneity [41].

The progression of PD is likely to mirror the underlying neuropathology, as suggested by the correlation between clinical and neuropathological features. When PD is diagnosed there are prominent motor symptoms, reflecting substantial dopaminergic neuronal loss in the substantia nigra [42]. A more updated view by Marras and colleagues proposes that the onset of disease occurs when a substantial proportion of dopaminergic terminals in the basal ganglia are lost [43]. The most influential view of PD progression based on neuropathology came in 2002, when Braak and colleagues proposed a progression staging scheme based on the α-synuclein pathological inclusions and its differential distribution in the brain [44,45]. They hypothesised damage in some brain areas to explain some of the progression hallmarks in PD. For instance, they proposed that Braak stages 5 and 6 (Lewy body pathological spread affecting limbic and neocortical structures) might explain impaired cognition among PD patients. Years later, Braak and colleagues found a correlation between the cognitive decline and the stage of Lewy Body pathology according to the

staging system [44,45]. Dementia in PD cases has been found to correlate with cortical Lewy bodies [46,47]. However, some other studies have not been able to replicate these findings [48]. Similarly, the neuropathological assessment in the Sydney Multicentre Study, revealed that some young onset PD cases with Lewy body pathology fit Braak staging, whereas some other rapidly progressive patients showed diffuse Lewy body disease [49]. These findings further support the heterogeneity of PD progression. Patterns of clinical progression not matching these proposed pathological stages might be a surrogate for yet unknown pathological hallmarks of PD not involving Lewy Body spread in the brain.

It is clear that α-synuclein aggregation is a hallmark of the main form of PD and its progression and spread. However, since the discovery of SNCA, other PD-associated loci have been discovered, which might explain pathogenic processes in forms of PD which lack Lewy bodies and Lewy neurites at autopsy [23]. In general terms, progression might relate to intrinsic cellular factors such as mitochondrial and lysosomal function or factors related to spread of a toxic protein.

b) Genetics of Parkinson's disease

PD aetiology can be understood in part through genetics. Even though there is no consensus on the quantification of the heritable component of the disorder and how this inheritance happens when PD causing and risk associated variants segregate together, the most updated estimates range between 22% and 40% [19]. Concordance rates in twin studies suggest the heritability of PD is 30% [50]. Heritability estimates based on PD common genetic variability exclusively, suggests the heritable component of idiopathic PD is 22%, of which only a fraction (16-36%) is explained by the largest PD risk genome-wide study [51].

PD genetics is complex, and it is likely that we have only uncovered a fraction of it. In addition, environmental and behavioural factors play a role on PD aetiology, and some of these factors may well interact with genetics (for example directly regulating gene expression, or shaping the epigenome), therefore making the PD genetic puzzle harder to complete as a result of the multidimensional spectrum causing the condition [9].

Regarding the genetics of Parkinson's, there are rare mutations that segregate in families and are known to cause PD [52,53]. These mutations are defined as rare as their frequency is very low in the general population (frequency < 1%). Nevertheless, taking them all together, the number of cases with a reported PD causing rare mutation is relatively common. Up to 15% of PD patients have a positive PD family history and 5–10% of these familial cases may have Mendelian inheritance [54]. However, PD genetics does not only span rare mutations causing disease. In addition, there are PD associated common genetic variants, that is genetic variants that have a frequency > 1% in the general population, which can be either protective or disease-causing [51]. Even though they are not disease causing mutations, when some individuals carry some of these common variants, they may develop Parkinson's related to a polygenic effect. PD risk variants have different magnitudes of effect, directionality, frequency, deleteriousness, and penetrance [4]. The vast majority of patients with PD are diagnosed as sporadic without a clear genetic (familial) cause [51].

i) Mendelian mutations in PD

The past few decades have witnessed the discovery of recessively and dominantly inherited genes responsible for rare monogenic forms of PD. Well-known, highly penetrant autosomal dominant mutations causing PD are found within the *SNCA*, *LRRK2* and *VPS35* genes, and autosomal recessive disease causing mutations are found in the *PRKN*, *DJ-1*, and *PINK1* genes [55].

According to Blauwendraat and colleagues, the term monogenic for PD is an oversimplification. Some carriers of any of these highly penetrant PD mutations may not develop PD (known as incomplete penetrance), which suggest that other genetic and environmental factors influence disease aetiology together with the well characterised highly penetrant mutations [4].

The first mutations in PD families was described by Polymeropoulus and colleagues. They found a mutation (p.A53T) in the fourth exon of *SNCA*, a gene located on chromosome 4, in a large Italian family, and replicated their findings in 3 unrelated Greek families with PD [56]. Subsequently, Singleton and colleagues examined a large family with autosomal dominant PD, and carried out quantitative real-time PCR amplification of *SNCA* exons to find an increase in gene dosage consistent with a gene

triplication [57]. Similarly, duplications of the α-synuclein gene have been found to cause PD [58].

Interestingly, as described earlier *SNCA* is the major component of Lewy bodies, and a reduction in the solubility of α-synuclein leads to the formation of filaments from insoluble alpha synuclein that aggregate into cytoplasmic inclusions, which contribute to the death or dysfunction of glial cells and neurons [59]. Although its neuronal function is unknown, it may have a role in synaptic vesicle dynamics, mitochondrial function, intracellular trafficking and might be a potential chaperone. α-synuclein acquires neurotoxic properties when it aggregates into insoluble α-synuclein fibrils characteristic of Lewy pathology [60]. α-synuclein may aggregate due to its overproduction as a result of the gene duplications and triplications that Singleton and colleagues, and Ibanez and colleagues described as mutations causing autosomal dominant PD. Another reason could be mutations in domains that lead to protein misfolding and oligomerization or alteration on the molecular pathways in which α-synuclein takes part [19,61].

Based on frequency, specific rare variants in *LRRK2* are the most important Mendelian cause of late-onset autosomal dominant PD, with a mutation frequency ranging from 2-40% depending on the population [55], which may reflect genetic diversity among different ethnic groups and geographical regions, as well as variability in sample sizes, study designs or mutation screening techniques of genetic studies. *LRRK2* G2019S is the most well characterised *LRRK2* disease causing mutation in PD [62–66]. The G2019S mutation occurs most commonly in European, North African, and Jewish families. G2019S is estimated to account for up to 30% of inherited PD cases in certain populations [67]. However, there are other pathogenic mutations at the *LRRK2* gene more frequent in Asian populations, such as the N1437D mutation in Chinese families, and I2020T in Japanese families [19,68]. In addition, G2019S mutation penetrance varies across age stratified groups, increasing up to 85% at 70 years old [67]. Interestingly, this variation in penetrance seems to be independent of the individual's ancestry at a fixed age of 80 years [69].

Different *LRRK2* missense mutations have been reported as disease-segregating mutations, and patients that harbour mutations have dopaminergic degeneration according to findings from *LRRK2*-autopsy cases [70]. The *LRRK2* encoded protein is

involved in autophagy (the process through which cells transports cytoplasmic components to the lysosome for their degradation and recycling), lysosomal function, and vesicular trafficking [71]. *LRRK2* protein pathogenic mutations concentrate around the kinase and GTPase domains, and disease-causing mutations in these regions increase protein kinase activity. A study using a rodent model expressing the human G2019S *LRRK2* sequence in neurons, has shown that dopaminergic neurons are progressively lost in the substantia nigra, which is associated with the level of *LRRK2* kinase activity [72]. Several groups have tried to explain the molecular mechanisms of pathogenic *LRRK2* mutations. Zimprich and colleagues firstly hypothesised that *LRRK2* may phosphorylate α-synuclein and tau proteins, leading to the accumulation followed by aggregation of unfolded α-synuclein and tau proteins in dying neurons [73]. *LRRK2* has several putative protein-protein interaction domains so it is plausible that mutations that alter these domains affect the contact with other proteins. In addition, dysfunction through mutations in the *LRRK2* kinase domain would also lead to changes on the proteins *LRRK2* might interact with and phosphorylate such as α-synuclein and tau proteins [55].

A more recent hypothesis by Alessi and Sammler based on data from recent years holds that *LRRK2* does regulate autophagy. This process seems to also be controlled by a subgroup of RAB family proteins, that are phosphorylated by *LRRK2* kinase, which ensure homeostasis and unaltered autophagy [71]. In addition, inflammation is also regulated by *LRRK2* and there are high levels of expression in immune cells such as macrophages and monocytes. Mouse models expressing the G20192 *LRRK2* mutation have been found to be protected from infection. In contrast, mice lacking *LRRK2* or expressing *LRRK2* inhibitors, have been found to be unable to clear out infections [74]. Understanding the protective role of *LRRK2* against infectious diseases, and more specifically, knowing if an antagonistic pleiotropy event occurs between *LRRK2* pathogenic mutations and PD and immunity, might be transferable knowledge to shed new light into PD disease aetiology. Whether *LRRK2* inhibitors have disease modifying effects on PD patients carrying pathogenic *LRRK2* mutations or in sporadic PD is under active investigation in ongoing drug trials [71].

Two other genes, *PARKIN* and *PINK1* have been found to cause early onset autosomal recessive PD [75,76]. *PARKIN* encodes an E3 ubiquitin ligase with an

amino-terminal ubiquitin-like domain and carboxyl-terminal ubiquitin ligase domain and resides in the cell's cytosol. *PINK1* encodes a serine-threonine protein kinase that localises to mitochondria. Both proteins work in the same pathway and participate in maintaining mitochondrial homeostasis [77], an organelle that as we already mentioned, is thought to play a central role in PD aetiology. First evidence of the link between mitochondrial dysfunction and PD was based on evidence from people developing the disease after illicitly using methyl-4-phenyl-1,2,5,6-tetrahydropyridine (MPTP) [78]. MPTP was found to oxidise to 1-Methyl-4-phenylpyridinium (MPP+), which causes the inhibition of complex I in the mitochondrial respiratory chain after its selective uptake in dopaminergic neurons [79].

ii) Common variation in PD

Genome wide association studies (GWAS) have been a powerful tool to better understand how genetic variability contributes to the development of disease. GWAS usually focuses on genetic variants with a minor allele frequency (MAF) higher than 1 or 5%. Therefore GWAS allows us to understand the common genetic architecture of complex diseases such as PD. In the past decades, several large association studies gathering samples of European ancestry have been conducted revealing genetic variants increasing the risk for PD [51,80,81]. In 2019, Nalls and colleagues conducted the largest GWAS to date totalling 37,000 cases, 18,600 UK Biobank proxy-cases (where the individual had a parent affected by PD), and 1,400,000 controls of European ancestry, and 7,800,000 single nucleotide polymorphisms (SNPs). This large scale study revealed 90 independent SNPs significantly increasing the risk of developing PD. They estimated based on a PD polygenic risk score (PRS) that the total SNP based heritability uncovered from their analysis, is about one third of the total SNP heritability of the condition [51]. This estimate suggests there is still a high percentage of the common heritability yet to be discovered. In addition, they performed several pathway, tissue, and cellular enrichment analyses across genes near the PD risk variants and found that genes were enriched in the brain. Interestingly, the expression of the selected PD genes were enriched in neuronal cells. Some of the 90 independent PD risk variants fell close to monogenic Parkinson's genes such as *SNCA*, *LRRK2*, *GBA1*, and *VPS13C*. The strongest associations found by Nalls and colleagues were at the *SNCA* and *MAPT* loci. This large PD case-control GWAS meta-

analysis has been recently expanded by Kim and colleagues incorporating several diverse ancestry populations. They identified 12 potentially novel risk loci, 9 that were shared across all ancestries and three that had heterogeneous effects across the different ancestry groups, hence were ancestry-specific. Based on fine-mapping they nominated 6 putative causal variants at 6 loci previously linked to PD [82].

Another large GWAS in PD patients of Asian ancestry, gathering 6,724 patients 24,851 healthy controls, identified 11 GWAS loci, reaching genome wide significance 9 of which overlapped with the Nalls and colleagues' European ancestry GWAS. Of the 78 SNPs nominated from the European GWAS, and that were polymorphic in the Asian GWAS, 63 (80.8%) were found to have the same directionality of association and 15 (19.2%) had an opposite direction. This suggests that there is an overlap in the common variability between the two different ancestry groups as well as consistent effect sizes, but also some differences such as 2 PD genetic risk factors reaching genome-wide significant only in the Asian ancestry specific cohort, as well as differences in LD haplotypes and allele frequency (AF) [83]. The presence of non-overlapping risk variants across ancestries is also shown in an analysis that gathered samples of Chinese ancestry. They genotyped several variants at four loci that have been reported to modulate the risk for PD (*SNCA*, *PARK16*, *LRRK2*, *BST1*). They found consistent effects of *SNCA*, and *LRRK2* variants and the risk for PD. However, they found *PARK16* variants to be associated with a lower PD risk. They did not find any effect between variants within the *BST1* locus and the risk for PD [84].

More recently, the largest GWAS in patients of African ancestry revealed a variant within *GBA1*, as the most significant genetic risk factor for PD in African and African admixed populations. They identified changes in *GBA1* expression which lead to decreased glucocerebrosidase activity (the protein encoded by *GBA1*), hence suggesting those expression changes as the potential disease mechanism increasing the risk in PD in African and African admixed populations. In a separate analysis, they further characterised the functional effect of the *GBA1* non-coding risk variant. They found that this variant, which is a key intronic branchpoint, alters the splicing of functional *GBA1* transcripts, reducing the levels of the protein, hence the activity [85]

GBA1 encodes the lysosomal enzyme glucocerebrosidase, and its dysfunction is linked to Gaucher’s disease (GD). GD is an autosomal recessive disorder and more

than 300 mutations, including insertion, deletion, point and missense mutations have been reported to cause the condition [86]. Moreover, some *GBA1* mutations both biallelic and single mutations, are a risk factor for PD.

Different studies conducted in PD cohorts have found heterogeneity in the clinical presentation of patients carrying *GBA1* mutations. There is evidence suggesting that *GBA1* pathogenic mutation carriers have a distinct clinical presentation compared to non-carriers. A well-powered study in a large UK observational cohort with 2.5% *GBA1* GD causing pathogenic mutation carriers, and 6.2% *GBA1* non-synonymous PD variants, with L444P being the most common pathogenic mutation, suggested that patients with *GBA1* mutations were 5 years younger at PD AAO compared to non-carriers. Moreover, *GBA1* mutation carriers more commonly had a greater risk of cognitive impairment, poorer response to dopaminergic treatment, lower α-synuclein levels as well as increased disease severity (higher Hoehn and Yahr -HY- score, a measure of disease severity) [87]. Other studies in non-European populations have reported consistent findings for age at onset, showing that *GBA1* mutation carriers are younger at PD onset [88,89]. However, there are discrepancies in the motor and cognitive presentations linked to *GBA1* carriers. A large multicentre study did not find differences in motor presentations between *GBA1* pathogenic mutation carriers and non-carriers [90]. Similarly, other well-powered studies have found an association between *GBA1* mutation carriers and development of dementia as well as cognitive decline [91,92].

The penetrance of *GBA1* PD genetic variants were estimated to range from 7.6% at 50 years to 29.7% at 80 years, based on the kin-method, an approach that leverages family data to calculate the probability that an individual with a certain genotype will show a particular phenotype, helping to assess the degree of penetrance. The penetrance of *GBA1*-PD variants is higher than that estimated on *GBA1*-GD patients and their relatives [93]. Blauwendraat and colleagues investigated genetic modifiers of *GBA1*-associated PD penetrance, using case-control GWAS based on *GBA1* mutation carrier status. Among the 90 independent variants found on the Nalls 2019 cc-GWAS meta-analysis, they found a strong association between the rs356219 polymorphism that passed Bonferroni correction and *SNCA* locus [94]. This *SNCA*-*GBA1* link and risk for PD is also plausible given that there is also a biological link

between glucocerebrosidase and α-synuclein, as they have been shown to interact in vitro as well as to influence the intracellular levels and processing of each other [95,96].

We have previously highlighted mutations at the *LRRK2* and *SNCA* genes that segregate in families and cause PD. In addition, common genetic variation at the *LRRK2* and *SNCA* loci has been associated with the risk of developing sporadic PD in [51]. Mutations that increase PD risk are located in the protein-coding gene and also in non-coding regions [67,97]. These non-coding variants are likely to have a biological effect based on the modulation of gene splicing and/or expression [98].

iii) The genetic architecture of PD progression

When studying PD genetics, there is an additional layer of complexity. Apart from the known disease causing Mendelian mutations [4,55] as well as common genetic variants that increase the risk of developing sporadic PD [51], it is possible that unknown common and rare genetic variants contribute to the high heterogeneity in progression trajectories. One key reason to make such a hypothesis is that some candidate PD genes analyses have successfully been associated with distinct progression trends. For instance, prospective studies looking at differences between *LRRK2* mutation carriers and non-carriers have found that patients carrying the G2019S mutation showed a slower motor decline [99]. Another study comparing *GBA1* mutation carriers versus non carriers showed the cohort carrying the mutation to have a more severe cognitive and motor decline [100]. It is clear that there are PD mutations that correlate with specific longitudinal PD traits.

Whether known PD risk genetics is also associated with PD progression is not fully understood yet. Iwaki and colleagues looked at the association of 31 PD risk SNPs with PD progression. Those 31 SNPs were nominated from three major PD risk studies showing variants that were significantly associated with PD risk [101–103]. Then, they looked at the association of these independent SNPs with PD clinical features. They used data from a total of 23,423 visits by 4,307 patients of European ancestry from 13 longitudinal cohorts. Variants in the *GBA1* gene were linked to daytime sleepiness and potential RBD changes. Furthermore, researchers identified a connection between the *GBA1* variant p.N370S and treatment-related challenges such as

wearing-off and dyskinesia. They also confirmed links between *GBA1* variants and declines in motor and cognitive functions. Additionally, genotype-phenotype associations were observed, including an intergenic variant near *LRRK2*, which was associated with accelerated motor symptom progression, and an intronic variant in *PMVK* associated with the emergence of wearing-off effects, which refer to the gradual return of motor and non-motor symptoms as the effectiveness of levodopa or other dopaminergic medications diminishes [104]. Therefore, the overlap between the common genetics of PD risk and progression is only partial and some common genetic variants influencing disease course might be unknown. However, a major limitation of this study is that it did not use the most up to date results from 90 independent risk loci associated with PD from the study conducted by Nalls and colleagues [51,80,81].

In the past years, there have been a number of GWASs performed to explore the effect of genetic variation on disease presentation and progression. The largest GWAS of PD AAO on 28,000 patients with PD showed that not all 90 PD risk variants are associated with AAO [105]. They found two genome-wide significant signals related to younger disease onset at the known PD risk loci *SNCA*, and the protein-coding gene *TMEM175*. In spite of the smaller sample size compared to the largest PD risk GWAS, which decreases the power of the study to reveal the complete PD AAO genetic heritability, the authors found a significant effect in only 6 loci based on a targeted analysis looking at the 44 SNPs that were genome-wide significant in the Chang and colleagues PD GWAS meta-analysis [106]. According to Blauwendraat and colleagues, based on AAO GWAS results, the mechanism that lead to early PD onset could be related to *SNCA* pathology, since *TMEM175* has been associated with increased α-synuclein aggregation and an increase in α-synuclein expression might also lead to an increase in α-synuclein aggregation [4]. A more recent study, led by the COURAGE-PD Consortium, added to the previous PD AAO meta-analysis a PD cohort of 8,535 PD patients of predominantly European ancestry, which led to the validation of the previously reported *SNCA* locus as well as the discovery of a novel locus, *BST1,* significantly associated with an earlier AAO [107]. This is a clear example showing that the genetic make-up of sporadic PD cases not only involve genetic variants that confer risk of disease but also the non-overlapping genetic determinants leading to differences on PD presentation as determined by the age at disease onset.

Liu and colleagues investigated whether genetics contributes to cognitive decline in PD. They accessed data from 15 different cohorts making up a total of 4,872 patients covering 36,123 study visits and carried out a longitudinal genome-wide survival study (GWSS) approach. They used Cox proportional hazard models (CPH) with covariate adjustment to investigate the influence of common and low-frequency genetic variants on cognitive decline by measuring the time to reach a Parkinson's disease with dementia (PDD) outcome from disease onset. They found a genome-wide significant association signal at the *RIMS2* locus with progression to PDD with a HR = 4.74. They further validated the association signal using a linear mixed effect model (LMM) and a different measure of global cognitive function in PD, the Mini Mental State Exam. They found patients carrying the lead *RIMS2* variant to decline more rapidly over time compared to non-carriers. *RIMS2* encodes the regulating synaptic membrane exocytosis 2 protein, a RIM family member, which is involved in docking and priming of presynaptic vesicles. In addition, they defined sub-threshold P-Values to investigate the overlap between PD susceptibility variants and the variants nominally associated with progression to PDD end point. They examined 505 variants, and none of them were significantly associated with susceptibility to PD, which suggests that there is little overlap between the genetic determinants of PD susceptibility and progression to PDD. Finally, they looked at the effects of *GBA1* and *APOE* on risk of dementia in patients with PDD. They found that patients carrying the *APOE* ε4 or a *GBA1* pathogenic mutation for Gaucher's disease or protein-coding variants associated with PD showed a faster cognitive decline compared to non-carriers. Finally they derived a polygenic hazard score (PHS) using the lead variant from each of the three prognosis loci they found to reach significance (including *RIMS2* lead variant), as well as with the inclusion of *GBA1* and *APOE [108]*. However, the analysis performed by Liu and colleagues has not been replicated [109], suggesting the association they found in their analysis could be driven by just one subset of the data (ie one large cohort with different inclusion criteria than the rest could lead to a more homogeneous profile compared to the rest of the cohorts), and not representative of the more general PD population. Nevertheless, it is worth highlighting they managed to derive a GHS comprising significant associations and develop predictive models that performed well in external cohorts that were not part of the meta-analysis.

Real and colleagues designed a study with a similar power than Liu's and colleagues in terms of sample size. They made use of the same definition of cognition to explore genetic variants that influence progression to PDD through a GWSS using CPH models and data from four longitudinal PD cohorts. They found three genome-wide significant loci, and the most significant SNP was the *APOE* ε4 allele-tagging SNP. *APOE* stands out as the primary genetic risk for Alzheimer as well as an earlier age of onset of disease [110]. Numerous studies have additionally demonstrated its involvement in cognitive deterioration and dementia among individuals with Parkinson's disease [111–113]. IIn addition, they found a novel association at the *LRP1B* locus, a receptor for *APOE*-carrying lipoproteins which is highly expressed in the adult human brain. Based on *APOE* and *LRP1B* interaction analyses, they found that carriers of both *APOE* ε4 and *LRP1B* rs80306347 risk alleles had a higher hazard of progression to PDD (HR = 8.08, 95% CI = 4.64–14.1, $P = 1.55 \times 10{-}13$) compared to carriers of *LRP1B* s80306347 (HR = 2.33, 95% CI =1.34–4.05, $P = 0.00273$) and *APOE* ε4 alleles separately. However, they did not find a significant interaction effect between the two alleles in a separate regression model, suggesting that the relationship between each allele and the outcome is likely independent, meaning the variables do not interact in a meaningful way in explaining the variation in the progression to dementia in PD. Moreover, a survival analysis was conducted controlling for *APOE* status. The findings revealed an elevated hazard of progressing to PDD among carriers of the *LRP1B* rs80306347 variant, which confirms that the impact of rs80306347 is independent of the influence of *APOE*. Finally, they could validate the effect of *GBA1* variants in PDD through a candidate gene analysis, which supports the idea of *GBA1* increasing the risk of progression to dementia [109].

Other analyses with different study designs have also investigated PD progression and how it is associated with genetics more exhaustively. Iwaki and colleagues studied the genetic impact on the trajectory of PD-related phenotypes using longitudinal data from 12 longitudinal cohorts in a total of 4,093 patients with and carried out GWAS for 25 cross-sectional and longitudinal phenotypes. They divided the analyses based on whether the PD progression outcomes were gathered under a continuous or binomial category. For continuous outcomes, they assessed progression through the longitudinal quantitative or ordinal scores of Hoehn and Yahr (HY), total and sub-scores of Movement Disorders Society Unified Parkinson's Disease Rating Scale

(MDS-UPDRS), Mini-Mental State Examination (MMSE), Montreal Cognitive Assessment (MoCA), and Schwab and England Activities of Daily Living (SEADL). They used LMMs to evaluate the association of variants for each of these continuous traits. For binomial outcomes, they assessed progression based on constipation, cognitive impairment, depression, daytime sleepiness, HY stage > 2, hyposmia, insomnia, motor fluctuation, RBD, restless legs syndrome, and a SEADL < 70. They used a combination of logistic regression and Cox proportional hazard models for binomial outcomes depending on outcome development rate at baseline to assess the influence of variants for each of the binomial outcomes. They found two variants reaching genome wide significance. An *SLC44A1* intronic variant was associated with reaching HY>2 more quickly. They also found an intergenic variant in chromosome 10 to be associated with a lower prevalence of insomnia at baseline. This variant is a significant expression quantitative trait loci (eQTL) for the α-2A adrenergic receptor. In candidate gene analysis, they replicated previous reports of *GBA1* coding variants (rs2230288: p.E365K; rs75548401: p.T408M) being associated with greater motor and cognitive decline over time, and an *APOE* ε4 tagging variant (rs429358) being associated with greater cognitive deficits in patients [114].

Tan and colleagues also explored the progression of Parkinson's using data from 3 deeply phenotyped longitudinal cohorts, totalling 3,364 patients with 12,144 observations (mean follow-up 4.2 years). Instead of looking at individual clinical assessments, they made use of principal component analysis (PCA) to derive the outcomes that would be later used in GWAS. They came up with a composite measure (PCA gathering motor and cognitive assessments in PD), as well as a motor and a cognitive score for each patient at each time point. They came up with the residual slopes of the PCA derived scores to remove the variance of progression trajectories explained by confounders using linear mixed effect models. Finally, they conducted a GWAS on the residual slopes through a multiple regression analysis. They managed to replicate previous findings regarding the association of *APOE* ε4 with a worse dementia progression pattern. In addition, they identified a novel signal in *ATP8B2* associated with motor progression based on a MAGMA gene-based analysis [115]. This gene encodes an ATPase phospholipid transporter (type 8B, member 2) and had never been reported to be associated with PD before. In addition, based on targeted analysis of PD risk variants, they found *GBA1* p.E326K to be nominally associated

with composite and cognitive progression, consistent with what Iwaki and colleagues found [116]. Unfortunately, they could not replicate the finding for the *SLC44A1* variant that was associated with progression to HY>2 in the Iwaki GWAS.

There are also other studies that have attempted to look at the progression of PD but with limited power due to small sample size. Ju Chunk and colleagues undertook survival analyses using Cox proportional hazard models. They found two associations that did not reach significance after Bonferroni correction, one with survival to cognitive decline (*CLRN3;* HR = 2.03, 95% CI 1.47–2.79, p = 4.08e−6), and the other with survival to motor decline (*C8orf4;* HR = 1.81; 95% CI = 1.42–2.31; p = 1.51e−6) [117]. None of these associations were replicated by the larger longitudinal GWASs I previously described.

In addition, other studies have focused on assessing genetic variants linked to candidate genes to understand how they influence progression. With respect to genes that influence motor decline, *GBA1* and *LRRK2* are not the only genes listed as potentially disease modifying targets. Stoker and colleagues accessed data from the CamPaIGN cohort (n=142) to explore concomitant genetic risk factors that could influence the progression of *GBA1*-PD. They found the rs356219 polymorphism at the *SNCA* locus significantly modulated the progression trajectory in *GBA1*-PD. Based on CPH models and Kaplan-Meier (KM) curves, they found that in particular the G/G genotype was associated with a worse cognitive decline. This effect occurred in *GBA1*-PD patients [118]. Another recent study explored the influence of several *SNCA* PD variants, independently of *GBA1*. They did find a minor effect of the known PD-risk variants rs356219 on motor progression as defined by UPDRS II score. In particular, they found that the G/G genotype was associated with a higher UPDRS II score. However, this association was not found in relation to UPDRS II rate of progression. They concluded that *SNCA* variants might have some effect on modifying disease progression but are not a major determinant of the PD clinical heterogeneity [119]. Rim and colleagues undertook a longitudinal study on 363 population-based incident PD cases diagnosed less than 3 years from baseline assessment to investigate the effect of *SNCA* on disease progression. They concluded that *SNCA* is a predictor of faster motor symptom decline in idiopathic PD based on their finding of a 4-fold increase in risk of carriers of the *SNCA*-Rep1 263 base pair repeat allele, a

promoter variant located in the SNCA-Rep1 microsatellite, which is among the most frequently investigated variants in *SNCA* [120]. Another study of 296 Chinese patients found that *SNCA* variants significantly contributed to the survival and severity of motor dysfunction [117]. Nevertheless, the impact of *SNCA* polymorphisms on disease progression is somewhat inconclusive as many studies have not been able to reach the sample conclusion [121,122].

Similarly, *APOE* is not the only genetic factor that has been associated with cognitive decline in PD. *COMT*, *BDNF*, *MTHFR*, and *SORL1* can also influence cognitive decline [123]. Another gene that has been implicated in cognitive decline is *MAPT*. Goris and colleagues found that development of PD dementia and cognitive decline were strongly associated with the inversion polymorphism containing *MAPT*. They also found a synergistic interaction between the *MAPT* inversion polymorphism and the single nucleotide polymorphism rs356219 from the 3′ region of *SNCA [124]*. Later on, Setó-Salvia and colleagues found that *MAPT* H1 was associated with PD and has a strong influence on the risk of dementia in PD patients [125]. Similarly to PD motor decline and the possible role of *SNCA*, there is controversy regarding *MAPT* impact on cognitive decline and progression to dementia, as many other well-powered studies were unable to replicate the findings [121,126].

2) Methods

a) Strategies and sources to capture disease progression

i) Longitudinal cohorts

In longitudinal studies, the first challenge is data collection. One of the most powerful data sources are **biobanks**, large repositories that contain biological data such as genotypic data, and store it associated with phenotypic data, so that it can be used in research [127]. When electronic health records (EHR) are available, multiple genetic research questions become feasible due to the large and deeply phenotyped samples available from biobanks [128]. In essence, biobanks enable the identification of loci and, subsequently, genes associated with various incident diseases, as well as those influencing drug efficacy or adverse reactions in an unbiased population sample. They

offer valuable insights into molecular targets, supporting the evidence-based creation of new drugs or diagnostic tools. Additionally, biobanks allow investigation of the interplay between genetics and treatment factors in disease progression, providing crucial medical information applicable to personalised medicine. An advantage worth highlighting relates to the removal of recall bias, in studying incident as opposed to prevalent cases and disadvantages include the lack of disease specific measures. A good example and widely used resource is the UK Biobank (UKBB), a prospective study gathering extensive genotypic and phenotypic data (including longitudinal follow-up) of over 500,000 participants aged 40-69 at recruitment [129]. Others include the Million Veteran Program, the BioBank Japan, All of Us Research Program, and FinnGen [130–133].

Cohort studies are another data source that enable us to assess how a condition evolves. Rothman and Greenland stated that in the field of epidemiology, 'cohort' is more often used to refer to those people that share a common experience or condition [134]. Cohorts are characterised as individuals from the general population with a shared attribute such as experiencing a specific health event. In such cases, the cohort design provides more information about health conditions measured by clinical assessments after the disease onset. This data assembly under cohorts enables the research community to investigate associations between multiple exposures and outcomes in a more specific way compared to a random sample [134].

There are two main types of cohort study. Prospective studies are those in which an exposure is assessed at baseline and study participants are followed up to record the development and progression of disease and mortality. Retrospective studies refer to those in which study participants are identified based on an inclusion criteria and exposures are assessed at baseline. Then outcomes of interest are studied during the historical observation period for those targeted samples [135]. Cohort studies can be used to determine the natural history and the prevalence of a condition. Normally, a study population free of disease / disease complication or an outcome is selected according to an exposure of interest and then followed up until the occurrence of the outcome of interest [136]. Cohort studies are particularly useful as they enable us to investigate single exposure, multiple outcomes associations, building up more insightful answers to hypotheses [137]. In addition, cohort studies are particularly appealing as opposed to case-control and cross-sectional study settings, in which

associations cannot efficiently separate causes and consequences [137]. Nevertheless, cohort studies also have their disadvantages. They suffer from selection bias. This normally occurs as the groups studied on a cohort should represent the underlying population and have the same exposures. In practice, this is usually not the case. In addition, losses to follow-up can cause difficulty, and high differences in follow up between recruitment groups on a cohort can bias results [137].

To study PD in particular, there are an increasing number of cohorts set up with different aims. Some PD cohorts aim to understand the natural history of PD. A good example of such cohorts is the Cambridgeshire Parkinson's Incidence from GP to Neurologist (CamPaIGN), a well-designed prospective study of PD evolution during 10 years on a population-representative incident cohort focusing on three milestones: postural instability (measured by HY 3), dementia and death [138]. A review from Heunzel and colleagues conducted in 2017 identified a total of 44 PD cohort studies with a published follow-up time of at least one year by using a PubMed search ("longitudinal" AND "Parkinson disease" AND "clinical"). All cohorts together made up a total of 14,666 participants, (cohorts' median: 138; range: 23–3.090), a median 1.5-year follow-up interval (0.5–4 years) and a median (planned) observational period of 5 years (1–20 years) were indicated. All 44 cohorts assessed motor functions, using UPDRS-III in 93% of studies. Similarly, cognitive function was measured in all cohorts identified [139].

Biobanks and observational studies are the primary data source for gene discovery in bioinformatics PD research. Target identification guides the time-consuming and dedicated phase of developing drugs modulating the disease related genes and ultimately the design of new **Randomised Clinical Trials (RCT)** for disease modification. For example, a Crohn' disease GWAS nominated the IL-12/IL-23 pathway to be associated with the development of disease [140]. This led to the design of clinical trials targeting that pathway [141].

RCTs are a type of prospective studies intended to measure the effectiveness and/or the safety of one or more interventions. An intervention such as treatment is allocated to two or more groups and the outcomes of interest are recorded so that comparisons can be made between the control and the treated groups. Each participant that is part of the RCT should have the same chance to be included in the intervention group

[142]. The randomisation incorporated in the clinical trial experiment is known to reduce bias when one wants to study cause-effect relationship thanks to the balancing of characteristics between groups in which a drug is being tested and compared against one another [143]. As a result, any differences observed between groups for the outcome of interest, can be attributed to the treatment. To further minimise bias, patient allocation in groups is anonymised (concealment), and participants, doctors, nurses and researchers are often blinded so that they do not know what treatment each participant is receiving. As with any experiment design, RCTs have their limitations, such as loss to follow up, time, high cost, and problems with generalisability (lack of representation of the intended population for which the RCTs was designed) [143].

In PD, clinical trials are intended to find drugs that can be proposed as either symptomatic treatments (ST; Improves or reduces symptoms of the condition) or disease modifying treatments (DMT; Delays or slows the progression of the condition by addressing the underlying biology of PD). In 1970 the U.S Federal Drug Agency (FDA) approved levodopa as the primary ST for PD [144]. In addition, monoamine oxidase type B (MAO-B) inhibitors, amantadine, apomorphine and dopamine agonists were tested before levodopa was approved [145–147]. However, all the currently approved drugs to treat PD target ST and none of them work directly in the underlying pathological biology, hence they have no impact in the progression of the disorder. However, thanks to the PD genetic research efforts, we have more knowledge of the possible biological pathways that either govern or influence the progressive neurodegeneration of Parkinson's. As a result, in the past years there has been a notable increase in clinical trials based on the understanding of potentially disease-relevant mechanisms of action [148].

During my PhD, I have mostly accessed PD patients data from Cohort studies and RCTs, based on data availability in the lab. Now, I will introduce in more detail 3 PD cohorts, 1 RCT, as well as 1 PD data source from a program called AMP-PD that has harmonised multiple PD cohorts and RCTs into one unified **biobank** [149].

Tracking Parkinson's [150]: The Tracking Parkinson's study is a multi-centre observational research initiative that enlisted patients from 72 centres throughout the UK. The recruitment criteria involve patients clinically diagnosed with PD who met the

UK Brain Bank diagnostic standards. Ethics approval was granted by the West of Scotland Research Ethics Service. The study adhered to the Declaration of Helsinki and is registered under NCT02881099 at ClinicalTrials.gov. This PD cohort study primarily recruited patients with recent onset, enrolling patients whose diagnosis was within 3.5 years. All participants underwent comprehensive clinical assessments recorded every 18 months, including motor, cognitive, and other non-motor evaluations. A second young onset PD group was established consisting of individuals diagnosed at the age of 50 or younger with a time from diagnosis exceeding 3.5 years. However, their assessments were conducted only at baseline, without longitudinal follow-up.

Oxford Discovery [151]: The Oxford Discovery study, officially known as the Oxford Parkinson's Disease Centre Discovery study, represents another observational multicentre investigation in the United Kingdom. Patients with PD were enlisted from neurology clinics located in the Thames Valley area. Eligible participants for the study were those who met the UK Brain Bank diagnostic criteria for PD and had received a diagnosis within the last three years. Ethical approval for the study was obtained from the Berkshire Regional Ethics Committee. Exclusions from participation were applied to individuals with non-idiopathic parkinsonism, dementia preceding PD by one year, or cognitive impairment hindering the acquisition of informed consent. Participants underwent standardised clinical assessments every 18 months.

Drug Interaction with Genes in Parkinson's Disease (DIGPD) [152]: DIGPD is a multi-centre longitudinal cohort study of PD patients. Patients were recruited based on the UK PD Society Brain Bank criteria that had a disease duration of less or equal than 5 years from disease duration at recruitment. Data was collected over 5 years by specialists in movement disorders. At every visit, specialists checked if patients met the UK PD BB criteria and filled out the standardised questionnaires. The cohort was approved by French regulatory authorities and an ethics committee, and conducted according to good clinical practices. All patients gave written informed consent (ClinicalTrials.gov NCT01564992). This study was set up to identify disease modifier genes as well as gene modifiers of treatment response and adverse events of parkinsonism drugs.

Simvastatin as a neuroprotective treatment for PD (PD-STAT) [153]: Carroll and colleagues set up a double-blind, randomised, placebo-controlled, multi-centre clinical trial to assess the possibility that statins might confer neuroprotection against PD. They did so as there is evidence from epidemiological and pre-clinical studies supporting the protective role of statins. In addition, simvastatin, a widely used cholesterol lowering drug with a well-established safety profile, has shown in various toxin and genetic cell culture and rodent PD models to influence several biological pathways that have been linked to PD such as neuroinflammation. Therefore, the PD-STAT experiment aims to define whether simvastatin could be used as neuroprotective therapy in PD. For that, they aimed to measure the futility of the drug in terms of prevention of the motor decline in PD patients, the validation the safety and tolerability of the drug in PD patients, the impact of simvastatin on activities of daily living and to distinguish symptomatic and disease modifying effects from simvastatin uptake [153]. The final results of the study showed that simvastatin was futile as a disease-modifying therapy in patients with PD of moderate severity [154].

The Accelerating Medicine Partnership in Parkinson's Disease (AMP PD) [149]: AMP program is a partnership between multiple biopharmaceutical and life sciences companies, the National Institute of Health (NIH), and non-profit organisations.

This initiative was set up to undertake a deep molecular characterisation and longitudinal clinical profiling of PD patient data and biosamples. The collection of such data is intended to enable researchers to identify and validate biomarkers of PD progression, prognostic and diagnostic. AMP-PD gathered data from well characterised cohorts with clinical data and biosamples available that were collected based on similar protocols and using common data elements. Among cohorts included in the latest release (release number 3) available when I last accessed AMP-PD data (20/01/2024), are: The MJFF and NINDS BioFIND study, Harvard Biomarkers Study (HBS), the NINDS Parkinson's Disease Biomarkers Program (PDBP), the LRRK2 cohort consortium (LCC), NIA International Lewy Body Dementia Genetics Consortium Genome Sequencing in Lewy body dementia case-control cohort (LBD), the study of Isradipine as disease modifying agent (STEADY-PD3), the study of Urate elevation in PD (SURE-PD3), and MJFF Parkinson's Progression Marker Initiative (PPMI). In addition, more recently, the Global Parkinson's Genetics Program (GP2) has joined the AMP-PD portal to provide a rich dataset [155].

PPMI is a multi-centre, international observational study [156]. PPMI is one of the most deeply phenotyped cohorts made publicly available to the community, with the primary objective of identifying and validating biomarkers that can aid in tracking the progression of Parkinson's disease. This initiative involves the recruitment of patients with Parkinson's disease at multiple centres across Europe, America, and Australia, adhering to the following selection criteria:

- Asymmetric resting tremor or asymmetric bradykinesia or two of bradykinesia, resting tremor, and rigidity
- Diagnosis within 2 years
- Hoehn and Yahr Stage I or II at baseline
- Untreated for PD, and not expected to require PD medication within 6 months at baseline
- Dopamine transporter (DAT) imaging showing DAT deficit
- 30 years or older at time of PD diagnosis

Throughout the study, participants underwent assessments every 3 months during the first year, followed by assessments every 6 months until the conclusion of the fifth year, and subsequently, assessments were conducted annually. Cognitive evaluations occurred exclusively during yearly visits. Motor assessments during annual visits were conducted in the "practically defined off" state, where participants refrained from taking PD medications since the night before the visit and for at least 12 hours prior. Since cognitive and "practically defined off" motor assessments were carried out annually, only data from annual visits were included in the analysis.

ii) Clinical instruments

The serial measures of clinical phenotype with questionnaires and structured clinical examinations are a reliable and accepted indicator of progression, and the baseline phenotype may predict future progression. For example, the extent of tremor in PD patients can be quantified using clinical scales such as MDS-UPDRS [157]. Several studies have reported an association between tremor dominant PD and a benign disease course [158–161]. Such indicators of progression are a valuable resource as there may be an association between the underlying pathophysiology and the

phenotype. In the study of Eggers and colleagues they linked the more benign disease course in tremor dominant patients with a less pronounced dopaminergic deficit [158].

Understanding the heterogeneity in PD progression is a primary aim of many observational cohort studies [150,156,162,163]. These longitudinal studies focus on gathering clinical assessments sensitive to changes. Measuring and modelling progression is central to developing effective disease modifying treatments and to understanding the underlying disease biology. Within the GP2 program [155] we have surveyed the use of Parkinson's assessments across global studies and selected the best established ones as we move towards one of our goals of creating the largest deeply phenotyped PD federated longitudinal cohort available. I summarise these clinical instruments in **Table 1**. They have been used widely across global clinics and are considered good definitions of different aspects of PD progression and severity.

Table 1. Summary of PD clinical measures and scales.

Clinical measure	Description
Hoehn and Yahr [164]	Motor progression based on symmetry, postural stability and gait
Rankin Scale [165]	Degree of disability or dependence in the daily activities of people who have suffered a stroke or other causes of neurological disability
MDS-UPDRS Part I [157]	Non motor experiences of daily living
MDS-UPDRS Part II [157]	Motor experiences of daily living
MDS-UPDRS Part III [157]	Motor signs
MDS-UPDRS Part IV [157]	PD motor complications
MoCA [166]	Cognitive assessment
SDM [167]	Shared Decision-Making assessment between a professional and a patient to evaluate psychometric speed
SCOPA-COG [168]	Scale sensitive to cognitive deficits in PD
RBD Screening Questionnaire [169]	Instrument to diagnose rapid eye movement sleep behaviour disorder (RBD) based on sleep behaviour measures
Epworth Sleepiness Scale [170]	Daytime sleepiness measure
Geriatric Depression Scale [171]	Clinical severity of depression among the elderly
Schwab England ADL [172]	ADL scale to determine dependence
UPSIT Olfactory test [173]	Sensibility of individuals to detector smells
PDQ-8 [174]	Quality of life in PD patients
King's PD pain scale [175]	Instrument to measure pain specifically in the PD population

Clinical measure	Description
MMSE [176]	Cognitive impairment measure
UPDRS Part I [177]	Mentation, behaviour and mood
UPDRS Part II [177]	Activities of daily living
UPDRS Part III [177]	Motor examination
UPDRS Part IV [177]	Complications of therapy
PD RFQ-U [178]	Instrument to measure the exposure to: caffeine, tobacco, alcohol, physical activity, head injury, residential and occupational histories, NSAID and hormonal medications, body habitus and pesticide exposure

Short description of all valid PD clinical scales summarised within the GP2 Cohort Integration Working Group.

Some of these scales are disease specific and some are used across conditions. Even though they may be considered good markers of different aspects of PD progression, there is no gold standard as to which test to use when studying clinical progression. Each assessment has its strengths and weaknesses to quantify progression.

When it comes to assessing PD motor progression, MDS-UPDRS III (PD motor examination) can be used to measure both response to levodopa treatment (symptomatic treatment) and the rate of change over time (progression) as exemplified by its widespread use in observational studies and RCTs. PPMI's original cohort consisted of *de novo* PD patients followed up during the course of 5 years, and the MDS-UPDRS scale was employed as one of the clinical assessments to measure motor and non-motor symptom severity in PD. A study using PPMI *de novo* PD patients was designed to characterise the progression pattern in untreated patients. They used the MDS-UPDRS scale to measure progression. They showed a linear increase of 2.4 (95% CI, 0.210–2.70 points per year) points per year in MDS-UPDRS part III total score (off medication), 0.92 points (95% CI, 0.80–1.05 points per year) for Part I (on medication), and 0.99 points (95% CI, 0.86–1.13 points per year) for Part II (on medication). Most of the changes in the MDS-UPDRS total score (a composite score made up by gathering each MDS-UPDRS subscale's score) were driven by the changes from MDS-UPDRS part III (estimated to account for 51% of MDS-UPDRS total score progression). Moreover, this study showed that the linear increase on the MDS UPDRS part III total score occurred in subgroups according to medication status. In the medicated group (patients that started taking dopamine medication at their 12-

months visit) a slower, linear increase was seen (1.8 points per year) as compared to a faster linear increase seen in the unmedicated group (progression of 4 points per year) [179].

In spite of the linear increase trend per year in the MDS-UPDRS scale reported on the PPMI *de novo* PD patients, whether the MDS-UPDRS scale is able to accurately capture disease progression during the early stages of the disease, is still uncertain. Regnault and colleagues used longitudinal MDS-UPDRS part II and III data in patients from the PPMI cohort, whose disease duration from diagnosis was ≤ 2 years, to investigate how well they explain the progression of early PD motor signs. They found both scales to have psychometric limitations which suggests a limited precision in measuring early motor signs. Limitation in the precision to measure early motor signs can decrease the sensitivity to detect differences in clinical change during the early stages of the disease. Particularly for MDS-UPDRS III, in spite of lacking single items to better measure changes in early PD, they found a clinically meaningful hierarchy explained by the scale, which makes it possible to be seen as a single metric across the severity continuum, hence MDS-UPDRS III could still be used as a basic measure to capture the basis of motor progression. On the other hand, MDS-UPDRS II was found to have more psychometric limitations, as the scale was not conceptually clear, even after an attempt of an scale reconceptualization, which may make the MDS-UPDRS II less appropriate for measuring motor symptoms in early PD as well as in studying disease severity across time [180].

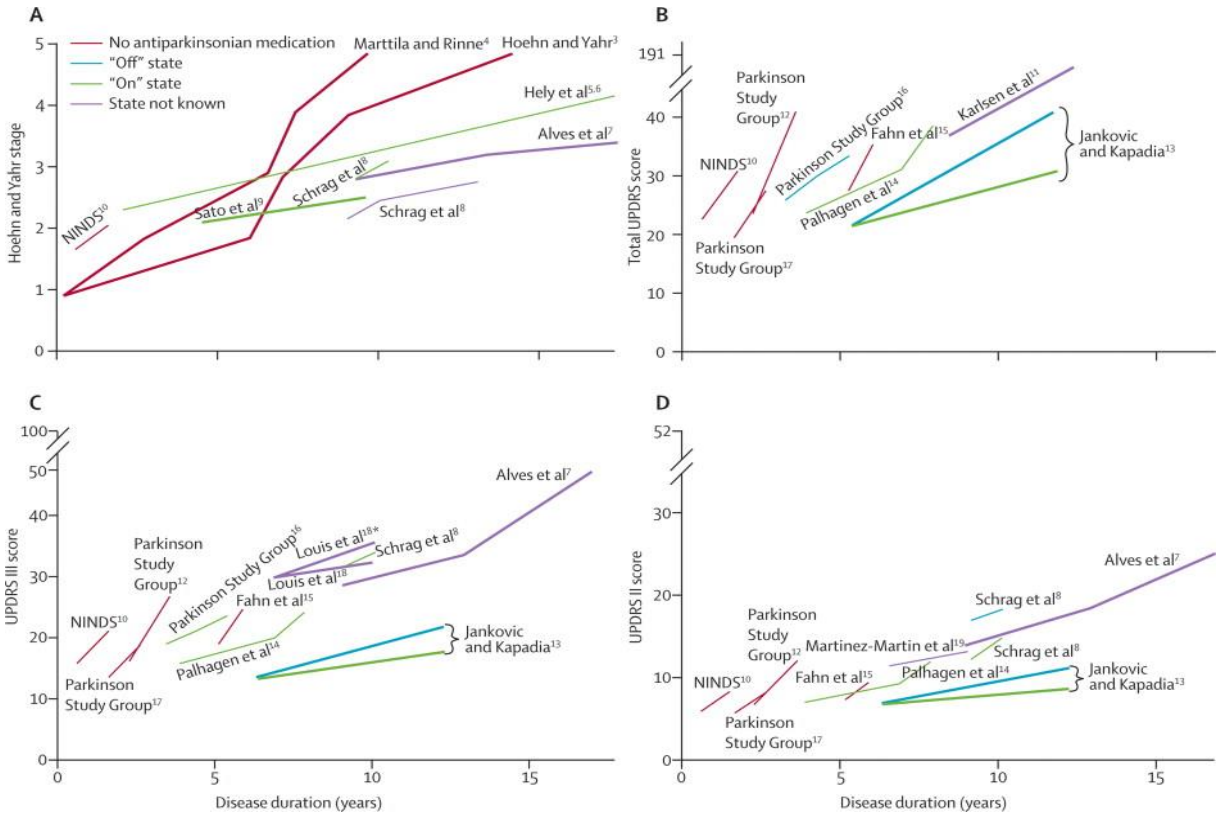
Evers and colleagues used a linear Gaussian state space model in a large observational cohort. With this type of statistical model, one can describe within-subject changes over time, and quantify estimates for the variance introduced by noise (measurement error and short-term effects), as well as the variance due to differences between individuals' progression results. The authors showed that the within-subject reliability, that is the rate of change per individual, across all parts of MDS-UPDRS was low. However, the within subjects reliability was favourable for parts II and III on the OFF state, as compared with part I and III on the ON state, with the scores measuring mobility, gait and posture, and rest tremor having the most consistent behaviour, outperforming scores quantifying bradykinesia, rigidity, and kinetic/postural tremor. Therefore, as previously suggested by authors of the MDS-UPDRS scale, analyses based on the subscales rather than on the composite MDS-UPDRS score

are preferred. In addition, because mobility and gait scores are most reliable in measuring individual subject changes, this knowledge could help to further split MDS-UPDRS subscales to more efficiently measure motor progression [181].

Although MDS-UPDRS part III is widely used in drug trials, other measures may outperform MDS-UPDRS part III or capture other important aspects of progression. Schrag and colleagues conducted a thorough comparison of clinical instruments in a community-based sample and a clinic-based sample, assessing the response to change over time of different scales measuring quality of life, disability, and impairment. This study showed that scales reflecting activity of daily living and functioning (SE, and UPDRS ADL part) were the most responsive to change over time, likely because overall function accounts for non-motor features as well as medication-driven motor components. Moreover, HY was slightly more responsive to change over 1 year than the UPDRS motor part III, for the clinic-based sample only [182].

Conventionally, clinical trials and observational studies have relied on face to face assessments at study centres at 1-6 monthly intervals. The advent of data collection from smartphones and wearable sensors could overcome widespread MDS-UPDRS scale limitations as they provide continuous and rater-independent measures of the patient's clinical state [181,183]. Despite the development of ambulatory devices to objectively measure progression, (MDS-)UPDRS, and HY, are still the most commonly used instruments to measure progression, and these scales are, to some extent, sensitive to change, with apparent differences according to disease duration (faster motor dysfunction in the first 5 years of the disease), and medication status (patients under PD medication show a decreased deterioration of motor signs) (**Figure 1**) [183].

***Figure 1**. Overview of motor progression across patients included in the placebo arms of treatment studies*.



Motor performance measured by (A) Hoehn and Yahr stage, (B) Total UPDRS Score, C) UPDRS III score, and (D) UPDRS II score. On the X-axis, disease duration in years. In red, studies without antiparkinsonian medication. In blue patients on antiparkinsonian medication with clinical measures on the OFF state. In green, patients on antiparkinsonian medication with clinical measures on the ON state. Thicker lines indicate more than 200 patients at study entry. The start of the line indicates the mean disease duration at the study inclusion. Figure reproduced from Maetzlet and colleagues [183].

When it comes to measuring global cognition in PD patients, MoCA and MMSE are widely used screening instruments. However, it is not clear whether such instruments are sensitive to changes over time. A study was designed to compare MMSE, MoCA, and SCOPA-Cog in PD patients without dementia at study enrolment across 6 North American movement disorders centres. Data for the instruments was collected yearly. They compared the assessments in terms of responsiveness over time, which can be understood as a way of quantifying the ability of an outcome to detect meaningful changes in a patient's health status over time. To measure the responsiveness over time, the authors used receiver operating characteristics (ROC) curves. They measured the area under the ROC curves (AUC) for MoCA (0.55 (95% CI 0.48–0.62)), MMSE (0.56 (0.48–0.63)) and SCOPA-Cog (0.63 (0.55–0.70)), with a larger AUC

meaning greater responsiveness to change. They did not find significant differences across the AUCs. Based on these results, they concluded that the sensitivity to detect decline in non-demented PD patients is poor. They hypothesise this might have an impact in clinical practice due to the lack of stable scores that capture cognitive decline [184].

Another study found that MMSE scores, as opposed to MoCA, declined significantly, which suggests that the MMSE may be more sensitive to cognitive decline [185]. In the OPDC cohort, Hu and colleagues found that MoCA was more sensitive in detecting changes than the MMSE, as they found that the MoCA declined significantly during study [186]. There is discrepancy between the studies that attempt to define a gold standard for measuring cognitive decline in PD. Other studies have also compared the sensitivity of the score to detect cognitive decline in a cohort of 102 Chinese idiopathic PD patients followed up for 30 months. Chen and colleagues found that cognitive performance significantly declined at 30 months as measured by both MoCA and MMSE. The annual decline was 0.82 for MMSE, and 1.02 for MoCA, suggesting that the MoCA scale might capture more cognitive decline compared to the MMSE scale [187].

Kim and colleagues investigated the capability of three clinical assessments, MoCA, DRS-2 and MMSE to predict disease progression on a group of nondemented PD patients with at least two clinical assessments over time. They found MoCA as the only outcome significantly associated with progression to PD with dementia (PDD) and faster time to dementia [188]. These results suggest that MoCA-based statistical and prediction modelling might be powerful to predict future progression to dementia and might be more powerful to capture dementia than the other two clinical instruments they compared (DRS-2 and MMSE).

iii) Statistical methods

There are several statistical methods that can be used to explore the impact of genetic variation on disease phenotype and progression. The goal is to be able to define the average impact of genetics on the rate of progression in a well-powered and bias free manner. In this section, I summarise the different statistical models I used in my thesis.

Linear regression

A linear regression is a type of statistical model which estimates the linear relationship between a quantitative outcome (dependent variable) and one or more explanatory variables (independent variables). We can distinguish between simple linear regression (only one independent variable) or multiple linear regression (more than one independent variable). Linear regression models make use of the linear function to estimate the model parameters for each explanatory variable (the so called regression coefficients, weights or “betas”) that can be used to predict the outcome of a dependent variable. To estimate the model parameters, linear regression models make use of “cost functions” on an optimization problem so the goal is to minimise such functions. The most widely used cost function in linear regression is the least squares approach, in which the goal is to minimise the sum of the squares of the residuals (the difference between the observed value and the value estimated by the model).

We can use these models to determine the amount of variation in the dependent variables attributed to the explanatory variables, as well as to determine the strength of their relationship. By using linear models in genetic studies, I can characterise the relationship between genetic variation and an outcome of interest such as a quantitative measure of PD severity based on a cross sectional clinical outcome or PD progression based on an average rate of change of a longitudinal assessment.

Logistic regression

Logistic regression is a powerful statistical modelling technique used to estimate the probability of a binary outcome based on one or more explanatory variables. At its core, this method employs the logit function—the natural logarithm of the odds—to transform probabilities from the bounded interval (0,1) to the entire real number line ($-\infty,+\infty$). This transformation allows for linear modelling of the relationship between predictors and the log-odds of the event of interest.

The model's foundation lies in the sigmoid curve, also known as the logistic function, which maps real-valued inputs to probabilities. The inverse of this function, the logit, serves as the link function in logistic regression, enabling the estimation of log-odds.

In binary logistic regression, the response variable is dichotomous, typically coded as 0 or 1. The model estimates the log-odds of the event (coded as 1) occurring, given a set of predictor variables that may be continuous, categorical, or a mixture of both. The relationship between these predictors and the log-odds is assumed to be linear.

Parameter estimation in logistic regression is commonly achieved through maximum likelihood estimation. This method identifies the values of the model parameters that maximise the likelihood of observing the given data under the assumed model.

In my research, I applied logistic regression to both traditional case-control studies and within-case genetic analyses. These applications allowed me to quantify the association between genetic variants and disease occurrence by estimating the log-odds of disease presence conditional on genetic markers. This approach provides valuable insights into the genetic basis of diseases and can inform risk prediction and personalised medicine strategies.

Generalised linear model (GLM)

GLM is a generalisation of linear regression. GLM can model dependent variables with any type of distribution (as opposed to simply normal distributions required in linear regression). They make use of a link function of the response variable which varies linearly with the predictors. Likewise GLMs unify various models such as linear regression and logistic regression. In case control studies, in which there is a binary response (disease or healthy condition), or in other words, a Bernoulli variable, a linear regression model is not suitable as probabilities are bounded on both ends. The log-odds function serves as a link function between the probability and the linear regression expression. This is because the log-odds function ranges between $(-\infty,+\infty)$, as I said earlier, so that linear regression can be applied, and once coefficients are estimated through linear regression, the log-odds can be easily converted back into probabilities.

Linear mixed effect model (LMM)

LMM is an extension of the ordinary least-squares (OLS) regression. LMM is used to incorporate hierarchical data such as serial measures. In longitudinal data, the OLS model assumption of observations sampled independently and randomly from the population is not met, as patient level observations are sampled from the same group

repeatedly, so there is non-independency in the data within groups. LMMs mitigate this assumption by taking into account the correlated nature of observations within groups. There are two sources of variance within hierarchical data, that is within groups (i.e. individual level serial observations), or between groups (i.e. patients with or without a candidate predictor variable, in this case a single nucleotide variant across patients).

LMMs incorporate fixed and random effects. Fixed effects match those from OLS or multiple linear regression, as it is a parameter associated with each covariate that is non-random and considered to be constant for the population being studied. Fixed effects are consistent at the group level (i.e. individuals). An example is the overall effect of the SNP under investigation that is consistent across individuals. This parameter is an estimation of the true coefficient in the population based on our data. Random effects are parameters that account for unexplained sources of variance (i.e. differences between individuals). Random variability can be included at two levels, the intercept and the slope. With a random intercept we allow for differences in the intercept between the population average and each individual intercept. With a random slope, we allow for differences between the population average slope and the individual slope. Therefore, in LMMs, the parameters are no longer fixed, but have a variation around their average values, and this usually provides a better fit and explains more variation than strategies based on OLS.

LMMs are able to account for unbalanced data, allocate individuals with incomplete records (individuals missing any time point during study duration), and are more informative as they capture the heterogeneity of complex traits over time, resulting in an increase of power to detect significant associations and reduction of false positive rates, as opposed to aggregated strategies.

Cox proportional hazard model (CPH)

Survival is a term used to refer to the time from a start point to the occurrence of an event (i.e, death, progression to a clinical milestone). Therefore, survival analysis refers to those statistical strategies to investigate the time for the occurrence of an event. When we want to use observational studies to perform a survival analysis on an event of interest, right-censoring might occur (study finishes and a patient has not experienced the event yet; a patient is lost to follow up). Right-censoring is efficiently

handled on survival analysis. Based on the observed survival times from observation studies, the survival probability can be estimated using the Kaplan-Meier nonparametric method. CPHs is a type of survival model to measure the association between the survival probability and multiple factors that may influence the survival time. The interpretation of the outcome-factor relationship is based on the hazard rate, which is the event rate at a time point (t) conditioning on surviving at least until that time point (t). A covariate coefficient greater than zero equals a hazard ratio greater than one, which indicates that the value of that covariate increases, the event hazard increases, therefore the survival time decreases as well.

Cox models are widely used in medical research to test treatment assignment, so that one can test the hazards of taking a medication against being untreated to assess the effectiveness of a drug on an outcome such as disease progression. Cox models are also used in genetic association studies to assess the impact of patients carrying a certain genotype, and the impact on the survival time while adjusting for confounding variables such as age (if we were assessing time to mortality, and one of the groups were older, that group would be more likely to die earlier due to the unaccounted effect of ageing on mortality).

There is one assumption for Cox models to generate unbiased effect estimates, and is that the hazards must be proportional during the study length. If the hazards of a given genotype on the subject are not constant over time, then conclusions about the survival time and outcome relationship through Cox models would not be valid and the model should be rejected. To check if this assumption is met, there are several statistical tests and graphical diagnostics that can be used. In my Thesis, I normally plotted the Kaplan-Meyer curves that enable to visually inspect the proportional hazards hold true. I also used the scaled Schoenfeld residuals (time-independent residuals) to correlate them with time and test for residuals-time independency.

iv) Algorithms

The standard approach to conduct genetic association studies is to apply the statistical models previously described at a genome-wide scale, which is testing genetic variants genome-wide against an outcome of interest. I can apply the same idea to explore

how a specific phenotype progresses over time, so use repeated measures during a study length and evaluate the impact of genetics on those progression trends.

Even though LMM makes possible the modelling of hierarchical data, they become computationally expensive when performed at a genome-wide scale, that is when performing ~6,000,000 independent tests. Recent studies have focused on finding more efficient approaches around LMM to shorten the compute time, making such type of large scale analysis possible. In 2012, Sikorska and colleagues explored several methods to decrease the compute time of longitudinal GWAS while producing accurate estimates. They compared the methods and found a conditional two step approach was the best performing method. This method was based on the idea of conditional inference. They estimated the longitudinal effects on the baseline characteristics omitting SNP information. In the case the reduced model is mis-specified by the effect of SNP cross-sectionally or longitudinally, the subject-specific slopes would contain information about the evolution of the outcome of interest for the different SNP alleles. Therefore, on a second step the best linear unbiased predictors of the subject-specific slope can be regressed on the SNP using a simple OLS [189]. Ning and colleagues developed the GMA method, which is composed of GMA-fixed based on a fixed regression strategy with eigenvalue decomposition, and GMA-trans, which applies a linear transformation of genomic estimation values for unbalanced (individuals may be recorded at different time points) and balanced (all individuals are measured at the same time points) longitudinal traits [190]. TrajGWAS is a method that scales linearly with the number of individuals. It allows us to assess the contribution of genetics to the mean level of biomarker trajectories or their fluctuations (or individual variability or within-subject variability), which are both a form of longitudinal trajectories [191]. Another approach, HiGwas, is based on a function-values approach to select significant SNPs based on lasso penalty and estimate their time-varying genetic effects that follow biologically interpretable functions [192].

There are two other novel algorithms that I have incorporated in my genetic association studies due to their very efficient reduction in computational time as well as the very accurate approximations to LMMs.

Simultaneous correction for empirical Bayesian estimates (SCEBE)

SCEBE is an algorithm adapted to explore genome-wide associations with longitudinal outcomes through mixed-effect modelling. With SCEBE, we fit a base mixed-effects model and used the predictors of random effects from the base model as phenotypes for GWAS through a linear regression model. Because the predictors of random effects are affected by shrinkage to population mean, as they are the weighted sum of the population and sample mean, using them as phenotypes would lead to biased estimations of the SNP effect estimated P-values. Yuan and colleagues quantified the bias in SCEBE and added it as a correction matrix, allowing us to generate unbiased SNPs metrics [193].

Genome-wide Analysis of Large-scale Longitudinal Outcomes using Penalization (GALLOP)

GALLOP is a high speed algorithm that enables the estimation of the cross-sectional and longitudinal SNP effects and the P-value of the test-statistic. GALLOP relies on the small SNP effects on outcomes on GWAS settings. We can estimate the variances in a base LMM and make the assumption of the model variances not changing after adding a given SNP due their very small effect sizes (the proportion of variance explained by a SNP on a LMM will be very small). Using the equivalence between a mixed model and penalised least squares, a system of many linear equations is set up and the result is a very sparse system with only the last rows and columns changing from SNP to SNP, which will result in a low memory use. This approach decreases the computational time by three orders of magnitude compared to the use of pure LMMs at the genome-wide scale [194].

v) Genetic approaches to disease progression

Here, I summarise some of the different approaches that research groups have used to investigate the impact of genetic variants on longitudinal traits.

Gorski and colleagues made use of multiple cohorts, including UK Biobank, to define for each individual the decline of estimated glomerular filtration rate (eGFR), which can progress to overt kidney failure. They used the annual eGFR decline to define genetic variants significantly associated with the annual decline [195]. Whereas these approaches are accurate and they are efficient surrogates of progression, they might

be underpowered as opposed to strategies that make use of all the repeated measures as continuous outcomes [196].

A genetic association study of early childhood growth made use of longitudinal growth traits from multiple cohorts based on a two-step approach. They used LMMs to derive sex-specific individual postnatal growth velocity and BMI curves in children from data collected from primary health care or clinical research visits. Then, they performed GWAS on six harmonised early growth traits and found four variants at four independent loci associated with three early growth traits, one of them, a newly discovered variant at the *LEPR/LEPROT* locus [197]. Adkins and colleagues investigated the common genetic variants predicting developmental trajectories of alcohol consumption in three longitudinal community samples. They used a two-step approach to first compute a subject-specific alcohol consumption trajectory adjusted on age, and then regressed on additive SNP effects based on linear regression [198]. Tan and colleagues carried out a GWAS on the rate of change in forced expiratory volume in the first second (FEV1) across 14 longitudinal, population-based cohort studies. The study encompassed 27,249 adults of European ancestry and employed a linear mixed-effects model for the analysis. They identified two novel genetic loci in association with the rate of change in FEV1 that harbour candidate genes related to lung function [199]. Allen and colleagues performed GWASs using LMMs with random slope and intercept with an (Time x SNP) interaction term, to identify genetic variants associated with declining lung capacity or declining gas transfer after diagnosis of IPF [200]. Smith and colleagues used Cardiovascular disease (CVD) risk factors recorded from childhood from the Bogalusa Heart Study, a longitudinal study focused on the early natural history of CVD. They used LMMs to estimate e. SNP and SNP x AGE interaction effects separately. They found genetic variants associated with CVD risk factors in a time-dependent (SNP x time effects on risk factors) and time-independent (SNP only effects on risk factors) fashion [201]. All these studies showed good power to investigate genetic association with longitudinal outcomes. Previous research found that efficient two-step approaches provide unbiased test-statistics and effect sizes of SNPs as opposed to regressing longitudinal traits on SNPs on LMMs genome-wide [193].

The algorithms we described in the previous section to investigate the genetic impact on longitudinal traits in large scale analyses are largely underused. He and colleagues

made use of the previously described tool, TrajGWAS to assess the influence of SNPs on the bone mineral density (BMD) trajectory mean as well as on the within-subject variability of BMD. They used data from 141,261 white participants from the UK Biobank with heel BMD phenotype data [202]. Yang and colleagues also investigated if during pregnancy and the postpartum period, genetic variants were associated with the mean and variance of platelet counts [203]. Benchmarking on these proposed methods to further prove the accurate approximation to LMMs and as well as to nominate the better performing algorithm is still needed.

b) GWAS concepts and methods

i) Genotyping and whole genome sequencing

Whole genome-sequencing (WGS)

WGS is a process through which the entire DNA sequence of an organism from both chromosomal and mitochondrial DNA is determined, although in practice WGS coverage ranges between 90-95%. During the last years, WGS has become more accessible thanks to the advent of new technologies such as next generation sequencing which entails improvements in massively parallel analysis, high throughput, and reduced costs.

Genotyping

Genotyping is a method to characterise the individual's DNA at certain genomic positions. Genotyping is distinct from DNA sequencing, which is a method to determine all nucleotides on a specific DNA fragment.

Microarrays are used to genotype thousands of different informative loci at a time, thanks to the ability to deposit different DNA sequences on a small surface, normally a glass slide. The microarray principle is based on complementary sequences binding to each other. Oligonucleotides with certain DNA combinations (probes) bind to the DNA of interest to detect sequence variants [204]. Therefore, when a sample complementary DNA is washed in the microarray, fragments of the molecule hybridise to a probe and the scanning software, called genotype calling, determine the genotype found on each probe [205].

Microarrays have evolved to include both common genetic variation and disease specific variants. This is possible thanks to the knowledge of the heritable component of multiple diseases. The NeuroBooster array (NBA) is a good example this type of array. It is designed to boost the genetic coverage of loci linked to neurological diseases. NBA contains a backbone of 1,914,934 genetic markers from the Infinium Global Diversity Array-8 v1.0, complemented with custom content of 95,273 disease-associated variants involved in a wide range of neurological conditions [206].

Genotyping data is commonly used in clinical and experimental studies. Its most extended application is in GWAS, in which all genetic variants across the entire genome are assessed for association with traits or diseases.

ii) Linkage disequilibrium (LD)

LD is the term used in population genetics to refer to the non-random association of alleles at two or more loci [207]. LD patterns are of importance in evolutionary biology as it provides clues about past events. Throughout the genome, LD reflects the breeding system, the population history and geographical subdivisions. At each specific genome region, LD reflects natural selection, gene conversion, mutation and other factors that influence the evolution of gene-frequency [208]. As an example, in some genomic regions, LD patterns correlate with recombination hotspots, so LD can be seen as a function of crossover distribution [209]. LD patterns are not constant, and vary across genomic regions as a result of stochastic factors such as different gene history across loci [210], and across populations [211].

For a pair of loci, the coefficient of LD is defined as the difference of the frequency of gametes carrying a pair of alleles at two loci (A and B) (Pab) and the product of the frequency of those two alleles (Pa x Pb). Linkage Equilibrium occurs when this difference is equal to 0. For more than two loci, pairs of loci are normally grouped in the so called haplotype blocks, which are non-overlapping loci in strong LD [208], separated by regions of recombination events [209,212]. This suggests an hypothetical division of the genome into regions of high LD separated by narrow recombination hotspots [213]. The HapMap project confirmed the generality of recombination hotspots in the genome, the large lengths of segments in high LD, and the low haplotype diversity [211].

The block-like structure of the genome was a revolutionary discovery as they were of practical use in case-control association studies, enabling the use of one SNP in each block as an approximation of association of all the SNPs on that haplotype block [214]. The applications that arise from LD knowledge include mutation and gene mapping, detecting natural selection, and estimating allele age [208]. The LD structure of the genome has been used to develop widely used statistical frameworks to correct genomic inflation in genetic association studies [215]. LD can also be used for imputation [216].

ii) Imputation

Imputation is a process used to infer missing genotypes from genotyping data. Imputation techniques rely on reference panels of tens of thousands of complete genomes from common ancestors and the LD structure of the genome. Likewise, reference and target genotyped samples can be matched to identify the shared patterns in DNA sequence, and the missing genotypes within the shared haplotypes can be inferred. To accurately impute missing DNA sequences from SNP array data, phasing is necessary since genotyping data is unordered [217]. Phasing is the process of deducing haplotypes by separating or 'phasing' maternally and paternally derived sequence information [218]. Imputation is usually performed as it increases the power of genetic association studies by increasing the number of variants that are available for hypothesis testing.

One of the major steps forward in imputation has been the efforts driven by large projects such as HapMaP, 1000 Genomes, UK10K, and the Haplotype Reference Consortium projects [219–222]. These large initiatives have been redefining and improving methods for the characterization of DNA genome-wide across several samples, reporting allele frequencies, types of DNA differences, as well as estimating the correlatory structure of the genome, which is possible due to the inherent LD structure of the genome, hence defining confident haplotype blocks. Based on the knowledge that such reference panels provide us with, microarrays can lead to yet confident and complete genomes in a very cost effective manner. The largest reference panel to date is the Trans-Omics for Precision Medicine (TOPMed), gathering 400,000,000 single-nucleotide and insertion or deletion variants across 130,000 samples at date of publication. Release 3 of the panel (the most up to date

and currently available to the scientific community), includes 133,597 reference samples and 445,600,184 genetic variants distributed across the 22 autosomes and the X chromosome.

At the same time that massive reference panels are generated, web servers for genetic data users are also available. This is a key point as increasing reference panel size also increases the computations cost of imputation, which would prevent every day users accessing and using them. The TopMed Imputation server, and the Michigan Imputation servers are the most powerful and widely used by the research community [223,224].

Similarly, this increase in computational cost has motivated the research community to develop optimised computational methods in multiple ways. By 25-03-2024, the most used imputation algorithms include IMPUTE2 [225], Minimac [226] and Beagle [227,228]. Eagle [229,230], SHAPEIT2 [231], and Beagle 5.1 can be used for imputation [232].

As part of this PhD, I individually imputed all the genotyped cohorts in the Michigan Imputation Server (MIS) [226]. In order to prepare data for imputation in the MIS specifically, I ran the Will Rayner tool for further quality checks according to the HRC Panel [233]. Prior to imputation, I updated strand, position, and reference / alternate allele assignment, as well removing A/T and G/C SNPs if MAF> 0.4, SNPs with > 0.2 allele frequency difference, and SNPs not present in the HRC Panel [234]. Then, I imputed it in the MIS, using Minimac4 [235] as the genotype imputation software, HRC as the Reference Panel for imputation, and Eagle v2.4 [236]. Similarly, I used TopMed Imputation Server for imputation of some cohorts.

iii) Quality control steps

Before and after data imputations, quality control is performed. Quality control is done at different stages and levels:

Sample level QC: At the patient level, I removed samples with low genotyping rates (<98%), sex mismatch between reported sex and the genotype derived sex, heterozygosity outliers (I considered samples as heterozygosity outliers if they deviated more than ±3 standard deviations (SD) away from the mean cohort

heterozygosity rate). To remove one of paired related individuals, using GCTA software (version 1.93.0 beta for Linux) [237], I created a genomic relationship matrix from pruned data between pairs of individuals, and I removed one of a pair of individuals with estimated relatedness larger than 0.125, equivalent to second degree relations. To deal with population stratification, I performed a principal component analysis (PCA) over pruned genotype data of each independent cohort merged with Utah residents with Northern and Western European ancestry (CEU), Han Chinese in Beijing, China (CHB), Japanese in Tokyo, Japan (JPT), and Yoruba in Ibadan, Nigeria (YRI) populations from the HapMap reference panel to identify non-European ancestry sample [238]. At first, I visualised each cohort with CEU, CHB, JPT, and YRI populations, so as to make a decision on the threshold of SD away from any of the mean 10 first PCs from the CEU population to consider non-European ancestry samples. Finally, a second filter was applied to further remove heterozygosity outliers, as well as samples with low genotyping rate (<95%) based on recalculated relatedness and missingness frequencies on the remaining samples.

Variant level QC: At a genotype level, I removed variants that had a missing rate higher than 0.05, variants with a minor allele frequency (MAF) of less than 0.01 or 0.05, and variants whose missing calls were not randomly distributed by testing whether missingness status could be predicted from genotype calls at the two adjacent variants. Moreover, I excluded variants with extreme Hardy-Weinberg equilibrium (HWE) deviations as they are indicative of sample contamination. (P Value<1e-10) [239].

Post-imputation QC: To only work with variants that were imputed with high confidence, I removed those with an estimated value of the squared correlation between imputed genotypes and true, unobserved genotypes (Rsquared or Rsq) < 0.8. Furthermore, I excluded variants with low genotyping rate (<95%), and MAF < 0.01, resulting in over 500000 SNPs available across cohorts.

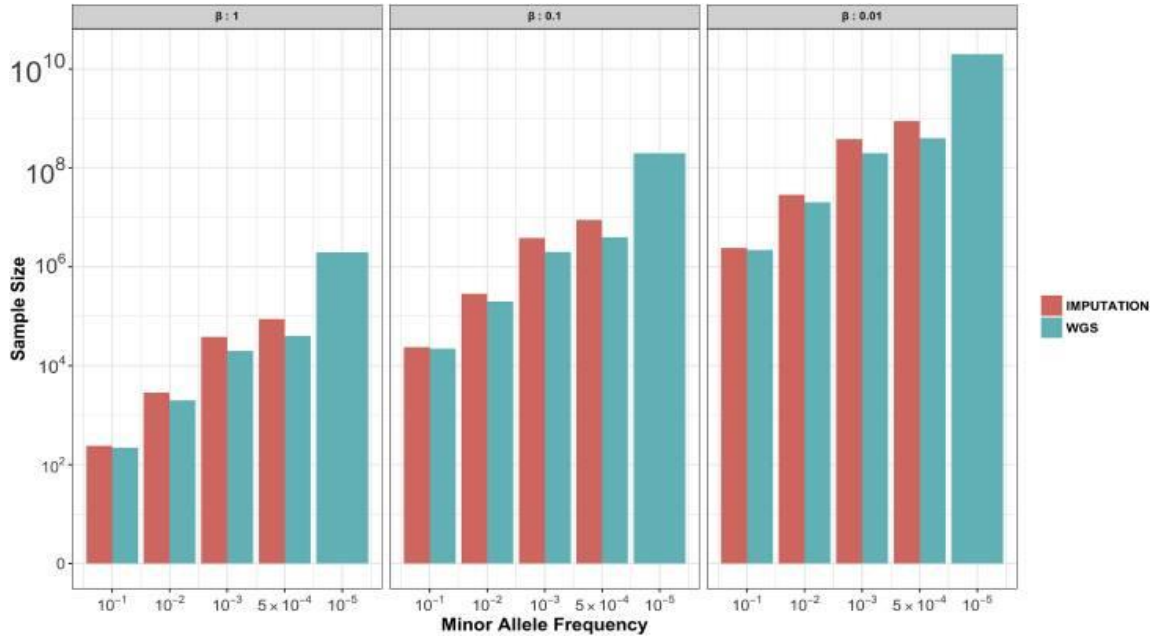
Post-meta-analysis QC: Once the meta-analysis was complete, I removed variants with MAF variability between cohorts higher than 15%, and also those variants showing high between-study heterogeneity according to the Cochran's Q-test (P < 0.05) and I^2 index (variants with an heterogeneity higher than 80%).

iv) GWAS

GWAS is the core approach in PD genetics research and in the general population genetics field, as it allows us to freely scan the genome in search of associations between genetic variants and disease with disease (in the case-control studies) or the severity of a phenotype and its progression. In GWAS, we can assess genetic variants under three different assumptions that fit any statistical model. Normally, GWAS focus on SNPs, usually those with MAF greater than 1% or 5%. A variant can be studied as being additive, that is, the effect of the polymorphism is cumulative. Therefore, the effect on the phenotype or trait will be higher if a variant is present as homozygous than heterozygous. Genetic variants can also be studied under dominant or recessive models. The dominant model assumes that a mutation in one allele is sufficient for the development of the phenotype and there is no cumulative effect. On the other hand, creating a recessive model assumes that a SNP must be present in the two alleles to have an effect on the phenotype.

The power of GWAS to detect significant associations depends on the sample size, the frequency of those variants associated with traits, the effect size of those variants, the heterogeneity of the trait studied, as well as the LD structure. As the sample size of the population studied, the effect of a disease-associated variant, and the allele frequency increase, the power of GWAS increases. Interestingly, it is worth noting that there is not much difference in power when doing GWAS based on imputed data from SNP arrays or WGS except for ultra-rare variants (frequency less than 1e-5) (**Figure 2**) [240]. For the estimation of the imputation accuracy, Vissner and colleagues used the average imputation Rsquared values reported by the HRC study in **Supplementary Figure 3** [222].

***Figure 2**. Sample Size required to detect association from Imputed (red) and WGS (blue) data.*



On the Y axis, sample size, and on the X axis minor allele frequency. Each plot represents the minimum sample size to detect an association of a SNP with a certain MAF, for different effect sizes expressed in phenotypic standard deviation units. Figure reproduced from Visscher and colleagues [240].

Therefore, increasing the experimental sample size will lead to new insights into the segregating variants significantly associated with complex PD phenotypes and derived traits such as progression, superficially for those whose effect and AF is lower as they can only be uncovered with larger sample sizes.

In a simple GWAS setting, I will test the same hypothesis for all SNPs that remain after all QC procedures. Taking into account the LD structure of the human genome this is thought to be equivalent to a million independent tests. Correction for multiple testing is needed in order to decrease the False Positives Rate. Currently, a genome-wide significance P-value threshold of 5e-8 is standard to report true common (MAF ≥ 5%) genetic associations, a stringent threshold adopted to avoid type I errors.

A crucial step after performing a GWAS is to check that there is no genomic inflation, suggesting population stratification and a systematic underlying difference between cases and controls. With Quantile-Quantile plots (QQ-plots) one can visualise the observed P-values from the GWAS against the expected P-values from a theoretical $\chi 2$-distribution. A diagonal line indicates the overlap between the observed and expected P-values so that an early departure from the diagonal indicates inflation,

normally attributed to population structure [241]. Lastly, Manhattan plots are used to graphically represent the results of all statistical tests conducted. The plot is named after the skyline of Manhattan due its resemblance to the vertical arrangements of buildings on the island. On a Manhattan plot, the X-axis represents the genomic position of genetic markers along the chromosomes, and the Y-axis represents the significance of association between each genetic variant and the trait or disease being studied. This is often measured as the negative logarithm of the p-value obtained from statistical tests. Therefore, in a Manhattan plot, each SNP is represented by a point on the graph.

v) Meta-analysis

The power of genetic association studies is dependent on sample size, hence we are limited by the sample size of each cohort. Meta-analysis is a statistical approach that leverages results from different studies increasing overall study sample size, and, as a result, decreases the standard error (SE) of the effects of variants on outcomes, as there is a closer representation of the more general PD population. Meta-analysis provides us with more reliable results for the association effect. If there are cohort-specific false positive associations (type I errors), these will not be significant in the meta-analysis as the results will not be supported by the other cohorts. Conversely, as a result of increasing sample size, small effects remaining undetectable (type II error) in small cohorts, may be picked up from meta-analyses [242]. Results derived from meta-analysis are statistically as efficient as joint participant data analysis [243]. Therefore, meta-analysis improves joint cohort analyses and reduces the use of resources [243].

There are two types of meta-analysis models, fixed effects, which makes the consideration that genetic factors have similar effects on the outcome between cohorts and that the observed variation happens by chance, and random effects, which considers that there is diversity among studies relating to true underlying allelic-effect heterogeneity [242]. To conduct meta-analyses I have used METAL software (version released on the 2011-03-25) [244]. METAL enables meta-analysis based on two different approaches. The first approach converts the effect directionality and P-value of a model variable into a signed Z-score across studies. This Z-score is combined across studies in a weighted sum, with the weight being proportional to the square-

root of the sample size for each study. A second approach is based on weighting the effect sizes of the model variable per study by their SE. The way I performed meta-analyses was based on a fixed-effects model weighted by β coefficients and the inverse of the SE [244,245]. I chose a meta-analysis over a merged analysis because of the heterogeneity in the inclusion and exclusion criteria across the clinical cohorts, and the differences in the genotyping approaches, as well as the statistical equivalence [243].

To correct summary statistics for any population stratification or cryptic relatedness bias, I applied genomic control correction to the cohort-specific summary statistics by computing the inflation of the test statistic, and then applying a correction to the SE.

vi) Polygenic risk score (PRS)

Variants nominated in GWAS tend to have a small effect on the phenotype of interest, and when they are assessed separately, their predictive capability is very limited. However, when we combine the effects from all the independent variants associated with an phenotype or trait, then this aggregated measure captures much of the heritability of the trait [246]. PRS is a method that has been developed to capture the aggregated effect of all those genome-wide variants associated with traits to increase the power and accuracy to predict phenotypes based on genetic variability alone [247,248]. More technically, PRS is calculated at the individual level as a sum of all the genotypes as genetic markers for a trait, genome-wide. Normally, genotypes are common (MAF > 0.01), biallelic SNPs, based on GWAS design. Those target genotypes are then weighted by their effect size inferred from GWAS results [248].

I used PRSice software (version 2) to compute PRS. I set a threshold of $P < 1e\text{-}6$ to include all independent nominal significant GWAS variants that make up the PRS [249]. I selected independent SNPs by clumping within ±250 Kb from the index SNPs (the most significant SNP on a genomic window). I used the SNP betas as the estimate to compute the PRS. Sex, standardised AAO, and the first 5 PCs were added as covariates to the PRS estimation process. To compute the LD estimates, I used the imputed cohorts from which I calculated the PRS, as they were large enough to provide accurate LD estimates (N > 500). To validate the PRS as an instrument to distinguish between PD patients with and without LiD, I derived time-dependent ROC curves,

under the assumption that different PRS loads might cause changes to time-to-LiD onset. I used the Inverse Probability of Censoring Weighting (IPCW) estimation of Cumulative/Dynamic time-dependent ROC curve from the 'timeROC' R package (version 0.4). To compute the weights, I used the Kaplan-Meier estimator of the censoring distribution.

c) Functional annotation for decoding GWAS

Although many GWASs have been conducted revealing novel associations with PD risk and traits, the interpretation of nominated variants remains challenging. In sporadic PD, as we depart from clear Mendelian inheritance patterns, the interpretation of GWAS is confounded by the LD structure of the genome and is limited to our functional understanding of the genome and available assays. In such "complex traits", there are many variants, hence genes, involved in disease, which can interact in an additive or non-additive way [250]. This scenario complicates the understanding of the underlying biological mechanism.

As stated by Francis Crick in the central dogma of molecular biology, genes are transcribed into messenger RNA, and then translated to protein [251]. Therefore, changes at the DNA level, that is mutations in protein coding genes, or mutations in non-coding regions, could result in the translation of aberrant proteins due to a change in the protein sequence, or a dysregulation in expression altering the RNA levels or splicing.

The majority of the genome is made of non-coding regions, where many regulatory elements exist and mediate the transcription of many genes, and this regulation can happen in *cis* (regulatory regions close to genes up to 1Mb) and *trans* (distal regulations that can happen even between loci in different chromosomes). Regulatory elements in the non-coding genome are enriched for disease-associated variants [252,253]. In addition, chromatin accessibility varies across the genome and cell types and there are loci transcriptionally more active than others [254], which suggests the non-coding genome activity is complex and genetic variation in the regulatory elements are linked with disease in a cell type specific manner.

Towards gaining novel insight into how regulatory elements control cell type specific gene regulation, large consortium studies have successfully provided maps of functional and regulatory elements [255,256]. Novel approaches have provided updated predictions of regulatory maps that link enhancers to genes based on extensive epigenetic assays at multiple tissues and cell types [257,258]. In addition, eQTL are derived from the combination of RNA-seq studies and genotyping or whole-genome sequencing studies. They provide insights into how loci are associated with gene expression and which genes are regulated by loci across the genome in humans [259,260]. eQTLs are enriched for trait association [261]. Nevertheless, none of these approaches are able to nominate causal variants, and therefore, statistical fine-mapping is needed to understand correlated structures due to LD.

In order to shed light into the underlying functional alterations linked to different PD traits, I list here the approaches I have recurrently used throughout the PhD to decode GWAS.

i) Fine-mapping

It is often the situation in which we nominate an LD block from a GWAS. In order to statistically support inferences made about the potential SNPs nominated to cause a specific trait, fine mapping tools are the gold standard approach to decode LD blocks by finding the genetic variant or variants responsible for complex traits [262]. Fine-mapping has been validated to confidently infer the causal variant or variants from GWAS summary statistics [263].

The three main statistical approaches to perform fine-mapping are based on heuristic methods, penalised regression models, and Bayesian methods. At a genomic level these approaches can be based solely on the LD structure and association statistics, or can incorporate functional data such data derived from expression and splicing analysis. These methods make use of different parameters: number of causal SNPs in a region as well as their effect size, the LD structure around the lead SNP, sample size, the SNP density and whether the causal variant can be measured or not. Bayesian methods are the most widely used, since simulation studies have shown fine mapping Bayesian methods to perform best [262]. Bayesian methods create 2^m models with SNPs as discrete variables, as causal (1) or non-causal (0). Based on this

posterior inclusion probability (PIP) indicating the probability for each SNP being causal, are computed. Moreover, a credible set (CS) can be derived through a parameter (α). This CS is just a cumulative PIP containing the minimum set of SNPs reaching the probability α. Moreover, some tools make use of a variety of functional annotations to increase the resolution of PIPs [262].

ii) GCTA-COJO

When interpreting a locus from GWAS results, usually the SNP showing the most significant statistical evidence for association is considered as the “top” SNP and the one that represents the locus. This top SNP might not be underlying causal SNP and the association arises just as a result of the correlation with the phenotype-causing SNP due the LD structure and background allele frequencies at the locus [264]. The top SNP is assumed to capture the maximum variation in the locus under study, and this is an assumption that may not hold true in two plausible scenarios: Despite the presence of only one causal variant in the locus, the top SNP might only partially capture the overall variation at the locus [264,265]. In the second scenario in which there is more than one independent causal variant at the locus, a single SNP is unlikely to capture all the LD structure between more than one unknown causal variant [264].

To efficiently annotate GWAS results at the locus level, it is necessary to uncover all independent causal variants to account for the total variation and the causal effect on the phenotype. GCTA-COJO is a tool that enables users to perform conditional and joint analyses based on a stepwise selection procedure to select the SNPs based on conditional P-values and likewise estimate the joint effects of all selected SNPs after model optimization [264]. In addition, GCTA-COJO enables us to perform association analysis conditioning on a given list of SNPs, to explore the conditional effects of all SNPs at a locus.

iii) Coloc

Colocalization is a powerful method to evaluate whether two independent signals at the same locus are consistent with a shared causal variant. Colocalization analysis can be performed between any pair of traits such as different GWAS traits or a case-control or phenotype GWAS vs a quantitative trait locus (QTL), which can represent any locus that is associated with the variation of a phenotypic trait. A typical setting

application that I have explored include a nominated locus from a GWAS, and an eQTL datasets containing loci associated with the variance in the levels of mRNA expression or splicing. When the two traits colocalize, it means they share one causal variant, hence, it is likely they also share biological mechanisms (i.e. in this example a causal variant from the nominated GWAS locus associated with the trait through the regulation in expression of a gene A).

There are different approaches to perform colocalization analysis. During my PhD, I have used coloc, an R package developed by Chris Wallace [266]. Coloc tests, under a Bayesian Inference framework, five different hypothesis for two datasets with the same allele frequency, and LD that is, samples from the same ethnic group:

- H0: No association with either trait
- H1: Association with trait 1, not with trait 2
- H2: Association with trait 2, not with trait 1
- H3: Association with trait 1 and trait 2, two independent SNPs
- H4: Association with trait 1 and trait 2, one shared SNP

For a region of Q variants, coloc constructs binary vectors for each trait of length Q, with 0 meaning no association and 1 meaning association. Then, it integrates all possible configurations, by using prior probabilities at the SNP level (prior probability of SNP associated with trait 1, trait 2, not with both traits). The computed probability of the data for each configuration together with the prior probabilities, can be used to compute the posterior probabilities for each hypothesis H [266].

A high enough H4 probability supports under this Bayesian framework that the two traits colocalize. Normally a PP H4 higher or equal to 0.8 is robust enough to be confident in that the two traits colocalize, in other words that the association signal for risk or the primary trait of interest, colocalizes with genetic risk for the second potential causal trait.

Recently a new framework to run coloc that incorporates the Sum of Single Effects (SuSiE) has been developed [267]. To assume that the single causal variant assumption holds, the new coloc framework makes use of SuSiE to partition the problem, so that in each loci-derived cluster, only one causal variant exists [268].

I used coloc software to test colocalization for all genes within ±1Mb from the GWASs lead SNPs using the eQTLGen and MetaBrain Cortex tissue meta-analysis eQTL data [269,270]. I used these two datasets as they are the largest blood and brain eQTL studies respectively, providing us with the greatest power to perform statistical co-localization tests. However, it is worth noting that the prior for H3 hypothesis (association with both phenotypic and expression traits, but distinct causal variants) is $\approx n(n-1)p1\ p2$, which scales with the square of n, resulting in H3 becoming more likely than H4 as the number of overlapping SNPs in the region tested increases [271]. This affects the colocalization tests against MetaBrain and eQTLGen meta-analyses. Therefore, I also performed two sensitivity analyses, adjusting the priors according to the number of overlapping SNPs [272], and also performing co-localization against PsychENCODE, which resulted in a considerable decrease in overlapping SNPs compared to the overlap against eQTL meta-analyses (MetaBrain and eQTLGen).

iv) FUMA

FUMA is a web-based platform that enables the interpretation of GWAS results by integrating a wide range of biological data. Using GWASs as input, FUMA allows us to gain insight into the biological implications of loci of interest. There are two separate steps within the FUMA framework [273].

SNP2GENE: is used to annotate SNPs according to their consequences in biological functionality, Combined Annotation Dependent Depletion (CADD) score, regulomeDB score, the chromatin state, the effects on gene expression, and chromatin interactions based on 3D structure chromatin data.

GENE2FUNC: provides the user with information regarding the putative biological mechanisms of the nominated genes from SNP2GENE step. For the gene set nominated for the phenotype or trait of interest, it gathers information about previous diseases associated, existing genes drug targets as well as genes' differential expression across a wide range of tissues from GTEx data. Moreover, an enrichment of the input genes (gene set enrichment analysis) in biological pathways and functional categories is also carried out.

v) echolocatoR

EcholocatoR is an R package that enables end-to-end statistical and functional fine-mapping as well as enrichment and annotation of results. I used the 'echolocatoR' R package (v 0.2.2) as a wrapper to perform fine-mapping based on ABF, FINEMAP, SuSiE, PolyFun [263,274–277]. I produced the 95% Probability Credible Set ($CS_{95\%}$). I reported the consensus SNPs at each locus, i.e. those that were included in the 95 $CS_{95\%}$ of at least two fine-mapping tools, therefore increasing the confidence in the nominated causal SNPs. I reported the PP as the mean PP across all fine-mapping tools. To account for SNP LD at each region, I used the precomputed LD matrix from the UK Biobank [278].

I also used echolocatoR to overlay the GWAS nominated loci with annotations of transcriptional activity using the assay for chromatin immunoprecipitation sequencing (CHIP-seq) data, and chromatin accessibility using the Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) data, and other gene expression regulatory information from transcription factor binding site marks assays. Primarily, I used tissue and cell type or line -specific genome-wide annotations from Roadmap, ENCODE, and FANTOM5 [255,279,280]. In addition, I accessed brain cell type-specific ChIP-seq data generated by quantifying H3K4me3 and H3K27ac epigenetic modifications, ATAC-seq data, and Proximity Ligation-Assisted ChIP-Seq (PLAC-Seq) data, which is a genomic assay that combines chromatin immunoprecipitation (ChIP) with proximity ligation to map long-range chromatin interactions mediated by specific proteins, such as transcription factors or histone modifications. This assay enables the identification of chromatin loops and regulatory interactions at high resolution, providing insights into the spatial organization of the genome and gene regulation [281].

vi) Cell and tissue enrichment analyses

MAGMA

To perform cell type enrichment analysis with PD traits, I used MAGMA software [115]. I first mapped SNPs to genes to obtain gene-level summary statistics based on a window size of 10kb upstream and 1.5kb downstream of each gene. Likewise, SNP level P-values can be aggregated based on the gene window size into a gene-level P-

value based on a SNP-wise mean association model, which uses a sum of squared SNP Z-statistics. Then, we can use MAGMA to test for association of the gene-level summary statistics of a trait with specificity matrices derived from tissue and cell -level expression data for each gene (specificity is defined based on a given expression in a cell type divided by the total expression of that gene based on its overall expression from all tissues or cell types). Specificity measures are then grouped into bins so that when testing for a positive association (one-sided test) between the bins and gene level summary statistics, I am evaluating whether an increase in the specificity level of a tissue or cell is associated with an an enrichment for common-variation for the GWAS traits being assessed. Confounding factors are taken into account by adding gene size, log(gene size), gene density and long(gene density) as covariates [115]. To perform enrichment analyses with MAGMA, I used the MAGMA.Celltyping R package hosted on GitHub which eases the automatization of large-scale enrichment analysis [282].

Stratified LD Score Regression (S-LDSC)

Linkage disequilibrium score regression (LDSC) was a tool developed to understand the inflation in test statistics in GWAS driven by true polygenic effects and bias such as cryptic relatedness and population structure [215]. The same year, another implementation of LDSC, S-LDSC was developed based on the prior knowledge that functional categories of the genome contribute disproportionately to the heritability of complex diseases [283]. S-LDSC is applied for partitioning heritability from GWAS, and it can be applied to perform cell type enrichment analyses based on S-LDSC method [283]. To do so, I generated annotation files for each cell type compatible with S-LDSC software with SNP level information. These annotation files contain information with the SNPs mapped to genes that belong to the 10% most specific genes for a given cell type. To derive the 10% most specific gene lists for each cell type, I processed the expression data described above to scale it to a total of 1 million Unique Molecular Identifiers (UMI) or 1 transcript per million (TPM) for each cell type or tissue. I only analysed genes with at least 1 million UMI or TPM in the cell type under study. The specificity measures were calculated by dividing the expression of a gene in a cell type by the total expression of that gene in all cell types. The analysis was limited to SNPs matching the Hapmap3 SNPs, as well as excluded the major

histocompatibility complex (MHC) due to its complex LD structure and high gene density (GitHub wiki) [284].

To map SNPs to genes, I extended gene ranges with 100Kb up and downstream to capture regulatory elements. Then, I added the mapped SNPs to the S-LDSC baseline model that consist of 53 functional annotations to take into account differences in heritability across the genome based on the activity and function of the region. I generated one annotation file per cell type.

S-LDSC computes the proportion of SNP heritability with each cell type from the annotation file, while taking into account the 53 annotations, therefore weighting the region's heritability according to the functional activity of the specific locus. Then, LDSC calculates an enrichment score and the coefficient Z-score P-value of the enrichment. The significance threshold was set to a 5% false discovery rate. For this analysis, I used custom scripts, mirroring the methods described by Bryois and colleagues [285].

vii) Mendelian Randomization

Mendelian randomization (MR) is a method to test the causal relationship of an exposure variable on an outcome driven by genetic variants. MR is normally based on an instrumental variable (IV) analysis, in which an instrument (i.e. genetic variant) is only associated with the outcome through its association with the exposure. MR makes the assumption that genetic variants provide a source of variation associated with the exposure and that is unrelated to the outcome. Therefore, the design of an MR analysis involves defining the association between a genetic variant (G) and an outcome (Y) which is used to test and quantify if an exposure of interest (X) influences the outcome, in the case in which the genetic variant is associated with the exposure and has no other path of association with the outcome [286].

There are three main conditions that need to hold for IV analysis to have a valid scenario in which I can test the null hypothesis that the exposure is not associated (or have effect) with the outcome [286].

- Relevance: IV is associated with the exposure
- Exchangeability: The IV does not have other mechanisms that influence the outcome other than through the exposure of interest.

- Exclusion restriction: The IV does not affect any other trait that has an effect on the outcome assessed.

d) Code availability

All the code for my analyses have been published on GitHub (https://github.com/AMCalejandro)

3) Genome-wide meta-analysis of Motor Progression in Parkinson Disease

a) Introduction

The majority of PD genetic studies have focused on case-control GWAS to explore the genetic factors contributing to the risk for PD [51]. However, little is known yet about the non-overlapping genetic factors that contribute to PD onset and progression across the multiple PD axes (cognitive decline, motor decline, non-motor and non-cognitive features).

In this study, I focused on modelling the early stages of motor Parkinson's disease, using the total score from the MDS-UPDRS part III. This validated scale is recommended for clinical trials to assess both the response to levodopa treatment and the rate of change over time [157]. Furthermore, I derived and explored axial and limb motor stages from the MDS-UPDRS part III scale, based on my hypothesis of a potential connection between different modules of MDS-UPDRS part III and specific pathological processes [287].

By using GWAS and meta-analysis, I aimed to identify genetic determinants associated with variability in motor progression and severity in the early stages of PD. This analysis led to findings significantly correlated with changes in the MDS-UPDRS part III scale. In addition, I performed functional annotation and fine-mapping analyses to unravel how the nominated genetic variants are associated with the regulation of gene expression and the fundamental biology underlying PD motor traits.

b) Methods

Code used in the analysis is available from github.com/AMCalejandro/EMPD (https://doi.org/10.5281/zenodo.7258985).

i) Study Design and data Quality Control

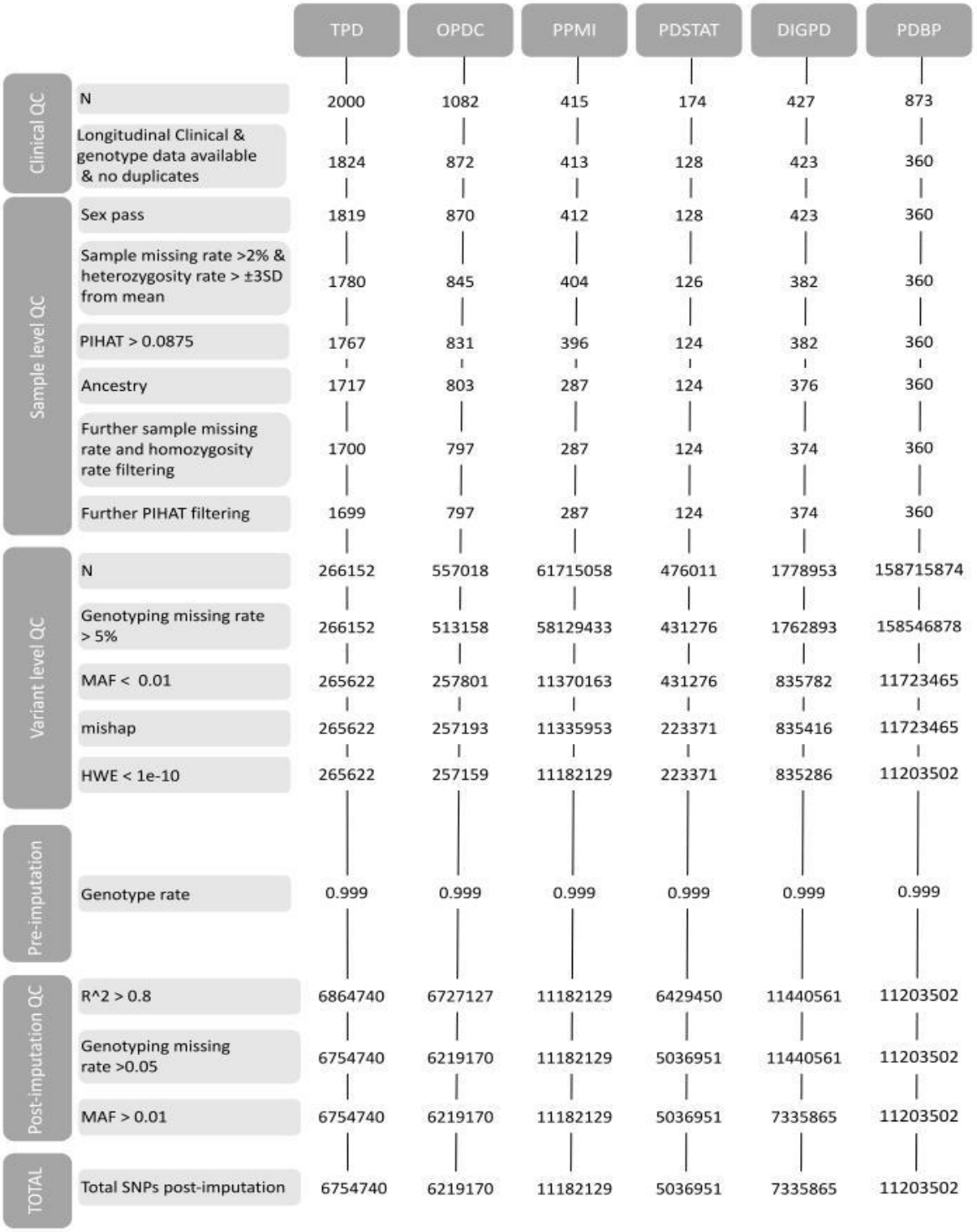
I examined six observational and interventional longitudinal cohorts of Parkinson's disease (PD), comprising a total of 4,971 patients with available genotyping or whole genome sequencing (WGS) data (**Table 2**).

***Table 2**. Study sample sizes and genotyping array.*

Study Name	Abbreviations	N	Genotyping array	Period of recruitment
Tracking Parkinson's	TPD	2000	Illumina HumanCoreExome array	Recruitment between 2012 and 2014
Oxford Parkinson's Disease Centre Discovery Cohort	OPDC	1082	Illumina HumanCoreExome -12 v1.1 or Illumina Infinium HumanCoreExome -24 v1.1	Study onset 2010. Recruitment is still ongoing.
Drug Interaction With Genes in Parkinson's Disease	DIGPD	427	Illumina Infinium Multi-Ethnic Global (MEGA)	Recruitment between 2009 and 2013
Parkinson's Progression Markers Initiative	PPMI	415	WGS	Recruitment between 2010 and 2012
Advancing Parkinson's Disease Biomarkers Discovery	PDBP	873	WGS	Recruitment between 2012 and 2014
Simvastatin as a neuroprotective treatment for Parkinson's disease	PD-STAT	174	Neurochip	Recruitment between 2016 and 2018

I selected cohorts based on the availability of longitudinal assessments using the Movement Disorder Society–Unified Parkinson's Disease Rating Scale part III (MDS-UPDRS). We kept individuals with matching clinical and genotyping data, removed duplicated samples, (**Figure 3**)

***Figure 3**. Quality Control flowchart.*



Each row shows a different QC step at different levels (Clinical QC, Sample level QC, Variant level QC, Post-Imputation QC) across each cohort displayed as a column. In addition, a metric of the genotyping rate prior imputation is shown in Pre-imputation. The resulting number of SNPs available for the study in each cohort is shown in TOTAL.

I defined limb and axial phenotypes based on established criteria using the MDS-UPDRS part III scale [288]. Additionally, the MDS-UPDRS III total score served as an overall measure of PD motor signs (**Table 3**).

***Table 3**. Limb, total, and axial PD motor measures derived from MDS-UPDRS.*

Motor Score	Scores from MDS-UPDRS III
MDS-UPDRS part III - Total	Speech (3.1) , Facial expression (3.2), Rigidity (3.3), Finger tapping (3.4), Hand movement (3.5), Pronation-supination movements of hands (3.6) , toe tapping (3.7), leg agility (3.8), Arising from chair (3.9), Gait (3.10), freezing of gait (3.11), postural stability (3.12), Posture (3.13) global spontaneity of movement body (Body bradykinesia) (3.14), postural tremor of the hands (3.15), kinetic tremor of the hands (3.16), rest tremor amplitude (3.17), constancy of rest tremor (3.18)
MDS-UPDRS part III - Limb	Rigidity (3.3), postural tremor of the hands (3.15), kinetic tremor of the hands (3.16), rest tremor amplitude (3.17) Finger tapping (3.4), Hand movement (3.5), Pronation-supination movements of hands (3.6) , toe tapping (3.7), leg agility (3.8), constancy of tremor (3.18)
MDS-UPDRS part III - Axial	Speech (3.1) , Facial expression (3.2), Arising from chair (3.9), Gait (3.10), freezing of gait (3.11), postural stability (3.12), Posture (3.13), global spontaneity of movement body (Body bradykinesia) (3.14)

In this study, I included longitudinal data from all data sources up to 36 months from the baseline visit, with the aim of gathering a subset of data with low missingness rate (< 50% missingness). Over 36 months, I found cohorts to have a missingness percentage higher than 50%. I used imputation techniques to address missing motor outcomes. For participants with incomplete MDS-UPDRS part III data I scaled up the limb, axial, and total scores. For each patient's time specific MDS-UPDRS part III measures, when no more than 20% of the total scores from each motor sub score (total, limb, axial) were missing at random, I scaled up the score summing the total score across motor sub scores, divided by the number of non-missing sub scores, multiplied by the total number of scores on each motor sub score. If more than 20% of the total scores per motor subscale were missing, I set the motor subscale as missing, and excluded that data point. On the other hand, if there were items in the MDS-UPDRS part III scale consistently missing (missing not at random), I scaled up the total motor score only when there were up to 3 measures missing not at random [289].

I conducted genetic QC at both the sample and variant levels, followed by imputation using the MIS, and post-imputation QC. I applied standard sample QC steps across cohorts using plink v1.9 [290] (**Figure 3**). A more detailed explanation on the QC steps is in the **Chapter 2 – Methods**.

Earlier research has indicated that levodopa enhances motor state examination and may potentially decelerate the progression of the disease [291]. Given that the observable motor improvement occurs a few hours after treatment and influences the Movement Disorder Society–Unified Parkinson's Disease Rating Scale (MDS-UPDRS) measure, I conducted comparisons among individuals within cohorts in the same state during each assessment. If cohorts had data available in the "OFF" state, I utilised longitudinal "OFF" vs "OFF" MDS-UPDRS part III scores; otherwise, I conducted "ON" vs "ON" comparisons. Additionally, I conducted sensitivity analyses by adjusting the motor scores based on levodopa dosage. To take into account the effect of dopaminergic treatment and doses on the motor scores derived from MDS-UPDRS part III, I performed a sensitivity analysis with the adjusted total, limb, and axial motor scores, using a correction factor according to the effect of levodopa dose on the MDS-UPDRS scale. It is well known that levodopa treatment improves MDS-UPDRS scores in the majority of PD patients [291]. To figure out whether any genetic association with the motor states was masked due to levodopa dosage, I used an equation that best predicted the effect of levodopa dose on MDS-UPDRS part III total over time to correct the motor scores by levodopa usage, provided by Dr Michael Lawton that best predicted the MDS-UPDRS III trend based on levodopa uptake. I used data from Tracking Parkinson's Levodopa challenge with motor subscores recorded at baseline before and after treatment in order to weight the effect of levodopa usage on the limb and axial motor states. I found that over the average difference in the MDS-UPDRS part III total score pre and post dose at baseline (9.9 points difference in average), 7.7 point change was explained by the limb composite score and 2.2 change due to the axial score. I used such weights on the equation that best predicted what I would expect to happen long term with levodopa usage, and I derived the adjusted longitudinal outcomes across cohorts.

ii) Statistical approaches

To assess the impact of genetics on both motor progression and baseline variability, I employed LMMs, a statistical model introduced in **Chapter 2 - Methods**. I explored changes in limb and axial motor severity and progression associated with genetic variants, focusing on an additive genetic effect.

Model 1: Disease Progression

$$\text{OUTCOME}_{ij} = \beta_0 + \beta_1 \text{TIME}_{ij} + \beta_2 \text{SNP}_{ij} + \beta_3 (\text{TIME} \times \text{SNP})_{ij} + \beta_4 \text{GENDER}_{ij} + \beta_5 \text{AGE\_DIAGNOSIS.STD}_{ij} + \beta_6 \text{PC1}_{ij} + \beta_7 \text{PC2}_{ij} + \beta_8 \text{PC3}_{ij} + \beta_9 \text{PC4}_{ij} + \beta_1 0 \text{PC5}_{ij} + (u_{0i} + u_{1i} \text{TIME}_{ij}) + \epsilon_{ij}$$

Where:

- β_0 is the fixed intercept.
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}$ are the fixed effect coefficients.
- u_{0i} is the random intercept for subject i.
- u_{1i} is the random slope for TIME for subject i.
- ϵ_{ij} is the residual error for subject i at time j.

Model 2: Disease Severity

$$\text{OUTCOME}_{ij} = \beta_0 + \beta_1 \text{SNP}_{ij} + \beta_2 \text{GENDER}_{ij} + \beta_3 \text{AGE\_DIAGNOSIS.STD}_{ij} + \beta_4 \text{PC1}_{ij} + \beta_5 \text{PC2}_{ij} + \beta_6 \text{PC3}_{ij} + \beta_7 \text{PC4}_{ij} + \beta_8 \text{PC5}_{ij} + u_{0i} + \epsilon_{ij}$$

Where:

- β_0 is the fixed intercept.
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8$ are the fixed effect coefficients.
- u_{0i} is the random intercept for subject i.
- ϵ_{ij} is the residual error for subject i at time j.

Disease progression model is a LMM with random variability at both the intercept and the slope level. I allowed for individual's intercepts to deviate from the global intercept as well as time individual's slopes to deviate from the global average time slope, while allowing correlation between the intercept deviations and time effect deviations within individual levels. I selected the disease progression model under the assumption that there are differences in patients' progression that could be explained from genetics.

Disease severity model is a LMM with random variability at the intercept level only. I allowed the individual's intercepts to deviate from the global intercept. It is adjusted by the confounding variables only. This model assumes that there is variability around

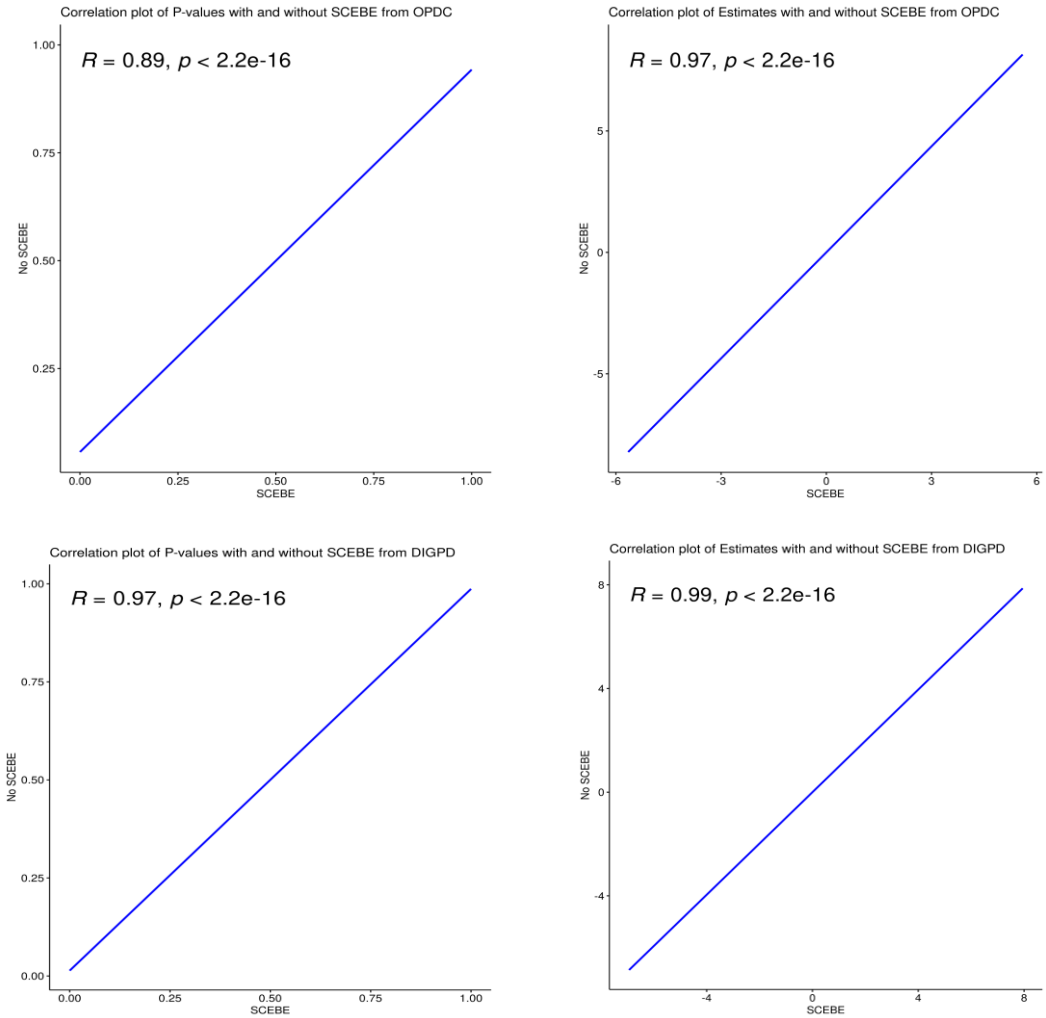
the motor stage patients were when the study began. In addition, the model assumes the motor decline would remain constant without differences between subjects. I selected the disease severity model under the assumption there is no unexplained variability in Parkinson's disease progression.

In the disease severity model, I examined the additive genetic effect (either an increase or a decrease) of SNPs on the average motor score while adjusting for the rest of covariates, or in other words, the additive genetic effect on patients disease severity. For the disease progression model, I investigated the genetic additive effect of SNPs on the motor rate of change.

I evaluated the power of LMMs for investigating SNP effects on changes in PD motor signs. The power of GWAS depends on the sample size, the frequency of those variants associated with traits, their effect size, the heterogeneity of the trait studied, and the LD structure. The power increases with sample size, variant effect size, and allele frequency. To estimate the power of LMMs in GWAS, I performed a power calculation across combinations of sample sizes, allele frequencies, and effect sizes in R. I carried out 10000 simulations and tested the association of 1000 dummy SNPs with different effect sizes (total MDS-UPDRS III rate of decline), different AFs for those SNPs, and for three different sample sizes. I reported the power as the number of times a SNP was found to be significantly associated with the outcome accounting for multiple testing (P = 0.05 / N SNPs), divided by the number of simulations.

For the disease progression model, I made use of the lmerTest R package (v. 3.1-3) and the Satterthwaite approach to approximate degrees of freedom, deriving p-values using restricted maximum likelihood (REML) due to its acceptable type-1 error rates [292]. Additionally, I employed the SCEBE [193] algorithm (v. 0.1.0) with REML and the lme4 R package (v. 1.1-30) to reduce computational costs by introducing unexplained variability at the slope level in the disease severity model. I validated SCEBE in two separate cohorts (**Figure 4**). All tests were two-tailed.

***Figure 4**. SCEBE validation in two independent cohorts.*



Pearson correlation plots between P-values and Coefficients derived with SCEBE approach (X-axis) and with lmeRTest using the Satterwhite approach to derive P-values (Y-axis). The two top figures are the correlation plots of models fitted with OPDC data. The two bottom figures are the correlation plots of models fitted using DIGPD data. I used MDS-UPDRS III total as the outcome of the model. Each plot shows the correlation value (R^2), and the significance of the correlation (P).

For meta-analysis of genome-wide association summary statistics, I used the METAL software (version released on 25/03/2011). The meta-analysis is based on a fixed-effects model weighted by β coefficients and the inverse of standard errors [244,245]. Additionally, I applied quality control to the meta-analysis results **as described in Chapter 2 - Methods**. Statistical significance was determined at the genome-wide level (p = 5e−8).

iii) Fine-Mapping and Functional Annotation

For each locus of interest, I implemented a conditional and stepwise model selection procedure to identify independently associated SNPs for each GWAS nominated locus [264]. Causal variant nominations were made through fine-mapping techniques [263,267,274,276,293] **as described in Chapter 2 – Methods**. To gain deeper insights into the regulatory mechanisms within these nominated loci, I cross-referenced each locus with (1) cell type-specific and general genome enhancer marks, (2) enhancer-transcription start site (TSS) interaction marks sourced from FANTOM5, and (3) transcriptional regulatory marks specific to brain cell types and distal enhancer-promoter interactions, using 'echolocatoR' R package (v 0.2.2) [281,294].

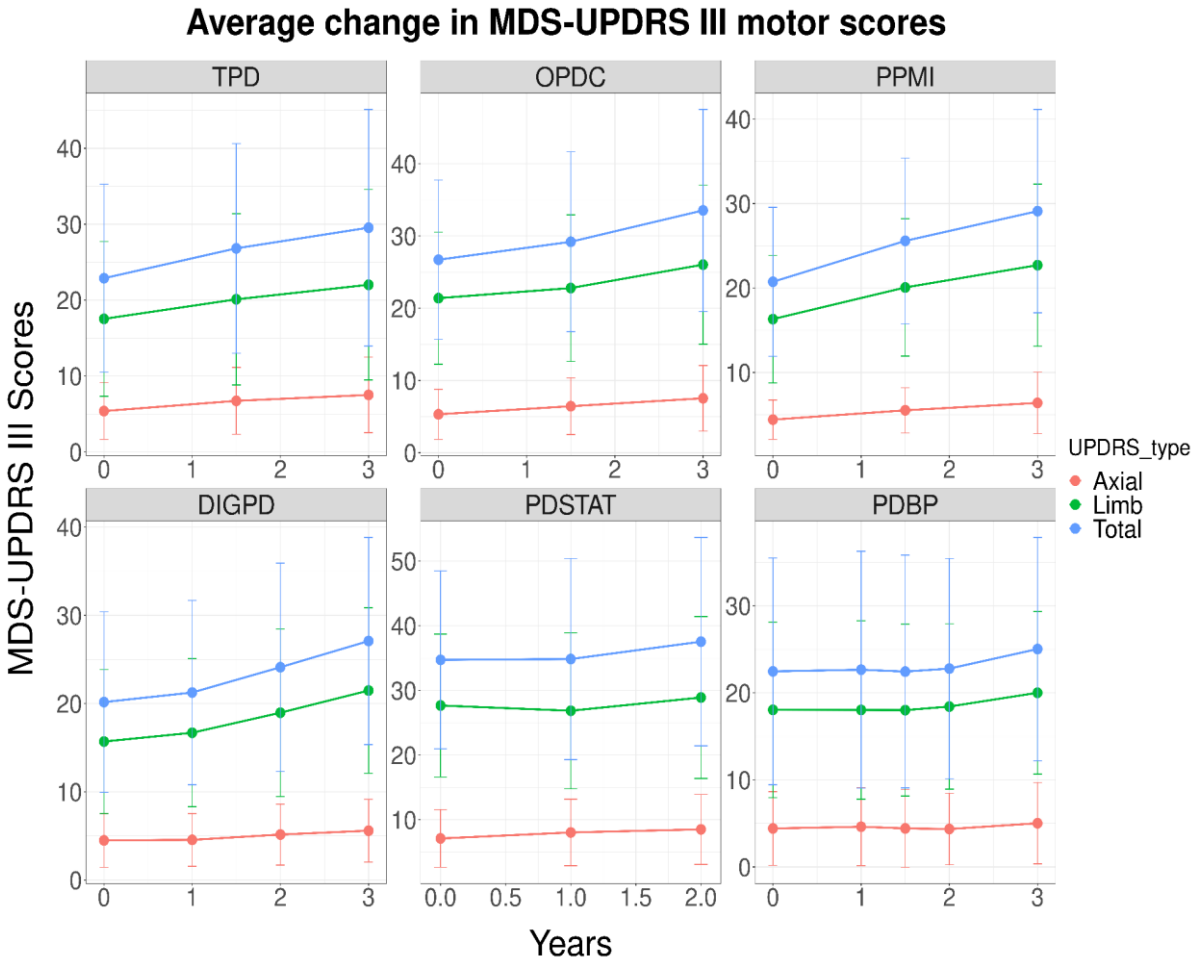
To assess whether causal variants might be linked to motor phenotypes via gene expression dysregulation, I conducted colocalization analyses using the coloc method against cis-expression quantitative trait loci (eQTL) datasets [266,269,295,296]. Furthermore, I employed FUMA, a web-based platform integrating a diverse range of functional annotation data (version 1.3.8) [273]. The LocusZoom tool (version 0.12) [297] was used to visually represent the LD structure of a given locus in relation to the lead SNP, along with the neighbouring protein coding genes and rRNAs.

c) Results

We explored the overall rate of change in MDS-UPDRS part III total, limb and axial scores explained in **Table 4**. There was variation across studies. We specifically studied the amount of change for the motor measures in each study by comparing the final score with the baseline score, divided by the baseline score, for MDS-UPDRS-total, axial and limb. We found that the axial score rate of change was the highest in TPD, OPDC, PD-STAT, and PDBP. The limb rate of change was the highest in PPMI and DIGPD. PD-STAT and PDBP had a lower rate of changes, which may be due either to longer disease duration, or to selection effects related to the inclusion of "benign" PD in patients with longer disease duration. We assessed this by fitting a LMM using data from TPD, and found a significant interaction between time and disease duration related to MDS-UPDRS total progression (β = -0.11, SE = 0.04, P = 0.01). Longer disease duration was associated with a lower total rate of change in MDS-UPDRS, which appears to be non-linear with extended disease durations.

Overall, we confirmed that the MDS-UPDRS derived measures increased, reflecting worsening motor impairment, from study entry up to 3 years (**Figure 5**). The MDS-UPDRS part III total yearly rate of change ranged between 2.37 - 3.01 points/year, which is consistent with previous reports[179].

***Figure 5**. MDS-UPDRS III Motor Scores Trajectories.*



Trajectory of the MDS-UPDRS III–derived motor scores across cohorts. In the x-axis, the time point at which the MDS-UPDRS III assessment was measured. Each plot shows the motor scores trajectories on each cohort highlighted in the label. The y-axis represents the average scores for each of the motor states. The bars represent the SD of the average motor scores.

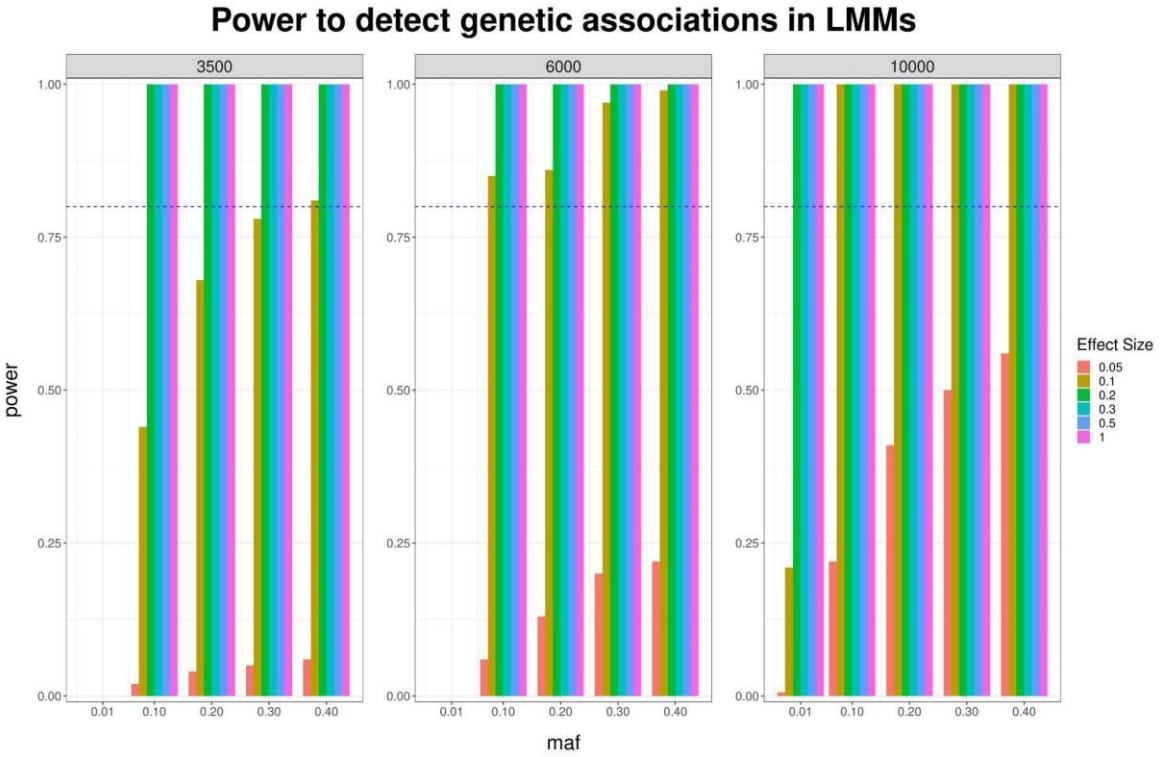
***Table 4**. Cohort demographics and motor scores rate of change.*

Study	N patients	N Observations	Visit interval, mo	ON/OFF	No. (%) male	AAD, years mean(sd)	AAB, years mean(sd)	Yearly total rate of change mean (sd)	total rate of change mean (sd)	limb rate of change mean (sd)	axial rate of change mean (sd)
TPD	1699	4349	18	ON	1101 (64.8)	66.20 (±9.24)	67.50 (±9.31)	2.7 (±4.69)	0.48 (±0.92)	0.54 (±1.25)	0.71 (±1.37)
OPDC	797	1978	18	ON	513 (64.37)	66.04 (±9.46)	67.25 (±9.57)	2.85 (±4.27)	0.50 (±0.77)	0.50 (±0.93)	1.02 (±1.76)
PPMI	287	1653	3 & 6 & 12	OFF	184 (64.11)	61.01 (±9.73)	61.59 (±9.70)	3.01 (±3.65)	0.84 (±0.94)	1.03 (±1.23)	0.67 (±1.05)
PD STAT	124	358	12	OFF	76 (61.29)	NA	66.08 (±9.37)	1.70 (±5.76)	0.06 (±0.41)	0.07 (±0.58)	0.27 (±0.61)
DIGPD	305	1005	12	ON	184 (60.33)	59.68 (±9.85)	62.59 (±9.70)	2.37 (±3.41)	0.56 (±0.87)	0.67 (±1.40)	0.52 (±0.95)
PDBP	360	2090	6	ON	222 (61.67)	59.9 (±10.90)	64.73 (±9.15)	1.05 (±4.11)	0.21 (±0.61)	0.30 (±0.90)	0.40 (±1.46)

Abbreviations = N, Number; AAD, Age at Diagnosis; AAB, Age at Baseline
Total rate of change per year: (Last visit score - Baseline score) ÷ number of years
Total/Limb/Axial rate of change centred: (Last visit score - Baseline score) ÷

My power calculation showed that the current LMM was well powered to detect high effect sizes (β ≥ 0.2) for a wide range of different MAFs, with a limit for variants with an allele frequency ≥ 1% (**Figure 6**). We performed a GWAS on each cohort to study PD motor progression and meta-analysed results separately using a genomic control to correct the test statistics of those cohorts that had genomic inflation (λ > 1 & λ < 1.2).

Figure 6. Power to detect genetic associations in LMMs.



The Y axis shows the power (0 to 1). The X axis shows the MAF of the SNP tested 10000 times. The header of each plot represents the sample size. Different colours represent the simulated effect size.

I evaluated disease progression and disease severity models for total, limb and axial progression. We did not find any significant genetic association with the PD limb motor progression or severity. For axial motor progression, I found one haplotype block that reached genome-wide significance (*GJA5* in chromosome 1) **(Figure 7a**; **Figure 7c**).

***Figure 7**. GWAS meta-analysis of motor axial progression.*

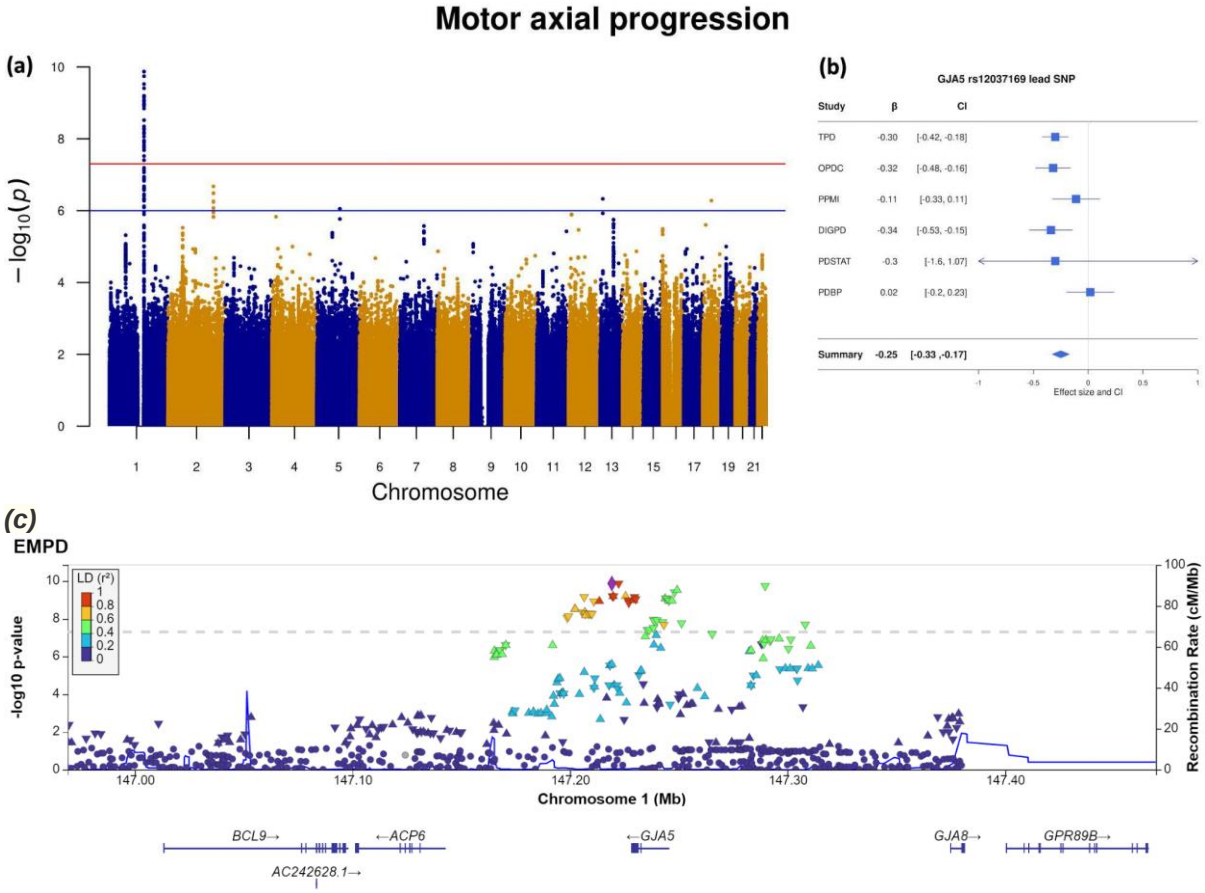


***Figure 7a**.Manhattan plot of the rate of axial change GWAS meta-analysis. On the X axis each of the 22 chromosomes, and each SNP P-value on the Y axis. The red dashed line indicates the genome-wide significance threshold P-value = 5e-8. The LD block that reached genome wide significance on chromosome 1 is on the GJA5 locus. Each dot corresponds to the P-Value of the conditional likelihood interaction term between SNP and time (SNP*time). There was no genomic inflation (λ = 0.99). **Figure 7b**. Forest plots for proxy variant rs12037169 within GJA5 locus under the GWAS meta-analysis using disease progression model for the axial outcome (I^2 = 40.1; Cochran's Q test: χ^2 = 9.64, df = 5, P = 0.10), annotated by study, effect size, and the corresponding 95% confidence interval. **Figure 7c.** GWAS locusZoom plot. LocusZoom plot centred around the lead SNP at the GJA5 locus. SNPs are coloured according to the LD (r^2) with the lead variant (purple). The corresponding degree of LD for each colour, is given in the plot label.*

This association was also found, at a lower significance level, for the MDS-UPDRS part III total. Given that there was no association with PD limb motor progression and severity, this relates to the inclusion of axial components in the overall MDS-UPDRS-III total score. Although the lead variant in the *GJA5* locus was not captured in the PPMI WGS data, I found proxy variants that were present in all cohorts. The lead proxy variant was rs12037169 (β = -0.25, SE = 0.04, *P* = 3.93e^{-10}) (**Table 5**). The association

test-statistic and directionality of each of these variants was consistent across cohorts (**Figure 7b**).

***Table 5**. Lead SNPs on the disease progression and severity GWASs.*

SNP	CHR	A1	MAF	BETA	SE	P-value	NEAREST GENE	TYPE OF VARIANT	MODEL
rs6593808	1	A	0.23	-0.28	0.04	1.35e-10	*GJA5*	intergenic	progression
rs12037169	1	A	0.25	-0.25	0.04	3.93e-10	*GJA5*	intergenic	progression
rs4073509	2	C	0.02	0.52	0.10	2.12e-07	*AC098872.3*	intergenic	progression
rs117239007	13	C	0.01	0.68	0.14	4.71e-07	*LINC00544*	intergenic	progression
rs36082764	17	T	0.42	-0.62	0.11	6.34e-08	*LINC00511*	ncRNA_intronic	severity
rs4721411	7	T	0.40	0.53	0.10	1.66e-07	*MAD1L1*	intronic	severity
rs10939702	4	T	0.45	0.57	0.12	8.10e-07	*WDR1*	intronic	severity

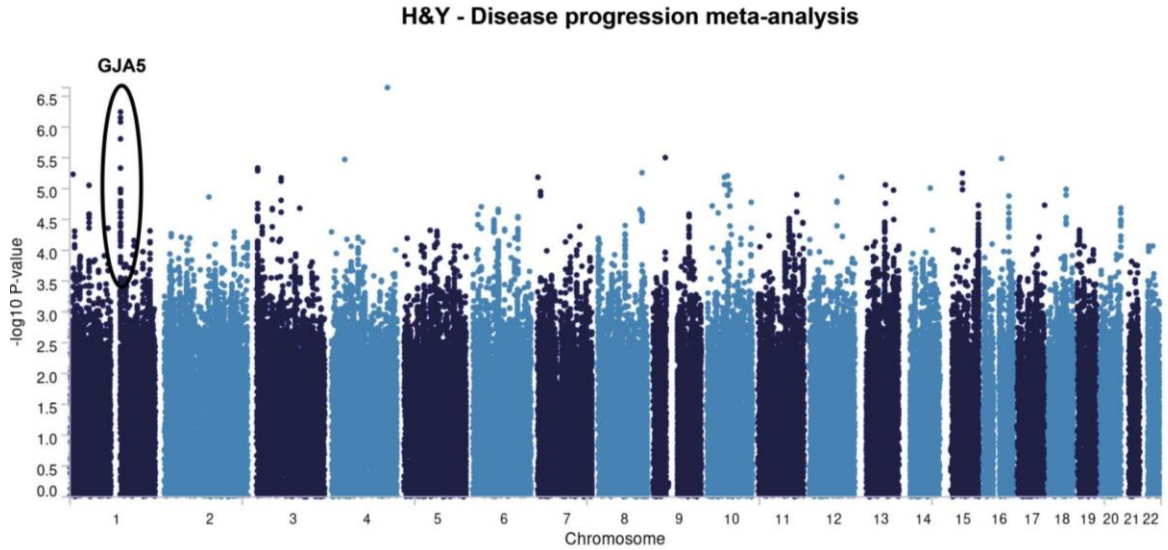
To assess whether levodopa presented a substantial confounding factor in my motor progression study, I adjusted patient motor scores as described in the **Methods section**. This equation, designed to best estimate the impact of levodopa dose on MDS-UPDRS Part III total scores over time (described in the **Statistical approaches of this Chapter)**, allowed us to correct the motor scores based on levodopa usage. To account for the influence of levodopa on limb and axial motor states, I incorporated data from the Tracking Parkinson's Levodopa challenge [298], which included MDS-UPDRS Part III scores recorded before and after treatment. I applied these weights to correct motor scores with respect to levodopa usage. Notably, I observed no significant alterations in the significance level or direction of effects for the tested SNPs. Furthermore, rs120371169 maintained a significant association with axial motor progression (**Table 6**).

***Table 6**. Metrics per cohort of lead SNPs found on the disease severity and progression GWASs meta-analysis.*

Cohort	rsID	Effect allele	MAF	Beta	se	P-Value	Levodopa Adjusted	Model
TPD	rs12037169	A	0.24	-0.30	0.06	1.67e-06	No	progression
OPDC	rs12037169	A	0.25	-0.33	0.09	1.75e-04	No	progression
PPMI	rs12037169	A	0.27	-0.10	0.35	0.44	No	progression
DIGPD	rs12037169	A	0.24	-0.29	0.11	1.84e-03	No	progression
PDSTAT	rs12037169	A	0.25	-0.33	0.25	0.18	No	progression
PDBP	rs12037169	A	0.23	0.07	0.18	0.59	No	progression
TPD	rs12037169	A	0.24	-0.30	0.07	8.41e-06	Yes	progression
OPDC	rs12037169	A	0.25	-0.35	0.09	6.20e-04	Yes	progression
PPMI	rs12037169	A	0.27	-0.21	0.39	0.6	Yes	progression
DIGPD	rs12037169	A	0.24	-0.28	0.11	1.84e-03	Yes	progression
TPD	rs36082764	T	0.44	-0.66	0.17	2.50e-04	No	severity
OPDC	rs36082764	NA	NA	NA	NA	NA	No	severity
PPMI	rs36082764	T	0.46	-0.66	0.19	6.50e-04	No	severity
DIGPD	rs36082764	T	0.38	-0.26	0.36	0.44	No	severity
PDBP	rs36082764	T	0.43	-0.65	0.28	1.00e-02	No	severity
PD-STAT	rs36082764	NA	NA	NA	NA	NA	No	severity
TPD	rs4721411	T	0.42	0.62	0.18	6.00e-04	No	severity
OPDC	rs4721411	T	0.41	0.65	0.24	8.00e-03	No	severity
PPMI	rs4721411	T	0.38	0.43	0.19	2.00e-02	No	severity
DIGPD	rs4721411	T	0.43	0.42	0.36	0.2	No	severity
PDBP	rs4721411	T	0.40	0.70	0.27	9.00e-03	No	severity
PD-STAT	rs4721411	T	0.42	-0.40	0.52	0.45	No	severity
TPD	rs36082764	T	0.44	-0.65	0.18	3.00e-04	Yes	severity
OPDC	rs36082764	NA	NA	NA	NA	NA	Yes	severity
PPMI	rs36082764	T	0.46	-0.70	0.21	8.00e-04	Yes	severity
DIGPD	rs36082764	T	0.38	-0.32	0.36	0.38	Yes	severity
TPD	rs4721411	T	0.42	0.62	0.18	6.00e-04	Yes	severity
OPDC	rs4721411	T	0.41	0.65	0.24	8.00e-03	Yes	severity
PPMI	rs4721411	T	0.38	0.33	0.21	9.00e-02	Yes	severity
DIGPD	rs4721411	T	0.42	0.38	0.36	0.29	Yes	severity

Hoehn and Yahr (HY), a metric capturing a patient's disease severity, can also be used over time to assess disease progression. As a way of validating the genome-wide significant association linked to PD axial motor progression, I used a disease progression statistical model incorporating HY as my longitudinal outcome to explore the contribution of SNPs to motor changes over time. Within the *GJA5* locus, at the same locus which was found significantly associated with axial motor progression, I identified an LD block approaching genome-wide significance (**Figure 8**). The lead variant in this block was rs36005900 (β = −0.08, SE = 0.0078, p = 5.7e−7). Notably, the directionality of the effects mirrored those observed in the axial motor progression GWAS. Furthermore, rs36005900 was in LD with the lead variant reported in the same locus for MDS-UPDRS III axial motor progression (D' = 0.8, R2 = 0.6).

***Figure 8**. Manhattan plot for disease progression GWAS meta-analysis using HY as the outcome.*



I then investigated whether there were independently associated SNPs at the *GJA5* locus. We did not find any signal other than the lead SNP in the selection procedure under a conditional and stepwise selection approach using GCTA-COJO. Under a single causal variant assumption, I then performed statistical fine mapping. I did not find a consensus SNPs (a SNP nominated to be causal by 2 different fine-mapping tools) at the *GJA5* locus. I found a total of 12 SNPs with support for causality of changes in motor axial progression, nominated from at least one fine-mapping tool (**Table 7**). I did not find an overlap between the GJA5 locus haplotype block and

regulatory marks from functional annotation datasets described in the **Chapter 2 - Methods.**

***Table 1**. Fine-mapping results using ABF, FINEMAP, SUSIE, and POYFUN_SUSIE*

LOCUS	SNP	P	leadSNP	ABF	FINEMAP	SUSIE	POLYFUN_SUSIE	Sup	mean.PP
GJA5	rs2353	1.3e-10	FALSE	0.13	0	1	0	1	0.28
GJA5	rs12032789	6.4e-10	FALSE	0.03	0	1	0	1	0.26
GJA5	rs1342711	6.4e-10	FALSE	0.03	0	1	0	1	0.25
GJA5	rs2352870	7.0e-10	FALSE	0.03	1	0	0	1	0.25
GJA5	rs10793706	8.5e-10	FALSE	0.02	0	0	1	1	0.25
GJA5	rs10793707	8.5e-10	FALSE	0.02	0	0	1	1	0.25
GJA5	rs12408247	8.5e-10	FALSE	0.02	0	0	1	1	0.25
GJA5	rs11552588	1.1e-09	FALSE	0.02	0	0	1	1	0.25
GJA5	rs35594137	1.1e-09	FALSE	0.02	0	0	1	1	0.25
GJA5	rs11576092	8.7e-09	FALSE	0.00	1	0	0	1	0.25
GJA5	rs1573101	1.4e-05	FALSE	2.5e-06	1	0	0	1	0.25
GJA5	rs4443942	9.8e-05	FALSE	4.5e-07	1	0	0	1	0.25
MAD1L1	rs3778978	4.5e-07	FALSE	0.02	NA	1	1	2	0.50
LINC00511	rs7213651	3.7e-06	FALSE	0.02	0.86	1	1	3	0.72
LINC00511	rs7218929	7.6e-06	FALSE	0.01	0.07	1	1	3	0.52
LINC00511	rs12950478	2.5e-05	FALSE	0.01	NA	1	1	2	0.50

Abbreviation = N, Sample size to do fine-mapping; t_stat = test statistic; CS = Credible Set; PP = Posterior Probability. mean.PP = the mean posterior probability from the four fine-mapping posterior probability.

I also explored expression quantitative trait loci (eQTL) datasets through the FUMA platform. We found that many of the GWAS significant SNPs within the *GJA5* locus were significant cis-eQTLs for *ACP6, a gene* located 105 kb from the lead SNP, in PsychEncode, and eQTLGen. In particular, we found that the lead variant was a significant eQTL in PsychEncode, and eQTLGen, and also rs12037169, the proxy significant variant found in all cohorts, was a significant cis-eQTL in eQTLGen (**Table 8**). We then carried out a colocalization analysis to evaluate whether there was colocalization between the GWAS axial progression results and eQTL GWAS for gene expression at the *GJA5* locus. I used cis-eQTL data from eQTLGen and Metabrain cortex tissue cis-eQTLs datasets and performed a colocalization test for any gene

within ±1Mb from the *GJA5* lead SNP. We did not find direct colocalization evidence for any gene, including *ACP6*. We found PPH3 (indicating separate significant associations for GWAS and eQTL analysis) to be the highest for the *ACP6* gene using default SNP priors (eQTLGen = 0.98, MetaBrain = 0.88). PPH3 was the highest for these two genes (PPH3 > 0.8), after we adjusted the priors according to the number of overlapping SNPs.

***Table 2**. GJA5 locus significant SNPs that are ACP6 eQTLs across different studies.*

SNP	CHR	POS	MAF	Nearest Gene	Gwas P-value	eQTL study	symbol	eQTL P-value
rs6593808	1	147219250	0.23	*GJA5*	1.349e-10	PsychENCODE	*ACP6*	1.18e-07
rs6593808	1	147219250	0.23	*GJA5*	1.349e-10	eQTLGen	*ACP6*	1.68e-14
rs2353	1	147222372	0.23	*GJA5*	1.349e-10	eQTLGen	*ACP6*	1.58e-14
rs7551148	1	147289707	0.25	*RP11-314N2.2*	1.815e-10	PsychENCODE	*ACP6*	1.23e-07
rs7551148	1	147289707	0.25	*RP11-314N2.2*	1.815e-10	eQTLGen	*ACP6*	1.08e-07
rs1495955	1	147249285	0.25	*RP11-433J22.3*	2.992e-10	PsychENCODE	*ACP6*	7.97e-05
rs1495955	1	147249285	0.25	*RP11-433J22.3*	2.992e-10	eQTLGen	*ACP6*	6.52e-17
rs12037169	1	147248057	0.25	*GJA5*	3.93e-10	eQTLGen	*ACP6*	8.64e-17
rs1857213	1	147219553	0.23	*GJA5*	6.383e-10	PsychENCODE	*ACP6*	1.18e-07
rs1857213	1	147219553	0.23	*GJA5*	6.383e-10	eQTLGen	*ACP6*	1.73e-14
rs1342711	1	147219835	0.23	*GJA5*	6.383e-10	PsychENCODE	*ACP6*	1.20e-07
rs1342711	1	147219835	0.23	*GJA5*	6.383e-10	eQTLGen	*ACP6*	2.10e-14
rs12032789	1	147220045	0.23	*GJA5*	6.383e-10	PsychENCODE	*ACP6*	7.91e-09
rs12032789	1	147220045	0.23	*GJA5*	6.383e-10	eQTLGen	*ACP6*	1.55e-14
rs36005900	1	147229662	0.23	*GJA5*	6.964e-10	eQTLGen	*ACP6*	1.92e-13
rs2352870	1	147206521	0.26	*GJA5*	7.019e-10	PsychENCODE	*ACP6*	1.18e-07

gwasP = P-value of a SNP in the GWAS study; eqtlP =P-value of a SNP in the eQTL study.

I then performed a colocalization analysis to evaluate whether there was colocalization between the GWAS axial progression results and eQTL GWAS for gene expression at the *GJA5* locus. No conclusive evidence of direct colocalization was identified for any gene, including *ACP6*. The highest Posterior Probability of Colocalization (PPH3), indicating distinct significant associations for GWAS and eQTL analyses, was observed for the *ACP6* gene with default SNP priors (eQTLGen = 0.98, MetaBrain =

0.88). Even after adjusting the priors based on the number of overlapping SNPs as described in **Chapter 2 - Methods**, PPH3 remained highest for these two genes (PPH3 > 0.8).

In addition, I investigated the genome-wide association of SNPs on average changes in limb and axial motor states using the disease severity model as highlighted in the **statistical approaches section in Chapter 2 – Methods**. No haplotype block reached genome-wide significance in this analysis. However, two distinct signals approached genome-wide significance, correlating with changes in average axial motor scores (*MAD1L1* on chromosome 7 and *LINC00511* on chromosome 17) (**Figure 9a**). The lead SNP in *MAD1L1* was identified as rs4721411 (β = 0.54, SE = 0.11, p = 1.6e−7), and the lead variant in the long noncoding RNA *LINC00511* was rs36082764 (β = −0.62, SE = 0.11, p = 6.3e−8) (see **Table 5**). We found the directionality and the effects of the lead SNPs to be consistent across the cohorts part of the meta-analysis (**Figure 9b and 9c**).

***Figure 1**. Motor severity GWAS Manhattan plot and lead variants forest plots.*

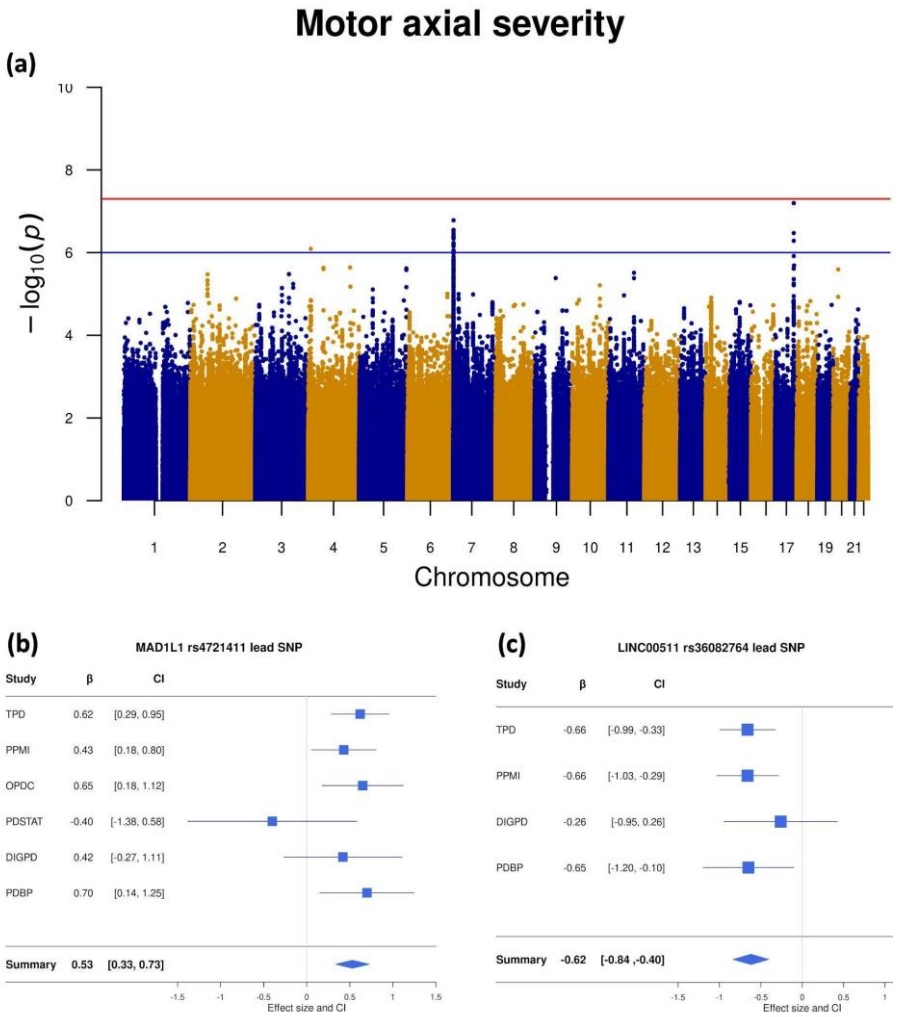
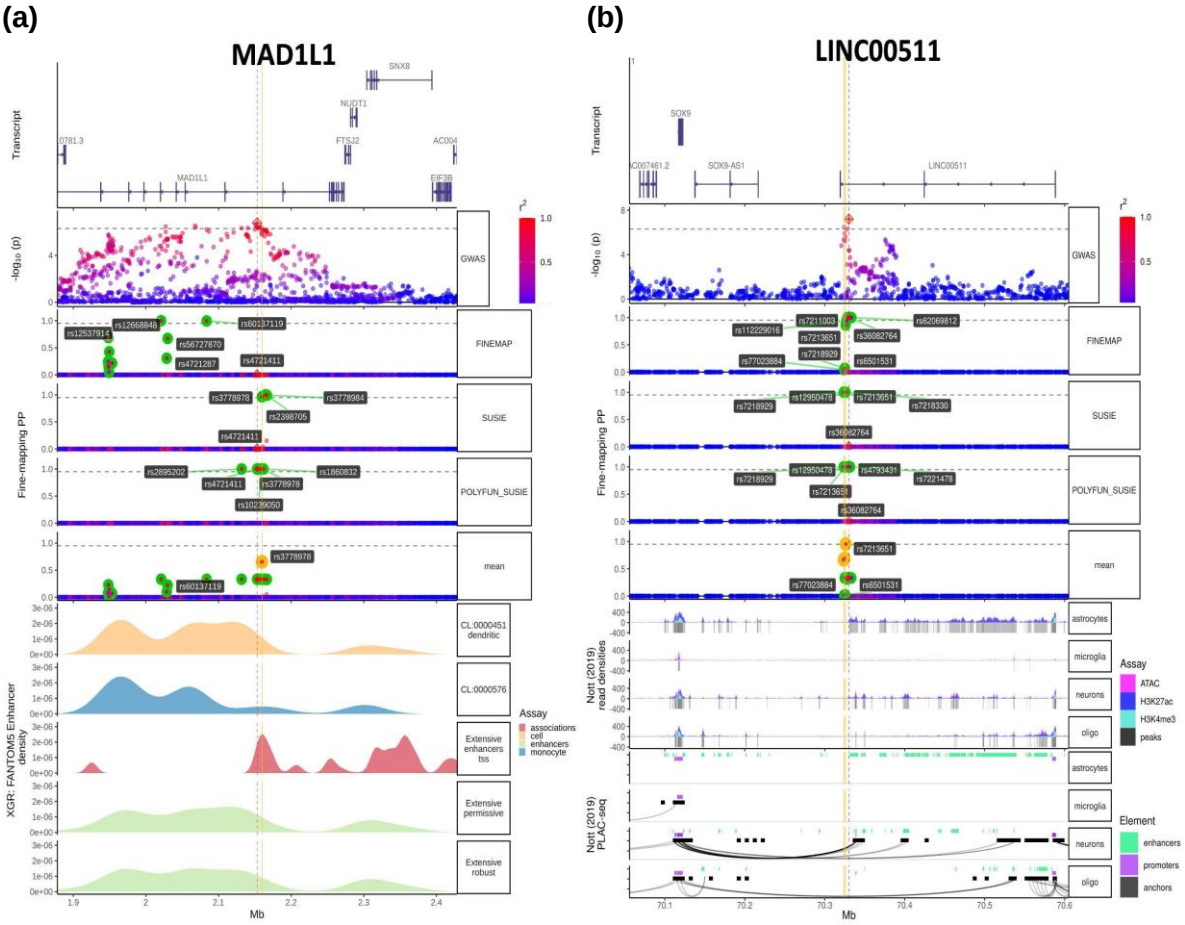


***Figure 9a**. Manhattan plot for the axial severity GWAS meta-analysis. The red dashed line indicates the genome-wide significance threshold P-value = 5e-8. The two LD blocks approaching genome wide significance are on the MAD1L1 locus in chromosome 7 and LINC00511 locus on chromosome 17. There was no genomic inflation (λ = 1.00). **Figure 9b**. Forest plots for lead variant rs4721411-T found at the MAD1L1 locus (right) under model A (I^2 = 0; Cochran's Q test: χ^2 = 4.01, df = 5, P =0.55) annotated by study, effect size, and the corresponding 95% confidence interval. **Figure 9c**. Forest plots for lead variant rs36082764-T found at LINC00511 locus under the GWAS meta-analysis using model A (I^2 = 0; Cochran's Q test: χ^2 = 1.07, df = 3, P = 0.78) annotated by study, effect size, and the corresponding 95% confidence interval.*

Subsequent fine-mapping at both loci identified rs3778978 in the *MAD1L1* locus as the causal SNP and a list of three SNPs (rs7213651, rs7218929, rs12950478) in the *LINC00511* locus as potential trait-causing SNPs. This fine-mapping effort narrowed down the spectrum of variants for further targeting in in vivo and in vitro analyses (**Table 7**). Notably, the *MAD1L1* fine-mapped causal variant and the lead SNP

overlapped with an active enhancer mark, suggesting an influence of the GWAS-nominated variants on the regulation of *MAD1L1* expression (**Figure 10a**).

***Figure 2**. MAD1L1 and LINC00511 functional annotation.*



*From top to bottom, transcripts plot, locus plot, the fine-mapping results, and the functional annotations specific assay we overlaid the GWAS locus with. In the locus plot, the SNPs are coloured in red as LD (given by R2) increases, and blue as the LD decreases. In the fine-mapping track, I highlight the SNPs with the highest posterior probabilities for each fine-mapping tool highlighted on the legend on the right hand side. In addition, I highlight in yellow the Consensus SNP. **Figure 10a.** We mapped the GWAS locus with FANTOM5 enhancer marks from the FANTOM project. All data from the FANTOM5 project was scanned to plot out the 5 datasets with present enhancer marks on the region of interest. From top to bottom, dendritic and monocytes cell type specific enhancer marks, bulk enhancer transcription start site interaction mark, and bulk enhancer permissive and robust marks. In the locus plot from the middle, from the extensive and enhancer enhancers marks, we can see how they overlap both the lead (red dashed vertical line), and the fine-mapped Consensus SNP (vertical yellow line). This overlap is also notable on the row of Enhancers-TSS interaction marks. **Figure 10b.** We mapped the GWAS locus with brain cell type specific regulatory element marks, the first 4 rows are the density marks (y-axis) from ATAC-seq assay (in pink), and CHIP-seq assays (H3K27ac in blue, and H3K4me3 in cyan), in astrocytes, microglia, neurons, and oligodendrocytes. The next four rows are the distal anchored chromatin loops (black curves) derived from the PLAC-seq assay.*

For LINC00511, an anchored chromatin loop was identified from the GWAS LD block in LINC00511 to a region containing the active promoter of neuronal *SOX9*, indicating that mutations in this distal regulatory region might alter *SOX9* expression specifically in neurons (**Figure 10b**).

I investigated eQTL databases through FUMA, discovering that both the lead variant and the fine-mapped nominated causal variant in *MAD1L1* were significant cis-eQTLs in BIOS and eQTLGen (**Table 9**). Subsequently, I conducted a colocalization analysis to assess the presence of a shared causal variant between the two traits **as detailed in Chapter 2 - Methods.** Despite an examination within a ±1 Mb range from the GWAS lead SNPs, no direct colocalization evidence was found for any gene. Unfortunately, there was no available cis-eQTL data for *SOX9*. In the *MAD1L1* locus, the Posterior Probability H3 (PPH3), indicating an association with both phenotypic and expression traits with distinct causal variants, reached the highest values (PPH3 in *MAD1L1*: eQTLGen = 0.97, MetaBrain = 0.98, PsychENCODE = 0.75).

Table 9*. cis-eQTL values of the Model A MAD1L1 locus lead SNP rs4721411.*

SNP	CHR	MAF	nearest Gene	gwasP	eQTL study	symbol	eqtlP
rs4721411	7	0.39	*MAD1L1*	1.657e-07	eQTLGen	*MAD1L1*	1.842e-57
rs4721411	7	0.39	*MAD1L1*	1.657e-07	BIOSl	*MAD1L1*	1.215e-12
rs3778978	7	0.38	*MAD1L1*	5.976e-07	eQTLGen	*MAD1L1*	1.265e-59
rs3778978	7	0.38	*MAD1L1*	5.976e-07	BIOS	*MAD1L1*	6.313e-12

gwasP = P-value of a SNP in the GWAS study; eqtlP =P-value of a SNP in the eQTL study.

Biological interpretation of nominated genes in relation to PD

ACP6 encodes Lysophosphatidic Acid Phosphatase Type 6, an enzyme that regulates lipid metabolism in mitochondria [299]. Changes in *ACP6* concentrations are found in Gaucher Disease (GD), although there is no clear link between *ACP6* levels in and GD progression. *ACP6* is highly expressed in astrocytes [300]. Mitochondrial dysfunction has been widely associated with PD aetiology [301].

MAD1L1 encodes the mitosis arrest deficient-like 1 protein, a component of the spindle-assembly checkpoint which prevents the onset of anaphase until chromosomes are aligned at the metaphase plate [302]. Recent GWAS have identified

MAD1L1 as a gene increasing the susceptibility for bipolar disorder and schizophrenia [303,304]. This variant is in high LD with the fine-mapping *MAD1L1* nominated variant (D' = 0.75) [305]. *MAD1L1* expression is measurable in several brain tissues [306]. A recent study investigated healthy adults carrying the *MAD1L1* rs11764590 risk allele [307]. Carriers showed alteration in the responsiveness and regulation of the mesolimbic reward system. Adults carrying the risk alleles showed significant hypoactivations of the ventral tegmental area (VTA), the bilateral striatum, and bilateral frontal and parietal cortices. Regarding PD in particular, a study including PD patients has shown that patients with more severe disease (measured in "OFF" and "ON " state), showed a fall in activation in the anterior cingulate cortex associated with reward expectancy [308]. A plausible explanation for this could be that *MAD1L1* PD mutation carriers, showing an impaired reward system, respond worse to dopaminergic therapy, hence developing with more severe axial signs.

It is known that enhancers are found in intronic and intergenic regions, as well as that introns act as gene regulators [309,310]. I have found evidence of the *MAD1L1* intron acting as an active enhancer and regulating and predicted to interact with a transcription start site (TSS). This together with the overlap found between eQTL and GWAS *MAD1L1* regional plots, suggest that this intron may play an active role in regulation in expression.

SOX9 is a SOX transcription factor (TFs) family member. The male sex determination gene (Sry) gave birth to this SOX family. SOX TFs regulate diverse cellular processes during development, as well as differentiation into tissues and organs. In addition, they play a major role in central nervous system development and adult neurogenesis[311]. Studies of *SOX9* gain and loss of function have demonstrated that *SOX9* is required for the formation of multipotent Neural stem cells (NSCs) and their maintenance in the central nervous system during embryonic and adult phase[312]. Moreover, *SOX9* regulates the transition from neurogenesis to gliogenesis during development, and it has been shown that when *SOX9* is not expressed, there was a reduction in astrocytes and oligodendrocytes, and a transient increase in motor neurons[313,314]. This is consistent with my findings suggesting when the distal regulation towards *SOX9* expression is altered in neurons, PD patients show a lower motor axial impairment, suggesting a connection between the CNS development and the adult neurogenesis.

d) Discussion

To understand the biology of motor progression in PD, we carried out a large well powered GWAS of PD motor progression. We have found one haplotype block at the *GJA5* locus that is significantly associated with axial PD motor progression. This association was consistent across individual cohorts included in my motor progression GWAS meta-analysis and was replicated in an analysis of H/Y supporting my findings. Further exploration of the GWAS significant signals in eQTL databases suggests that the GWAS hits may control the expression of *ACP6*, an enzyme that regulates lipid metabolism in mitochondria [299].

We used the MDS-UPDRS III (PD motor examination) scale, a sensitive measure of motor progression over time which has been widely studied in observational and interventional studies of PD. A study of untreated *de novo* PD patients in the PPMI study, followed up for 5 years to assess the progression of MDS-UPDRS, showed a linear increase of 2.4 points per year in MDS-UPDRS part III total score [179]. In this study, we observed a similar yearly rate of change for the total MDS-UPDRS score across the studies I included in my analysis (2.3 points/year on average) (**Table 4**). We have used linear mixed effect models to investigate the common genetic variability associated with the severity and progression of distinct PD motor aspects. This concept may be consistent with PD subtypes studies having a differential motor severity and progression [35,315–319]. Another aspect, of this differential approach to PD symptomatology is that limb and axial PD motor components may have a different cellular and pathophysiological basis, with axial and limb motor symptoms related to cholinergic and dopaminergic dysfunction respectively [320,321].

We corrected all models by AAO, and sex and PCs as confounding variables. We performed a fixed effects meta-analysis as opposed to a pooled analysis to further account for between cohorts heterogeneity, as cohorts we included had different inclusion and exclusion criteria, and were either genotyped with different microarrays or whole genome sequenced. My results are not confounded by levodopa response, as defined in my sensitivity analysis. In this dataset we have identified common genetic variability which determines axial, but not limb motor progression.

The lack of association between common genomic variation and the MDS-UPDRS limb subscale could be due to a combination of limited power and the levodopa effect in early disease. Evers and colleagues reported that measures of mobility, tremor, gait and posture, were consistent and reliable measures of PD progression [181]. Because these measures are well represented in the axial score (except for tremor), this may be better powered to assess progression. Moreover, the limb signs may be more sensitive to levodopa use than the axial signs, making it possible that true genetic associations with limb motor progression were masked. Lastly, we found the individual cohorts with the largest sample size had a higher axial rate of change compared to the limb rate of change (**Table 4**). A separate GWAS meta-analysis assessing the PD genetic contribution to the disease motor severity and subsequent functional annotation, identified *MAD1L1* and *SOX9* as candidate genes associated with PD axial motor severity. Nevertheless, these potential associations did not reach genome-wide significance and further analysis in distinct PD cohorts are needed for validation.

Strengths of my study include the large sample size, and the consistency of my results across cohorts and across different measures of axial motor progression. Potential limitations of my identification of *ACP6* as the relevant gene at the *GJA5* locus include the lack of colocalization between the phenotype and expression GWAS although these analyses are current limited by the sample size of eQTL datasets and the lack of cell specific gene expression data.

We hypothesise that expression of *ACP6* is important in the function in cell groups relevant to axial progression in PD including the pedunculopontine nucleus, and that therapies directed towards mitochondrial lipid metabolism may be relevant to the disease modification. Further replication, in independent cohorts genotyped in the global Parkinson's genetics program (GP2.org) will help to determine the importance of this region and further analysis of this biochemical pathway may provide new insights into the pathogenesis of PD progression.

4) Genetic meta-analysis of levodopa induced dyskinesia in Parkinson's disease

a) Introduction

The development of levodopa-induced dyskinesia (LiD) is a major clinical problem for PD patients and multiple pharmacological and neurosurgical approaches have been developed to try to prevent, attenuate or treat LiD. Dopamine is lost from the nigrostriatal pathway, which manifests as bradykinesia, muscular rigidity, rest tremor and postural instability [21,322]. There are several symptomatic treatments for PD motor symptoms, with the metabolic precursor of dopamine, levodopa, being the "gold standard" drug. Levodopa improves motor function as measured by UPDRS or the more recent MDS-UPDRS, widely used standard clinical assessments to evaluate the motor state in PD patients [291]. A comparison of an early levodopa treated group against a delayed treated group showed no difference in the rate of motor progression, suggesting that levodopa itself is not disease modifying or disease accelerating [323]. One of the major drawbacks of long-term levodopa treatment is that many PD patients experience levodopa-related motor complications, such as wearing off, dystonia and dyskinesia [324].

The prevalence of LiD varies across academic- and industry-led studies, averaging at around 20-40% after four years of levodopa treatment. There are two major LiD subtypes: peak-dose dyskinesia, which occur during the therapeutic window of levodopa treatment, and diphasic dyskinesia, which present at the start and end of a dose cycle [325].

Levodopa treatment is necessary for LiD development, but there are likely to be several other mediating factors [325]. Based on research in animal models, it is hypothesised that pulsatile delivery of oral levodopa, presynaptic nigrostriatal degeneration and intact striatal neurons are needed for the development of LiD [325]. Major risk factors for the development of LiD include young age at onset (AAO), female gender, low body weight, disease severity, disease duration and treatment duration (from the initiation of levodopa) as well as the total dose of levodopa [326,327]. Disease duration and treatment duration are closely related and delayed start study designs have evaluated the effect of delaying the initiation of levodopa, showing an

association between longer delay and a decreased risk of LiD [328]. There is increasing evidence that suggests genetics plays a role in the susceptibility to LiD. Rare variants in genes such as *PRKN*, *PINK1*, and *DJ-1* have been reported to be associated with higher rates of dyskinesia [329–331], although patients with autosomal recessive PD usually have early onset disease, which is in itself a risk factor for LiD. A study which corrected for age and disease duration variability did not replicate the findings of a higher LiD susceptibility among *PARK2* mutation carriers [332].

Common variation may also influence the risk of developing LiD. Variations at the DRD2, COMT, MAOA, BDNF, SLC6A3 and ADORA2A loci have all been reported to influence the risk of developing LiD [333–342]. Recently, an exome-wide association study of LiD in PD found that variants in *MAD2L2* and *MAP7* loci were associated with LiD, and replicated the association of the opioid receptor gene *OPRM1* [343]. Due to the high heterogeneity in the genetic determinants that regulate LiD, validation in large cohorts is needed.

Here, I investigated the genetic determinants of LiD by performing a meta-analysis of genome-wide survival to the development of LiD in five different cohorts, and assessed previously reported loci. I also performed functional genetic annotation to better understand the nominated loci. Lastly, I have investigated the predictive power of a PRS, and explored baseline clinical features that were significantly associated with the development of LiD in PD using a stepwise regression approach.

b) Methods

The source code with all materials and methods are available on GitHub (https://github.com/AMCalejandro/LID-CPH.git;https://doi.org/10.5281/zenodo.8139563). The README explains each step of the workflow to conduct the analysis and a link to each relevant pipeline or protocol.

i) Patients data and LiD definition

I accessed clinical and genetic data from the Tracking Parkinson's (TPD) [150], Oxford Parkinson's Disease Centre Discovery Cohort (OPDC) [344], Parkinson's Progression Markers Initiative (PPMI) [156], Parkinson's Disease Biomarkers Program (PDBP) [163], and simvastatin as a neuroprotective treatment for PD trial (PD-STAT) [153]

studies (**Table 10**). Each subject provided written informed consent for participation according to the Declaration of Helsinki and all cohort studies were approved by the relevant ethics committee.

***Table 10**. Study sample sizes and genotyping array.*

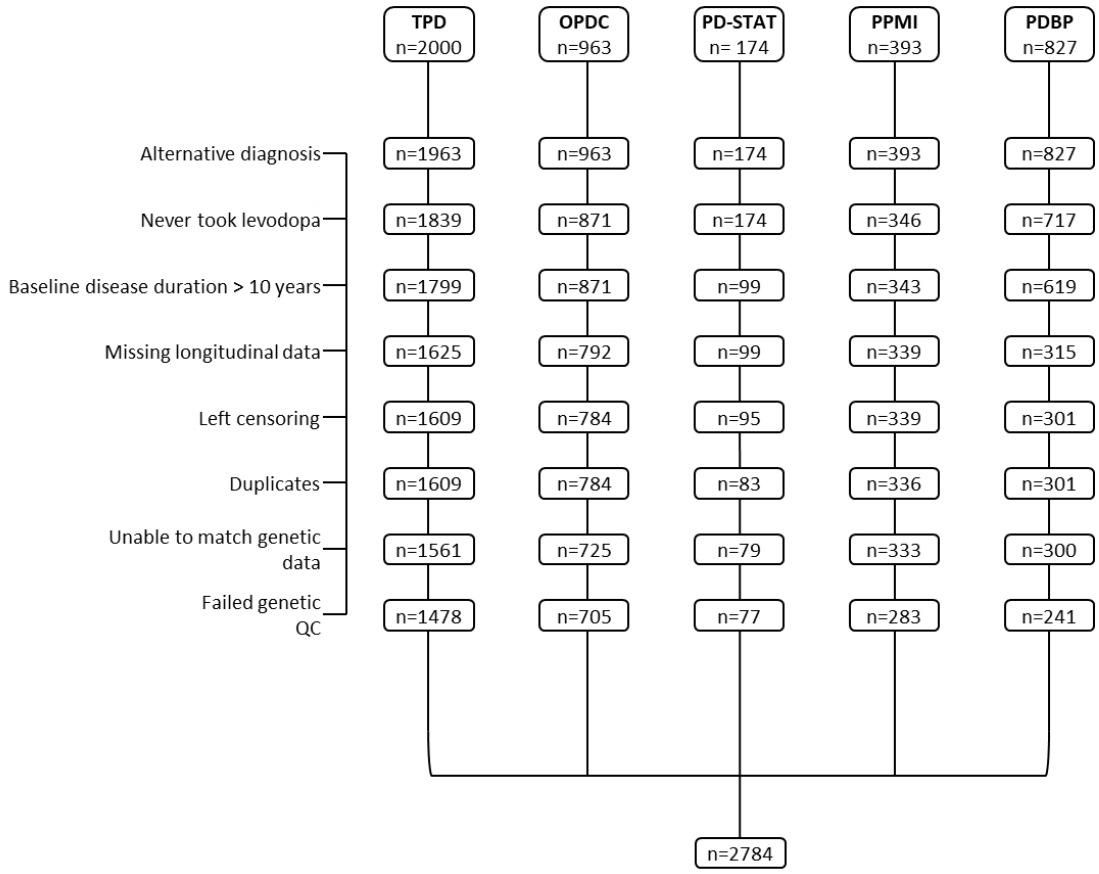
Study Name	Abbreviations	N	Genotyping array
Tracking Parkinson's Disease	TPD	2000	Illumina HumanCoreExome array
Oxford Parkinson's Disease Centre Discovery Cohort	OPDC	1082	Illumina HumanCoreExome-12 v1.1 or Illumina Infinium HumanCoreExome-24 v1.1
Parkinson's Progression Markers Initiative	PPMI	415	WGS
Advancing Parkinson's Disease Biomarkers Discovery	PDBP	873	WGS
Simvastatin as a neuroprotective treatment for Parkinson's disease	PD-STAT	174	Illumina Neurochip

WGS = Whole Genome Sequencing

I carried out clinical data QC on each cohort independently (**Figure 11**). Levodopa is necessary for PD patients to develop LiD [325], therefore I excluded those who were not exposed to levodopa. In addition, I removed patients who had a disease duration at study entry of more than 10 years from disease onset, patients without longitudinal data (patients with less than two clinical records available), and those with missing genotype data.

I defined PD patients as having dyskinesia if they reached an MDS-UPDRS item 4.1 score equal to or higher than 2 which is equivalent to a range of 26%-50% of the waking time with dyskinesia, and the first appearance of LiD was defined as the event time. Patients were excluded if they had dyskinesia at study entry, as time to the development of dyskinesia could not be established.

***Figure 11**. Quality control flowchart**.** We highlight the number of samples remaining after applying the multiple QC steps on each cohort we included in this study.*



ii) Genotype data quality control and imputation

To perform quality control (QC) at both the sample and genotype levels, I used PLINK v1.9 [290]. Each quality control step and the imputation approach was performed as described in **Chapter 2 – Methods**.

iii) Whole-genome sequencing data

The PDBP and PPMI cohorts included in this study were whole-genome sequenced using Illumina HiSeq X Ten Sequencer. More information can be found in https://ida.loni.usc.edu/login.jsp. WGS data was QC'ed using the same pipeline as the array-based data**.**

iv) Statistical analyses

I used the R programming language (version 4.3.0) to perform all the statistical analysis [345]. I studied the association between genome-wide genetic variants and time to develop dyskinesia from self-reported age at PD motor onset with Cox proportional hazard (CPH) regression models under a genetic additive model, using the 'survival' R (version 3.3-1). All tests were two-tailed. To investigate the power to detect an association under a Cox regression model with the current sample size, as well as to perform a simulation on the relationship between power and allele frequency (AF), SNP hazard ratios (HR), and sample size, I used the R package survSNP (version 0.25).

I ran time-to-LiD GWAS in each cohort separately, adjusting by AAO (or AAD in the cohorts where AAO was not available), gender, and first 5 PCs, using as my outcome the midpoint between the visit the threshold was met and the previous time point. Multiple studies indicate that the risk of dyskinesia relates to disease severity. To improve the power to detect a genetic association, I explored the goodness-of-fit of the model in each cohort independently after adding the following baseline covariates, which provide surrogate measures of disease severity and dopaminergic denervation at baseline: levodopa or LEDD dose, disease duration from onset to baseline assessment and baseline motor score as measured by MDS-UPDRS part III. For each cohort, I selected the model which provided the most accurate prediction of LiD based on the Akaike Information Criteria (AIC). I used the resulting model as the main model in my analysis. I summarised the nominated set of covariates in each cohort (**Table 11**). I verified that the proportional hazards assumption held true by assessing the independence between scaled Schoenfeld residuals and time through the cox.zph function from the 'survival' package. Schoenfeld residuals are obtained by subtracting the individuals' covariate values at the time "t" and the corresponding risk-weighted average of covariates among all those that are at risk at the time "t". Then, they are scaled by performing a variance-weighted transformation. A non-significant relationship between the scaled residuals and time reveals proportionality of the hazards in the model.

***Table 11**. List of covariates added on both the basic and adjusted model across cohorts*

Study Name	Covariates in basic model	Covariates in the adjusted model
Tracking Parkinson's Disease	AAO, GENDER, 5 PCs	AAO, GENDER, 5 PCs, BASELINE DISEASE DURATION, BASELINE MDS-UPDRS-III total, BASELINE L-DOPA DOSE
Oxford Parkinson's Disease Centre Discovery Cohort	AAO, GENDER, 5 PCs	AAO, GENDER, 5 PCs, BASELINE MDS-UPDRS-III total, BASELINE LEDD
Parkinson's Progression Markers Initiative	AAO, GENDER, 5 PCs	AAO, GENDER, 5 PCs, BASELINE MDS-UPDRS-III total, BASELINE DISEASE DURATION
Advancing Parkinson's Disease Biomarkers Discovery	AAD, GENDER, 5 PCs	AAD, GENDER, 5 PCs, BASELINE MDS-UPDRS III total, BASELINE DISEASE DURATION
Simvastatin as a neuroprotective treatment for Parkinson's disease	AAO, GENDER, 5 PCs	AAO, GENDER, 5 PCs

I performed a meta-analysis using METAL as described in **Chapter 2 - Methods**. I applied a post meta-analysis QC step to remove genetic variants that were present in less than 3 out of 5 cohorts, with less than 1000 variants, as well as variants with high minor allele frequency (MAF) heterogeneity across the cohorts (MAF > 0.15). In addition, I accounted for high heterogeneous variants by removing those with a significant Cochran's Q test as well as those with an I2 index higher than 80%.

Statistical significance was assessed at the conservative threshold of $P = 5 \times 10^{-8}$, derived from a Bonferroni correction accounting for the number of independent tests and the LD structure of the genome [346].

I proved that the model met the proportional hazard assumption after including significant SNPs using the cox.zph function from the 'survival' package. I evaluated whether signals were replicated across different cohorts with the R package 'forestplot' (version 2.0.1).

v) Sensitivity analyses

To validate the genome-wide significance findings, I performed four sensitivity analyses to assess if the best model described above led to an unbiased testing of the null hypothesis of no association between all genome-wide SNPs and time-to-LiD. The first sensitivity analysis was designed to compare the basic and adjusted models. I tested whether high deviations in the SNP estimates and P-values arose after accounting for disease severity and dopaminergic denervation at baseline by measuring the correlation between the basic and adjusted GWAS meta-analyses. Next, I performed two separate sensitivity analyses to test whether either levodopa dose or the PD motor severity (as measured by MDS-UPDRS part III) at the time point where LiD were first documented, were confounding my findings. I performed this sensitivity analysis in Tracking Parkinson's, the largest dataset. I performed a CPH GWAS on the Tracking Parkinson's cohort adjusting by: a) known confounders, b) known confounders + motor severity (as measured by MDS-UPDRS part III) c) known confounders + levodopa dose. I compared the SNP metrics from the three models for the lead SNPs on the loci that reached genome-wide significance on the time-to-LiD GWAS meta-analysis. Lastly, because the PDBP cohort did not have age at onset available and I used age at diagnosis (AAD) in the CPH model, I reran the time-to-LiD GWAS meta-analysis excluding PDBP to confirm that this cohort was not inflating the SNP test-statistics.

vi) Post-GWAS analyses

I performed fine-mapping **as described in Chapter 2 – Methods** to nominate causal variants at each locus that reached genome-wide significance. To evaluate the potential effect of SNPs on candidate loci on the control of gene expression I also used echolocatoR to map GWAS nominated loci with epigenetic marks form the brain cell type -specific marks by Nott and colleagues, and Uniformed transcription factor binding sites from ENCODE.

To investigate whether there were several independently associated SNPs at each GWAS nominated locus, I performed a conditional and stepwise selection procedure with GCTA-COJO (version 1.93.0 beta for Linux) [237]. I used the Accelerating Medicines Partnership: Parkinson’s Disease (AMP-PD, v.2.5) data [347] (*n* = 10,418)

as the reference panel to estimate the correlation between SNPs. The reference sample was subjected to the same QC steps as described above, needed to get unbiased LD estimates [264].

I used the 'coloc' R package (version 5.1.0) to perform colocalization analysis between loci significantly associated with progression to LiD and SNPs defining gene expression in the region. I used cis-eQTL data from MetaBrain cortex tissue [270] (N = 6,601 individuals) and blood cis-eQTLs from eQTLGen (N = 31,684) [269]. The strategy I followed to perform colocalization is explained in more detail in **Chapter 2 - Methods**.

I used FUMA (version 1.3.8) to further characterise the nominated loci by querying GWAS Catalogue to retrieve uncharacterised GWAS loci SNPs in my meta-analysis and to get positional mapping information based on MAGMA [115]. I used a threshold of $P < 1e\text{-}6$ to nominate tag SNPs. Additional SNPs that were in high LD with tag SNPs were inferred using European samples 1Kg Phase3 reference panel (with $r2 > 0.6$ and independent from each other with $r2 < 0.6$).

vii) Candidate gene analysis

In order to validate variants that have been reported in previous studies to be associated with time-to-LiD or LiD risk, I accessed the LiDPD website (Date accessed: 12/01/2023) and downloaded a list of curated variants from the literature. I explored these in my time-to-LiD GWAS meta-analysis [348].

viii) LiD prediction modelling

I used PRSice software (version 2) to compute a polygenic risk score (PRS) [349]. I used the summary statistics of my time-to-LiD meta-analysis as base data and the Tracking Parkinson's cohort as target data. I chose the Tracking Parkinson's cohort as it is the single largest cohort, which reduces the SE of the PRS estimates, leading to more confident estimates. I then replicated the association of the nominated SNPs composing the PRS in the second largest cohort I had access to, OPDC, resembling a discovery / replication study design, although in this case the OPDC data had contributed to the LiD PRS. Further details on how we ran PRS for my analyses is available in **Chapter 2 – Methods**.

Next, I used a stepwise logistic regression model with a custom script using the ‘stats’ R base package (version 4.2.2) to find whether any baseline clinical variable was significantly associated with LiD status. I used data from the Tracking Parkinson's cohort, as it is deeply phenotypically characterised (number of baseline covariates = 702). After removing variables with high missingness rate (missing rate > 10%) or categorical variables with only one level, I defined a total of 502 baseline features (including the PRS). Then, I created a base logistic regression model (adjusted for sex the first 5 PCs and standardised AAO). At each step of the stepwise regression approach, I refitted the base model with each of the baseline predictors individually, and selected the model with the variable that decreased AIC the most. I ran the model until no variable further decreased the AIC, or until the AIC score was equal to 1. Once the model was fitted, I selected only those predictors that were significantly associated with the binary outcome, applying the conservative Bonferroni correction accounting for the number of predictors assessed. I set the significance threshold as 0.05 / 502 = 1e-4. To account for class imbalance in the evaluation of classifiers, I computed precision recall curves using the ‘PRROC’ R package (version 1.3.1)

c) Results

i) Cohort clinical features and prevalence

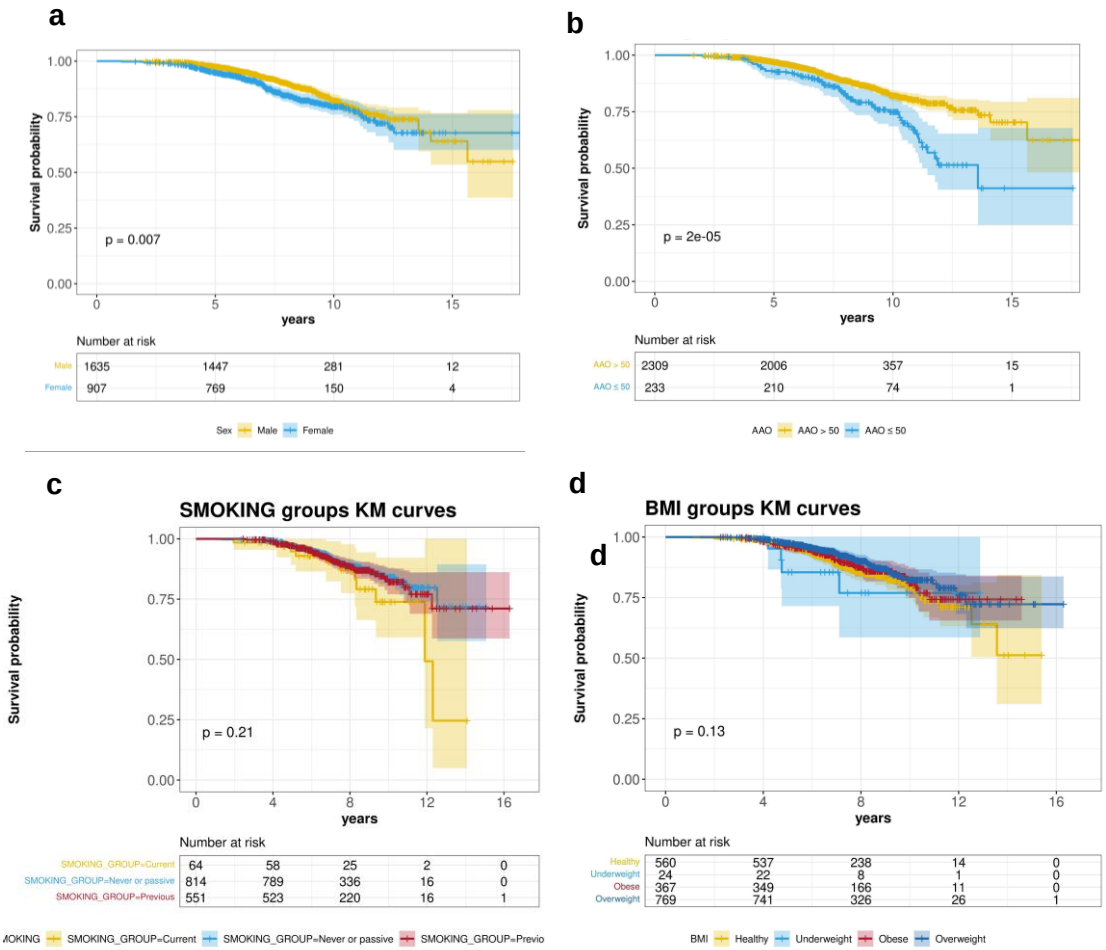
Across all cohorts (n= 2,784 PD patients), the incidence of LiD was 14% (**Table 12**), except in the PPMI cohort where it was 21%. This is consistent with the effect of age at onset on LiD [350–352], given that PPMI is a *de novo* study that recruited younger patients. I did not exclude any patient from the PPMI cohort due to left-censoring.

Table 12. *Cohorts summary statistics.*

COHORT	PD patients Post-QC (n)	Follow up, years	No(%) LiD	No(%) left-censored	No(%) male	Time to midpoint event (mean ± sd)	AAO, years (mean ±sd)	AAB, years (mean ± sd)	Disease duration at baseline from onset, years (mean ± sd)	MDS-UPDRS part III at baseline (mean ± sd)	Levodopa dose at baseline (mean ±sd)
TPD	1478	7.5	177 (12)	16 (1)	945 (64.3)	7.47 (2.2)	64.43 (9.16)	67.29 (9)	2.86 (1.6)	22.36 (11.7)	217 (197)
OPDC	705	9.0	92 (13)	8 (0.8)	451 (64)	7.87 (2.9)	64.35 (9.47)	67.21 (9.3)	2.85 (1.7)	26.27 (10.8)	280 (205)
PPMI	283	9.0	82 (21)	0 (0)	259 (66)	8.28 (2.3)	60.16 (9.93)	62.08 (9.8)	1.92 (1.3)	21.38 (9.1)	0 (0)
PD STAT	77	2.0	10 (13)	4 (4.9)	48 (62)	8.77 (2.8)	57.23 (8.7)	64.84 (9.2)	7.61 (1.7)	28.86 (11.6)	NA
PDBP	241	5.0	33 (14)	16 (6)	149 (62)	5.93 (2.7)	NA	64.58 (9.3)	2.85 (2.5)	20.9 (11.1)	414 (207)

I explored the effect of demographic and clinical factors previously reported to be associated with LiD. I merged baseline clinical data from all the cohorts. I found that patients with younger PD AAO (grouped as people with age at onset higher than 50 years and lower or equal than 50 years), had a higher probability of developing LiD than older patients along the time interval from disease onset to study end (HR = 1.8, SE = 0.14, *P* = 2e-5) (data excluding PDBP as AAO was not available). Female PD patients showed a consistent increase in the probability of developing LiD during a 12.5 years' time interval (**Figure 12 a and b**). Body mass index (BMI) was available in PPMI and Tracking Parkinson's, and smoking status data was available in the Tracking Parkinson's cohort only. I did not find a significant increase in the probability of developing dyskinesia either for PD patients with low baseline BMI nor for PD smokers at baseline (**Figure 12 c and d**).

***Figure 12**. LiD risk factors Kaplan-Meier curves.*



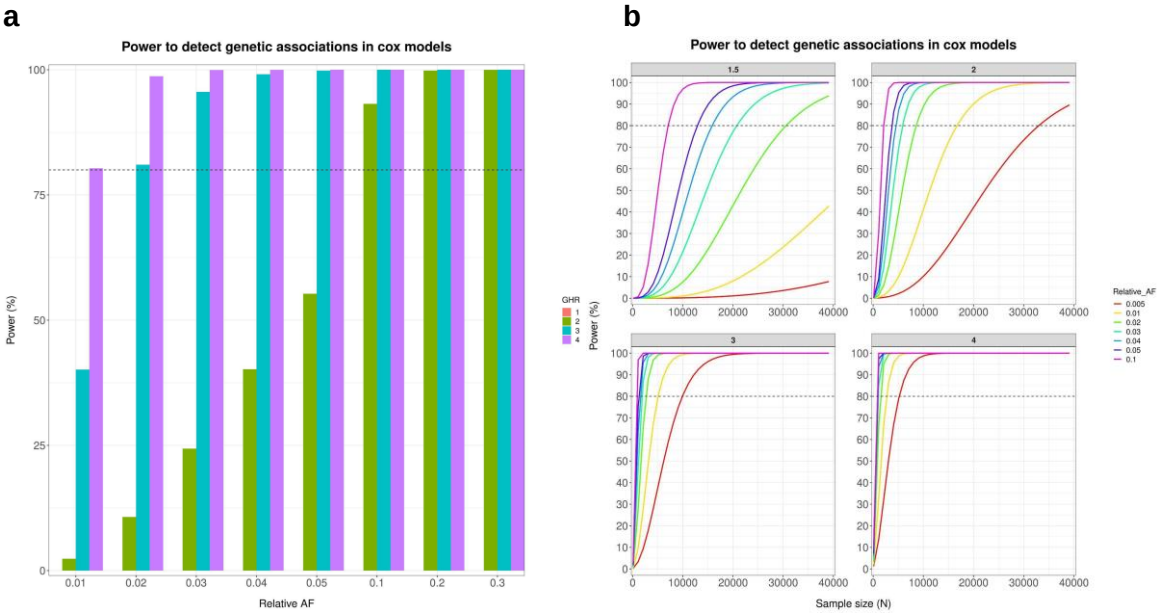
Kaplan-Meier curve for Survival probability (LID free probability) based on gender (A), age at onset (AAO) (B), smoking status (C), and smoking status baseline body mass

index (BMI). The P-value (P) showing the significance of differences on the survival probability is given on each plot. Number at risk represents the number of PD patients remaining on the study at the different time points (0, 5, 10, 15 years). The colour expansion on each curve represents the confidence interval (CI).

ii) Power analysis

I performed a power analysis to estimate the power to find a genetic association between time-to-LiD and genome-wide SNPs with the current sample size and LiD event rate, and to evaluate how this varied with a range of genotype hazard ratios (GHRs) and AFs. I was well-powered (80% power) to detect genetic variants associated with the development of LiD with a HR equal or higher than 2 and a MAF as low as 0.01 (**Figure 13a**). In addition, I performed a simulation to show as the sample size increases, the power to detect rarer associations improves. As I increased the simulated sample size to 18000, I achieved 80% power for genetic variants with a MAF lower than 0.01, and with a HR lower than 2 (**Figure 13b**).

***Figure 13**. Power calculation and simulation.*

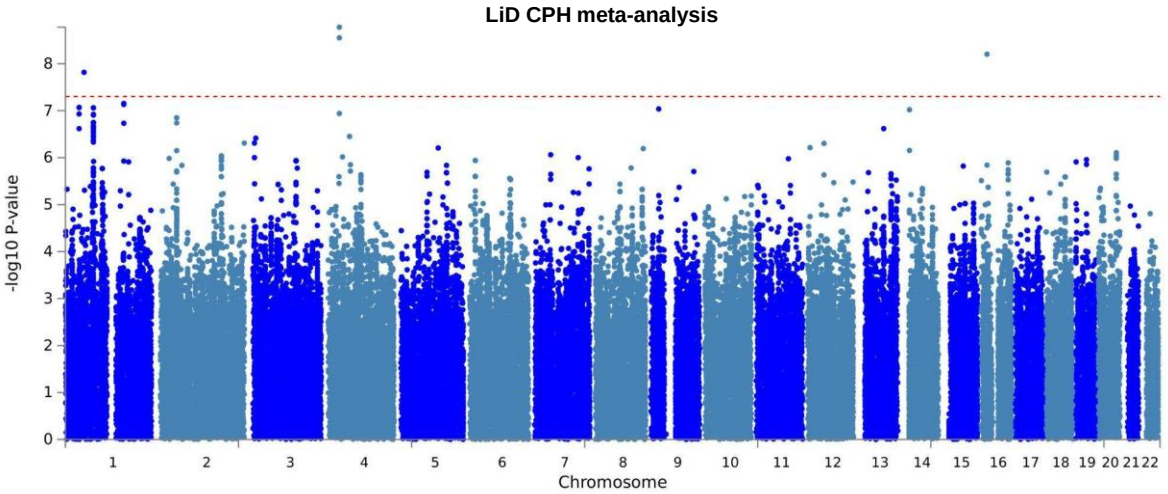


*Power calculation and simulation to detect genetic association with time to develop LiD as a function of sample size, relative allele frequency (AF), and genetic hazard ratio (GHR). **Figure 13a**. Power calculation (y-axis) for the current sample size based on different AFs (x-axis) (graph label) ; **Figure 13b**. Power simulation to explore the increase in power (y-axis) to detect lower GHR (graph grid) and relative AFs (graph label) as I increase the sample size (x-axis).*

iii) Time-to-LiD GWAS

I ran time-to-LiD GWAS independently for each cohort, using the first appearance of LiD as the outcome. I meta-analysed results controlling for genomic inflation. In addition, I estimated genomic inflation on the time-to-LiD meta-analysis and did not find significant genomic inflation (λ = 1.02). I identified three loci significantly associated with time-to-LiD onset in the meta-analysis of the adjusted model on chromosome 1 , chromosome 16 and chromosome 4 (**Figure 14**).

Figure 14. LiD CPH GWAS meta-analysis.



The meta-analysis was conducted using a Cox proportional hazards model in each cohort separately, and results were meta-analysed. Genome-wide significance was set at 5e-8 and is indicated by the red dashed line.

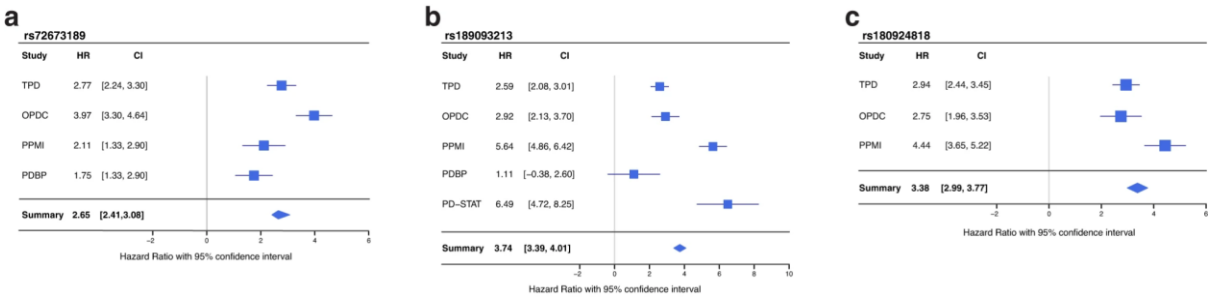
The most significant SNPs at each loci were rs72673189, rs189093213, rs180924818. **rs72673189** (HR = 2.77 , SE = 0.18 , *P* = 1.53e-8) in chromosome 1, is a variant in the third intron of the *LRP8* gene. **rs189093213** (HR = 3.06, SE = 0.19, *P* = 2.81e-9) in chromosome 4 was found in the non-coding RNA *LINC02353 (PCDH7 1.2Mb downstream)*. **rs180924818** (HR = 3,13, SE = 0.20 , *P* = 6.27e-9) in chromosome 16 was found very close (0.15Mb upstream) to the 3'-UTR of the *XYLT1* protein coding gene in a non-coding region of the genome (**Table 13**).

***Table 13**. Independent significant SNPs with a P-value lower than 1e-7.*

CHR	BP	SNP	MAF	BETA	HR	SE	SNP P-value in the Adjusted model	SNP P-value in the Basic model	Number of SNPs	Nearest gene	Type of variant
4	32435284	rs189093213	0.02	1.12	3.06	0.19	1.673e-09	6.15e-08	3	*LINC02353*	ncRNA intergenic
16	17044975	rs180924818	0.03	1.14	3.13	0.2	6.265e-09	8.20e-08	3	*XYLT1*	intergenic
1	53778300	rs72673189	0.03	1.02	2.77	0.18	1.527e-08	2.65e-08	2	*LRP8*	intronic
1	168645690	rs79432789	0.05	0.77	2.16	0.14	7.037e-08	2.47e-06	4	*DPT*	intergenic
1	39646765	rs71642678	0.01	1.61	5	0.3	8.555e-08	1.89e-07	12	*MACF1*	intronic
1	80950480	rs12133858	0.04	0.76	2.14	0.14	8.692e-08	1.01e-06	48	*RP11-115A15*	intergenic
9	22664277	rs77115593	0.02	1.26	3.52	0.24	9.192e-08	4.37e-07	1	*LINC02551*	ncRNA intronic
14	22020490	rs139943801	0.03	1	2.72	0.19	9.522e-08	2.63e-07	1	*RBBP4P5*	intergenic

The direction of the effects was consistent across the meta-analysed cohorts in which the SNPs were present (**Figure 15**).

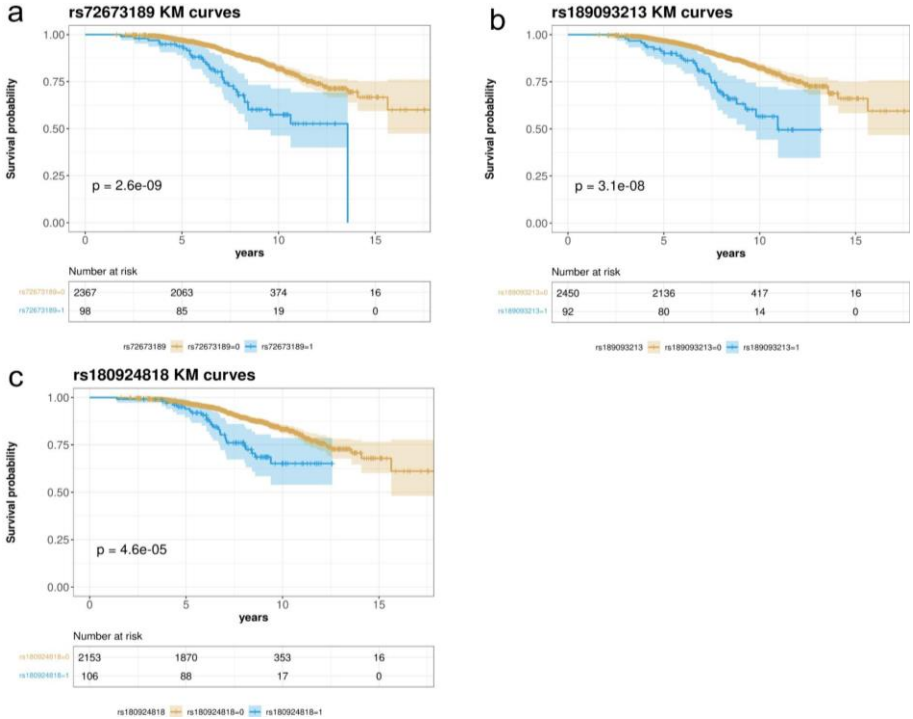
***Figure 15**. Forest plots of lead genetic associated variants.*



***a**, LRP8 rs72673189 variant (I^2=0; Q:χ^2=0.24, df=3, P=1.53e-08). **b**, LINC02353 rs189093213 variant (I^2=21.4; Q:χ^2=5.09, df=4, P=1.67e-09). **c**, XYLT1 rs180924818 variant (I^2=0; χ^2=0.77, df=2, P=6.27e-09). I^2 = I^2 Index of heterogeneity, HR = Hazard ratio, P = P-value, Q = Cochran's Q test of heterogeneity, df = degrees of freedom.*

To visually represent the survival probability of patients carrying the lead SNP on each locus I found in my meta-analysis, I extracted each patient's genotypes and showed the difference in the probability of LiD between carriers and non-carriers through Kaplan-Meier curves (**Figure 16**).

Figure 16**. Survival curves of candidate SNPs **.



***a**, Kaplan-Meier curve for Survival probability (LiD free probability) based on rs72673189 carrier status in PD patients. **b**, Kaplan-Meier curve for Survival probability (LiD free probability) based on rs189093213 carrier status in PD patients. **c**, Kaplan-*

Meier curve for Survival probability (LiD free probability) based on rs180924818 carrier status in PD patients. The blue curve represents genetic variant carriers, whereas the yellow curve represents non-carriers. p = p-value. Number at risk represents the number of PD patients remaining on the study at the different time points (0, 5, 10, 15 years). The colour expansion on each curve represents the confidence interval (CI).

iv) Sensitivity analysis

The three variants found to significantly increase LiD susceptibility in the adjusted model approach remained associated in the basic model including only known confounders (**Table 14**). I found the correlation of the SNP metrics between the basic and the adjusted model to be high (SNP P-values Pearson correlation coefficient = 0.87; P-value < 2e-16) (SNP Effect size Pearson correlation coefficient = 0.97; P-value < 2e-16). This indicated that adding additional predictors based on baseline variation increased the power to detect SNP-outcome associations, presumably by explaining other sources of variance in the model, and that there was no source of confounding given by disease duration and severity measures (suggested by the high correlation in the SNP metrics).

***Table 14**. Sensitivity analyses lead SNP P-values in the basic CPH model for the TPD cohort*

CHR	BP	SNP	A1	A2	Basic model P-value	Levodopa model P-value	MDS-UPDRS III model P-value
1	53778300	rs72673189	A	G	1.96E-04	2.39E-04	2.50E-04
4	32435284	rs189093213	A	G	6.32E-03	1.89E-03	3.49E-03
16	17044975	rs180924818	G	A	1.69E-04	2.62E-04	1.21E-04

Using data from Tracking Parkinson's only, I investigated whether these associations could be confounded by levodopa dose or the disease stage at the LiD event time point. For each of the genome-wide significant SNPs, I repeated the CPH analysis adjusting for levodopa dose or disease stage as measured by MDS-UPDRS part III at the first visit when the LiD threshold was reached or at the last available visit for patients who did not develop LiD during the study length. I did not find a change either in the hazard ratio or the test-statistics that could suggest an unaccounted source of confounding. Finally, excluding PDBP from the meta-analysis did not significantly change the lead SNP's hazard ratio and significance levels (**Table 15**).

***Table 15**. Lead SNP P-values in the CPH model including and excluding PDBP cohort in the basic and adjusted models.*

CHR:POS	SNP	A1	A2	MAF	HR	SE	P-value	N	PDBP	MODEL
4:32435284	rs189093213	A	G	0.02	3.08	0.19	1.673e-09	2687	YES	ADJUSTED
4:32435284	rs189093213	A	G	0.02	2.73	0.18	6.154e-08	2784	YES	BASIC
4:32435284	rs189093213	A	G	0.02	3.29	0.19	6.24e-10	2446	NO	ADJUSTED
4:32435284	rs189093213	A	G	0.02	2.88	0.19	2.989e-08	2543	NO	BASIC
16:17044975	rs180924818	A	G	0.98	0.32	0.20	6.265e-09	2687	YES	ADJUSTED
16:17044975	rs180924818	A	G	0.98	0.35	0.19	8.197e-08	2784	YES	BASIC
16:17044975	rs180924818	A	G	0.98	0.32	0.20	6.265e-09	2446	NO	ADJUSTED
16:17044975	rs180924818	A	G	0.98	0.35	0.19	8.197e-08	2543	NO	BASIC
1:53778300	rs72673189	A	G	0.02	2.76	0.18	1.527e-08	2610	YES	ADJUSTED
1:53778300	rs72673189	A	G	0.02	2.72	0.18	2.654e-08	2707	YES	BASIC
1:53778300	rs72673189	A	G	0.02	2.93	0.19	1.505e-08	2369	NO	ADJUSTED
1:53778300	rs72673189	A	G	0.02	2.83	0.19	4.214e-08	2466	NO	BASIC

v) Functional annotation

I performed fine-mapping using ABF, SuSiE, FINEMAP, and Polyfun-SuSiE, and found Consensus SNPs on each CPH GWAS nominated loci (**Table 16**).

***Table 16**. List of fine-mapped consensus SNPs on each locus.*

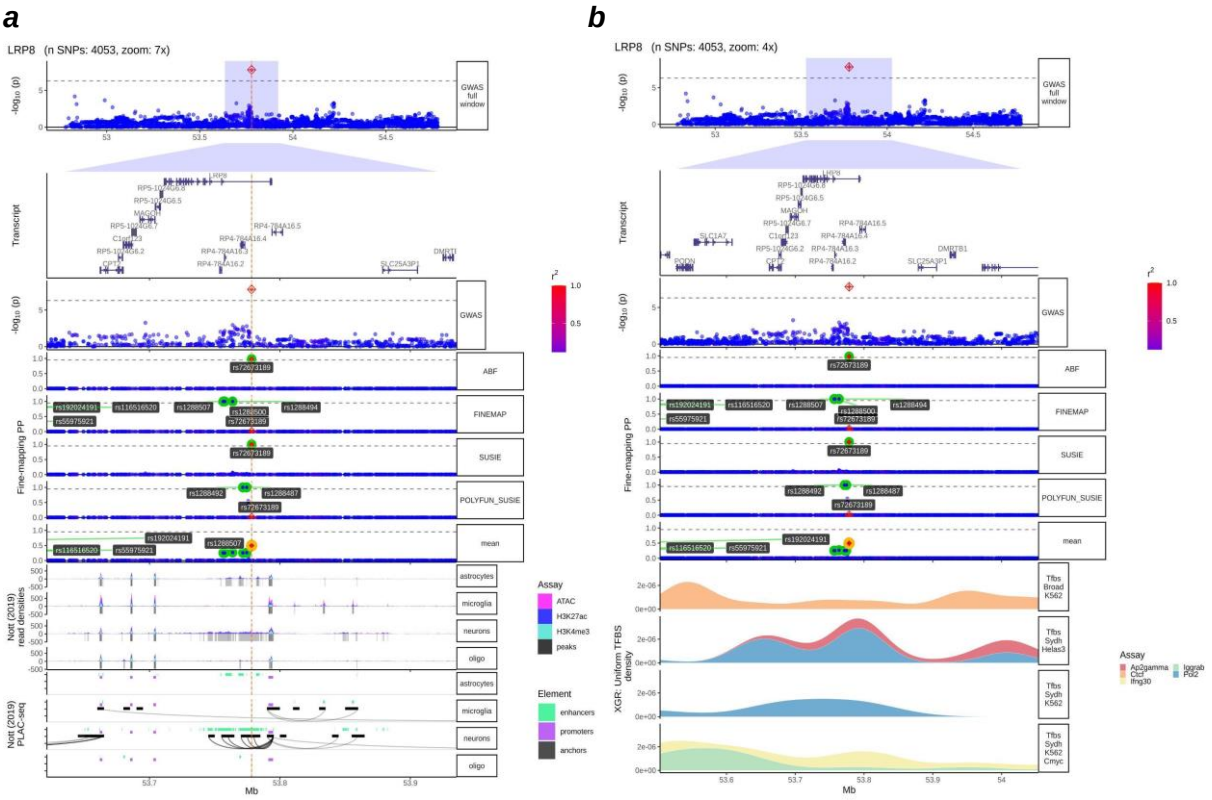
SNP	Locus	CHR	P	Effect	SE	A1	leadSNP	ABF	FINEMAP	SUSIE	POLYFUN_SUSIE	Sup
rs72673189	*LRP8*	1	1.5e-08	1.01	0.18	A	TRUE	1	0	1	0	2
rs180924818	*XYLT1*	16	6.2e-09	-1.14	0.20	A	TRUE	1	1	1	1	4
rs137895239	*XYLT1*	16	3.1e-05	0.88	0.21	A	FALSE	0	1	1	1	3
rs142441980	*XYLT1*	16	1.4e-06	0.88	0.18	A	FALSE	0.01	1	1	1	3
rs17207399	*XYLT1*	16	2e-04	-0.47	0.13	C	FALSE	1	1	1	1	3
rs189093213	*LINC02353*	4	1.7e-09	1.12	0.19	A	TRUE	0.61	0.96	1	1	3
rs10023843	*LINC02353*	4	0.1	0.55	0.36	T	FALSE	0	0	1	1	2
rs139511855	*LINC02353*	4	4.5e-05	-1.09	0.27	A	FALSE	0	0	1	1	2
rs147573196	*LINC02353*	4	2.5e-06	1.20	0.25	A	FALSE	0	0	1	1	2

SNP	Locus	CHR	P	Effect	SE	A1	leadSNP	ABF	FINEMAP	SUSIE	POLYFUN_SUSIE	Sup
rs28858724	*LINC02353*	4	0.03	-0.34	0.17	A	FALSE	0	0	1	1	2

leadSNP: Whether a given SNP is the locus lead SNP.
<tool>.CS: The posterior probability that a SNP is casual of the LiD phenotype.
Support: The number of fine-mapping tools that nominated the Consensus SNP
mean.PP: The mean SNP wise PP across fine mapping tools
mean.CS: If mean PP is greater than the 95% probability threshold (mean.PP > 0.95), then mean.CS is 1, else 0.

I found the lead SNP at each locus to be Consensus SNPs, which are those selected by at least two different fine-mapping tools. I plotted each locus found to have at least one variant significantly associated with time to reach LiD against brain cell type-specific epigenomic data. I found that the lead (and fine-mapped SNP) at the *LRP8* locus belonged to a neuronal specific chromatin accessible region, which is a target region for DNA-associated proteins, as measured with the ATAC-seq and CHIP-seq (H3K27ac and H3K4me3) assays (**Figure 17a**). I also found this SNP to be part of a neuronal specific enhancer-promoter interaction within *LRP8,* as defined by PLAC-seq (**Figure 17a**). This implies that this specific *LRP8* intronic signal is an active neuronal enhancer of the *LRP8* expression, forming an anchored chromatin loop recruiting the transcription machinery to the *LRP8* transcription start site (TSS). In addition, I found suggestive evidence that the lead SNP lies in a transcription factor binding site (TFBS), as defined by the ENCODE project (**Figure 17b**).

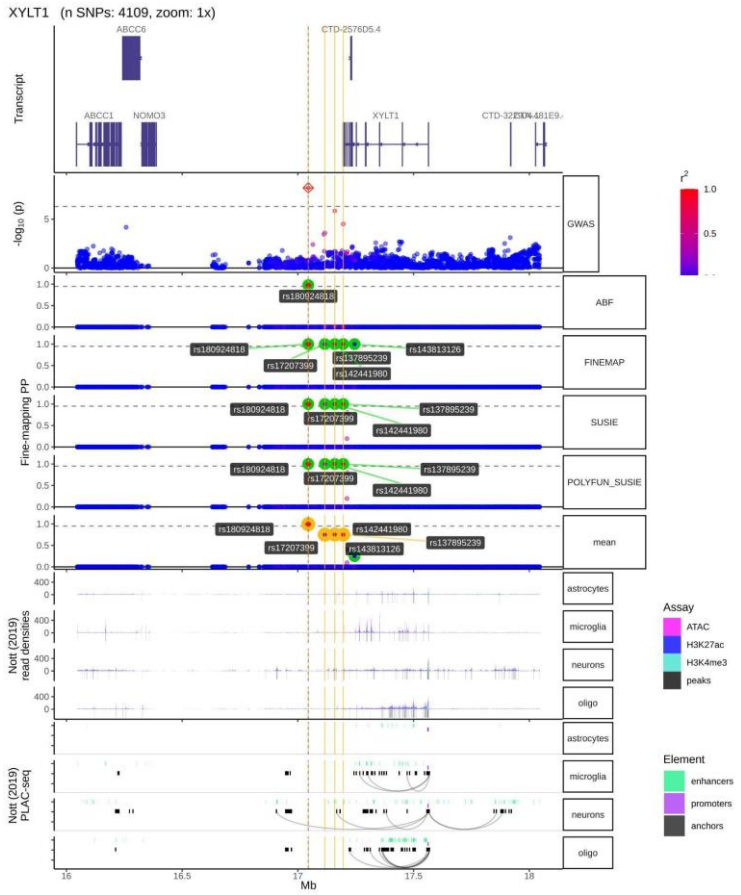
***Figure 17**. LRP8 functional annotation.*



From top to bottom, transcripts plot, locus plot, the fine-mapping results, and the functional annotations specific assay we overlaid the GWAS locus with. In the locus plot, the SNPs are coloured in red as LD (given by R2) increases, and blue as the LD decreases. In the fine-mapping track, I highlight the SNPs with the highest posterior probabilities for each fine-mapping tool highlighted on the legend on the right hand side. In addition, I highlight in yellow the Consensus SNP. ***Figure 17a.*** *The last track contains cell type specific regulatory element marks, the first 4 rows are the density marks (y-axis) from ATAC-seq assay (in pink), and CHIP-seq assays (H3K27ac in blue, and H3K4me3 in cyan), in astrocytes, microglia, neurons, and oligodendrocytes. The next four rows are the distal anchored chromatin loops (black curves). I see how, only in neurons, there is a chromatin loop forming from the LRP8 GWS and the fine-mapped consensus variant towards the LRP8 promoter (purple).* ***Figure 17b****. The rows in the last track show the transcription factor binding sites (TFBS) densities (y-axis) measured on different cell lines and laboratories. XGR finds the top 5 transcription factors (TF) with the highest binding activity in the track genomic window. These top 5 TF are displayed in the Assay label.*

Similarly, I found that some of the fine-mapped SNPs (including the lead SNP) in the *XYLT1* locus were forming chromatin loops towards the *XYLT1* promoter, as measured by the PLAC-seq assay, suggesting that regulation of this gene associated with susceptibility to LiD (**Figure 18**). I did not find any functional regulatory marks at the *LINC02353* locus.

***Figure 18**. XYLT1 locus fine-mapping and brain cell type specific regulatory marks.*



From top to bottom, transcript plot, locus plot, the fine-mapping nominated variants across fine-mapping tools, brain cell type specific regulatory element marks. In the locus plot, the SNPs are coloured in red as LD (given by R2) increases, and blue as the LD decreases. In the fine-mapping track, I highlight the SNPs with the highest posterior probabilities for each fine-mapping tool (ABF, FINEMAP, SUSIE, POLYFUN_SUSIE). In addition, I highlight in yellow the Consensus SNP with the highest mean Posterior Probability (mean). In the cell type specific regulatory element marks, the first 4 rows are the density marks (y-axis) from ATAC-seq assay (in pink), and CHIP-seq assays (H3K27ac in blue, and H3K4me3 in cyan), in astrocytes, microglia, neurons, and oligodendrocytes. The next four rows are the distal anchored chromatin loops (black curves). I see how, only in neurons, there is a chromatin loop forming from the XYL71 GWS and the fine-mapped consensus variant towards the LRP8 promoter (purple).

Next, I performed colocalization analysis in all genes within 1Mb from lead SNPs with *P* < 1e-7. I found suggestive support for colocalization between the LiD GWAS meta-analysis signals and ci-eQTL data from Metabrain Cortex (PP H4 > 0.7 on the unadjusted colocalization analysis; PP H4 > 0.5 on the colocalization analysis after adjusting the priors based on the number of overlapping SNPs in the locus of interest) for the *DNAJB4* gene on chromosome 1 (**Table 17**). I did not find evidence of colocalization in the *XYLT1, LRP8 n*or the non-coding RNA loci

***Table 17**. Colocalization hypotheses posterior probabilities*

HGNC	nSNPs	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4	ratio_PPH4_PPH3
DNAJB4	4840	6.76E-05	7.03E-05	0.23	0.24	0.52	2.17
ZNF697	2881	2.60E-18	2.62E-19	0.72	0.07	0.21	2.93
LORICRIN	3572	5.25E-02	4.97E-03	0.74	0.07	0.14	1.96
USP33	4552	4.54E-05	4.72E-05	0.43	0.44	0.13	0.30
STXBP3	4434	1.07E-01	1.24E-02	0.72	0.08	0.08	0.92
CLCC1	4311	7.44E-10	6.81E-11	0.85	0.08	0.07	0.92

nSNPs: Number of overlapping SNPs between for each locus between the eqtl and the GWAS traits
PP.<hypothesis>: The posterior probability for each coloc hypothesis
ratio_PPH4_PPH3: The ratio of the H4 and H3 posterior probabilities (ratio = H4/H3

A few loci approaching genome-wide significance (GWS) in chromosome 1, were in proximity with *DNAJB4*. Therefore, I decided to investigate if the single causal variant assumption holds in the *DNAJB4* locus, necessary to validate the colocalization signal in *DNAJB4*. I ran GCTA-COJO under stepwise and conditional model selection procedures. I filtered all SNPs within the *DNAJB4* locus that were used to perform the colocalization analysis and that matched the AMP-PD reference panel (4590 out of 4840 SNPs included in the colocalization analysis). After performing the stepwise selection procedure assuming complete LD between SNPs that are more than 10Mb from each other, and setting a collinearity cutoff of 0.9, only the lead SNP in the locus retained nominal significance (rs278853, MAF = 0.26, β = 0.40 , se = 0.08, *P* = 4.07e-6). Similarly, running an association analysis on each of the 4590 SNPs conditioning on the lead variant (rs278853) did not show any of these SNPs to be nominally significantly associated, confirming the single causal variant assumption and that the results obtained with coloc on the *DNAJB4* locus were unbiased. Lastly, to understand whether the *DNAJB4* signal was independent of the GWS *LRP8* locus signal, I ran an analysis conditioning on the genome-wide significant *LRP8* SNP (rs72673189). I found that rs278853 remained nominally associated ($P = 4.40 \times 10^{-6}$), indicating these two signals were independently associated with the risk of developing LiD.

vi) Candidate variant analysis

I determined whether previously reported variants in the LiD literature (from LiDPD) had an impact on the time to LiD (**Table 18**). I found *ANNK1* and *BDNF* variants to be

nominally significantly associated ($P < 0.05$) with the time to dyskinesia. Nonetheless, ANNK1 or BDNF variants did not reach the significance threshold after applying Bonferroni correction according to the number of SNPs tested ($P < 2e-3$).

***Table 18**. Candidate variants analysis.*

Gene	SNP	MAF	BETA	SE	P-value	Direction	Publication
ANKK1	rs1800497	0.21	0.24	0.09	8.89E-03	+++--	Rieck et al. 2012
ANKK1	rs2734849	0.50	0.18	0.08	2.11E-02	+++++	Rieck et al. 2012
BDNF	rs6265	0.18	0.19	0.10	4.95E-02	+++-+	Foltynie et al. 2009 Kusters et al. 2018
DRD2	rs2283265	0.17	0.16	0.10	1.06E-01	+++--	Rieck et al. 2012
DRD2	rs6277	0.46	0.08	0.08	2.73E-01	----+	Rieck et al. 2012
DRD2	rs1076560	0.17	0.15	0.10	1.42E-01	+++--	Rieck et al. 2012
PRKCA	rs4790904	0.22	-0.14	0.10	1.43E-01	-++++	Martin-Flores et al. 2018
RPS6KB1	rs1292034	0.42	-0.13	0.08	1.08E-01	-----	Martin-Flores et al. 2018
OPRM1	rs1799971	0.12	-0.13	0.12	3.04E-01	+-+++	Strong et al. 2006
EIF4EBP2	rs1043098	0.49	0.06	0.08	4.67E-01	+-+--	Martin-Flores et al. 2018
SLC6A3	rs393795	0.20	0.07	0.10	4.72E-01	-++++	Kaplan et al. 2014 Purcaro et al. 2018

Gene	SNP	MAF	BETA	SE	P-value	Direction	Publication
RICTOR	rs2043112	0.40	0.05	0.08	5.50E-01	+++-+	Martin-Flores et al. 2018
HRAS	rs12628	0.35	-0.04	0.08	5.89E-01	+-++-	Martin-Flores et al. 2018
RPS6KA2	rs6456121	0.30	0.04	0.08	6.29E-01	++--+	Martin-Flores et al. 2018
COMT	rs4680	0.47	-0.03	0.08	6.65E-01	++---	Bialecka et al. 2004 de Lau et al. 2011 Hao et al. 2014 Cheshire al. 2014
PRKN	rs1801582	0.16	-0.04	0.11	7.01E-01	-+-++	Martin-Flores et al. 2018
FCHSD1	rs456998	0.49	-0.03	0.08	7.17E-01	+-+-+	Martin-Flores et al. 2018
DRD3	rs6280	0.33	0.02	0.08	7.63E-01	+-+-+	Lee et al. 2011
ADORA2A	rs3761422	0.37	0.02	0.08	7.71E-01	+-++-	Rieck et al. 2015
ADORA2A	rs2298383	0.40	0.02	0.08	8.39E-01	-+-++	Rieck et al. 2015
HOMER1	rs4704559	0.09	-0.03	0.13	8.31E-01	+++--	Schumacher-Schuh et al. 2014

Direction: Indicates the directionality of the effect of the variant across substudies included on each study

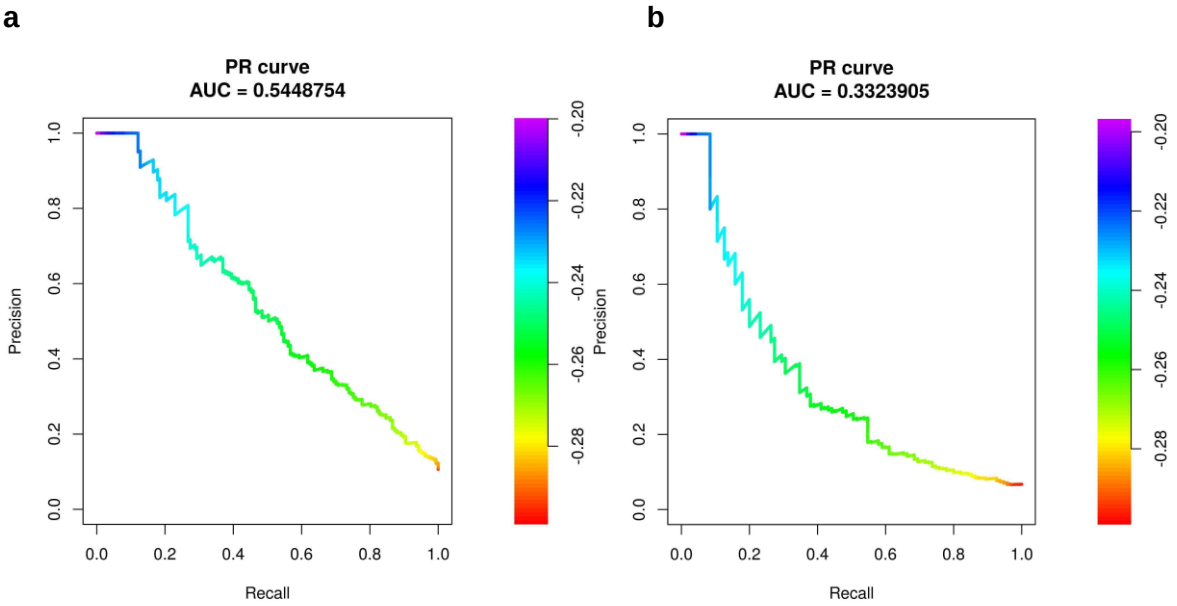
LRP8, also known as Apolipoprotein E Receptor 2 (ApoER2), is part of the low-density lipoprotein receptor family [353]. In addition, using western blot analysis based *LRP8* knockout mice models, have shown that *LRP8* knockout increases the

phosphorylation level of the microtubule-stabilising protein tau encoded by *MAPT* [354]. A previous retrospective study including 855 Caucasian PD patients found a suggestive association between the H1b *MAPT* haplotype and a higher likelihood of dyskinesia at an initial visit [355]. In the case of *XYLT1*, a previous study has found a regulatory effect of a *XYLT1* variant on the mRNA levels of *GBA1* in the substantia nigra and cortex [356]. I investigated whether *MAPT* variants (rs1800547; rs242562; rs3785883; rs2435207) were associated with the time to LiD. In addition, I explored whether *APOE* and *GBA1* variants increased the risk to develop LiD [357]. I did not find an association between time to LiD and *APOE* variants rs429358 and rs7412, or *GBA1* rs2230288 variant (E326K), or *MAPT* rs1800547, rs242562, rs3785883, rs2435207 variants. In addition, I explored genetic associations from *PINK1*, *DJ-1,* and *PRKN* intergenic variants. Whereas I did not find any genetic variant associated with time to LiD on the PINK1 locus, I found 26 *DJ-1* intergenic variants on the with a P-value < 0.05 (rs1641433611 lead SNP; HR = 1.84, SE = 0.2, *P* = 4e-4). Similarly, I found 162 intergenic variants with a P-value < 0.05 in the PRKN locus (lead SNP = rs113276175; HR = 1.84, SE = 0.2, *P* = 4e-4).

vii) PRS is capable of distinguishing patients that develop LiD.

I nominated a total of 67 independent SNPs to compute the PRS in the Tracking Parkinson's cohort. I then validated the proposed SNP set on the OPDC cohort by measuring the ability to distinguish LiD PD patients. I found that genetic data as summarised by PRS, without any other clinical or demographic data, could accurately distinguish PD patients that developed LiD at 10 years from disease onset in two separate cohorts: Tracking Parkinson's (AUC 83.9) and OPDC (AUC 87.8). At 10 years after PD onset, I found that 16% of patients had LiD in the Tracking Parkinson' cohort, and 18% of patients had LiD in the OPDC cohort. Class imbalance can lead to inaccurate evaluation of classifiers. Therefore, I also computed precision recall curves (PROC) as large class imbalance can lead to biassed ROC curves when assessing the performance of a classifier. I found the PROC AUC to be lower in both TPD (AUC = 54.49) and OPDC (AUC = 33.24) (**Figure 19**).

***Figure 19**. PROC curves for PD and LiD patients PRS.*



***a)** PROC PRS in Tracking Parkinson's and **b)** PROC PRS in OPDC. The color scale on the right side of the plot gives an indication, which classification threshold results in a certain point on the curve. PR = Precision recall. Precision = Positive predictive value. Recall = Sensitivity.*

viii) Baseline predictors of LiD development.

I used Tracking Parkinson's data at baseline in a stepwise regression approach using a logistic model. I then filtered out from the final model predictors that were not significantly associated after applying Bonferroni correction ($P < 0.05 / 502 = 1e\text{-}4$).

In addition to the PRS, which was significantly associated with a increase of the odds of LiD (OR = 10.01, SE = 0.57, $P = 1.07e\text{-}30$), I found that anxiety at baseline (as measured by the Leeds Anxiety and Depression Scale [358]) was significantly associated with a increase of the odds of LiD (OR = 1.14, SE = 0.03, $P = 7.4e\text{-}5$). I also explored clinical features previously reported as being associated with an increased or decreased LiD risk. Sex, AAO, and 5PCs were added in the base model of the stepwise regression approach. Consistent with previous studies as well as with my CPH model highlighted above, younger AAO increased the LiD odds (OR = 2.41, SE = 0.04, $P = 4e\text{-}3$). However, sex was not found to be significantly associated in my final model including PRS and Leeds anxiety status.

Neither smoking status nor BMI were selected on the stepwise regression approach, consistent with what I found when I individually explored known LiD risk factors (**Figure 12**). Interestingly, I also found that PD family history was selected in the stepwise regression analysis, and was nominally significantly associated with an increase in the odds of LiD (OR = 1.62, SE = 0.14, *P* = 6.9e-4).

Finally, I attempted to replicate the association between dyskinesia state and anxiety using the State-Trait Anxiety Inventory [359] available in PPMI. I did not find the Trait Anxiety Score to be significantly associated with LiD patients in PPMI (OR = -0.03, SE = 0.04, *P* = 0.44).

iv) Patients with LiD have an average higher cognitive scoring

I assessed the cognitive status of LiD patients because of the association between the *LRP8* nominated locus and *APOE*. I explored whether the cognitive state differed between patients developing LiD and patients who did not develop LiD during the study length using the Wilcoxon rank sum non-parametric test with continuity correction, as I observed the data was not normally distributed. In addition, I also looked into differences in the MDS-UPDRS part III scores between the two groups, using the unpaired two samples t-test to compare the mean of two independent groups. I compared the LiD group (N=172) against the non-LiD PD group (N=1318) using data from Tracking Parkinson's alone as it is the largest deeply phenotyped cohort I had available. I did not find differences in the average MDS-UPDRS part III total score, neither at baseline nor at the visit when patients first developed LiD (or the last available visit in cases who did not develop LiD) (**Table 19**). However, PD patients who did not develop LiD through the study had a significantly lower MoCA score on average at baseline, as well as at the final visit (**Table 19**).

***Table 19**. MoCa and UPDRS score comparison between PD-LiD and PD groups*

Variable	method	p.value	statistic	PD group mean(sd)	PD-LID group mean(sd)
moca_bl	Wilcoxon rank sum test with continuity correction	2.1e-05	73754.5	25.16(3.31)	26.09(3.56)
moca_visit	Wilcoxon rank sum test with continuity correction	0.01	77249.5	24.34(4.77)	25.35(4.11)
updrs_III_bl	Welch Two Sample t-test	0.15	-1.45	22.2(11.6)	23.72(12.1)
updrs_III_visit	Welch Two Sample t-test	0.25	1.15	31.91(16.7)	30.42(14.0)

moca_bl = Moca average scores for the LiD and PD group at baseline
moca_visit = Moca average scores for the LiD and PD group at the time LiD was developed or at the last visit available
updrsIII_bl = MDS-UPDRS III averages scores for the LiD and PD group at baseline
Updrs_iii_visit = MDS-UPDRS III average scores for the LiD and PD group at the time LiD was developed or at the last visit available

d) Discussion

I have performed an untargeted genome-wide study to define genetic variants associated with the time-to-LiD in PD, using a CPH model under a genetic additive effect and analysed the effect of genetic and baseline clinical variation on the development of LiD. I found genome-wide significant associations with the time-to-develop LiD at the *LRP8*, *LINC02353* and *XYLT1* loci. These associations were consistent across all the cohorts included in the meta-analysis. I also performed a candidate gene analysis, exploring genetic variants reported to be associated with LiD risk in my large GWAS meta-analysis. I found that genetic variability in *BDNF* and *ANKK2*, were nominally associated with LiD. I did not replicate any other variant associated with LiD risk.

LRP8 expression is enriched in brain tissues such as the neocortex, cerebellum, hippocampus and olfactory bulb [353]. *LRP8*, together with *VLDLR*, is a mediator of the Reelin pathway, which contributes to development of the central nervous system as well as to facilitate neuronal migration [360,361]. LiD develops in the context of ongoing neuronal loss, and synaptic/signalling changes related to dopamine therapy. My finding suggests the changes in the Reelin pathway and neural development / plasticity may be important in the development of LiD.

In addition, the *LRP8* protein stabilises the microtubule-stabilising protein tau and it has been shown that knocking out *LRP8* in mice increases tau phosphorylation [354]. Post-hoc functional annotation analysis revealed a chromatin loop between an enhancer at the third intron of *LRP8* (where the lead variant was found) and the *LRP8* promoter, thus providing functional support for *LRP8* as the causal gene at this locus. In addition, a colocalization analysis, looking at all genes within ±1Mb from all GWAS variants with P-value < 1e-7 revealed a second association in chromosome 1 with the *DNAJB4* gene. Conditional analysis further confirmed that both regions were in LD, hence both *LRP8* and *DNAJB4* were independently associated with the time-to-LiD. I also found a similar event of distal regulation in the *XYLT1* locus, although the chromatin loop did not perfectly match with the GWAS signals, making the functional annotation analysis inconclusive. Moreover, I found that the two GWAS nominated signals overlapped with Transcription Factor Binding Sites marks from the ENCODE project, adding further support for the transcription machinery being recruited in the GWAS loci and regulating both genes expression after forming the enhancer-promoter distal chromatin loops. Nevertheless, whereas I found a chromatin loop suggesting regulation of *XYLT1* and *LRP8* gene expression, I did not find statistical support for gene regulation based on the colocalization Bayesian framework.

The three nominated protein coding genes have been previously reported to be functionally associated with putative PD genes, which may provide an insight into the development of LiD. *LRP8* encodes the low-density lipoprotein receptor-related protein 8, and it has been found to be associated with *APOE*. In addition, the *LRP8* protein stabilises microtubule-stabilising protein tau and it has been shown that knocking out *LRP8* in mice increases tau phosphorylatIon [354]. *DNAJB4* gene encodes a molecular chaperone tumour suppressor, and member of the heat shock protein-40 family. Mutations in the DNAJ family protein have been reported to cause or increase the risk of several neurological disorders, including Parkinson's disease [362]. *XYLT1* encodes a xylosyltransferase enzyme which takes part in the biosynthesis of glycosaminoglycan chains. A previous study has found a regulatory effect of a *XYLT1* variant on the mRNA levels of *GBA1* in the substantia nigra and cortex [356]. I did not find support for colocalization with eQTLs nor evidence suggestive of epigenetic regulation of genes in the *LINC02353* locus. *PCDH7*, the nearest gene coding protein gene, encodes a protein with an extracellular domain

containing 7 cadherin repeats. This gene has been described as a potential PD biomarker [363].

At an individual patient level, treatment strategies including levodopa and non-levodopa therapies, and the use of deep brain stimulation (DBS) are determined by the emergence of motor complications including LiD. The ability to develop a predictive algorithm to enhance clinical care would improve the outlook for PD treatment. Here, I have shown that both clinical and genetic variables have the potential to have a high predictive value for the development of LiD. This will need to be validated in further cohorts and I hypothesise that the integration of further 'omics data (e.g. RNA and proteomics), using machine learning may lead to the definition of an accurate predictive model for defining PD patients at risk of developing dyskinesia.

I have analysed a large dataset with detailed clinical, drug exposure and genetic data. I have carefully tested for confounding by PD age at onset, gender, population structure and shown that my results are free of confounding effects as well as demonstrating they are consistent across cohorts. Because the dose of levodopa may be a major confounder in my study, I tested the effects of adjusting for levodopa dose on a sensitivity analysis, and found that the lead SNPs on *LRP8*, *LINC02353* and *XYLT1* loci remained significantly associated with the outcome, concluding that levodopa treatment was not a confounder in my study design. Likewise, adjusting for the MDS-UPDRS part III total score at the time of LiD development did not change the significance levels of the lead SNPs, suggesting that my findings were not confounded by motor severity or progression.

Although this is a large study there are limitations based on sample size. According to my sample calculation, I would be 80% powered to detect associations with the LiD phenotype from variants with a MAF of 0.01 when I reached a sample size of 18000 patients. In addition, my results are limited to individuals of European ancestry and I have not explored whether there is a shared common genetic variability associated with changes in LiD survival across different ancestries. Expanding this analysis to PD genetic datasets with deeply phenotypic data available from initiatives such as the Global Parkinson's Genetic Program (GP2) will give us new insight into the genetics of PD LiD patients as well as serve as a valuable resource for validation of findings [155].

MDS-UPDRS 4.1 is a simple but widely used measure which documents the appearance of LiD. Potentially, more detailed scales such as the Unified Dyskinesia Rating Scale[364] would provide a more accurate measure of the extent and impact of LiD, which would improve future GWAS.

Overall, I have found new evidence of common genetic variability associated with the time-to-LiD. I have been able to map genes at nearby risk loci, as well as provide fine mapping support of potential causal variants for LiD traits. Likewise, I hope to help design personalised medicine strategies that prevent PD patients developing dyskinesia according to their genetic burden which could be tested with the proposed PRS in this study. Similarly, I hope to help understand the molecular pathways that lead to LiD. Targeting nominated genes might allow the development of LiD treatment strategies. Further investigation regarding the overlap between anxiety GWAS and my GWAS might help understanding common causal pathways between the two conditions. Understanding shared mechanisms will help us prevent medication adverse events affecting non-targeted pathways and to fine-tune current treatments.

5) Global large-scale analysis in Parkinson's disease using long-gwas provides new insights into the genetic determinants of Parkinson's disease phenotypes.

a) Introduction

Complex genetic diseases are thought to develop due to the effect of multiple common genetic risk variants and environmental risk factors, rather than highly penetrant single gene variants. Genome-wide association studies (GWAS) have been used to define the genetics of complex traits and diseases [51,365,366]. Despite the upper limit of genetic studies given by the heritability of the trait, and how much of it can be captured by the study designs of trait-disease-specific GWAS and data availability, estimates of variant effects can be used to predict the genetic predisposition to disease in individuals who have not yet developed the condition [367]. Moreover, insights from GWAS have proven to be transferable when developing clinically actionable strategies to deal with the development and progression of disease. The *PCSK9* gene exemplifies the successful translation of GWAS knowledge into FDA-approved disease-modifying treatments. After the discovery of *PCSK9* mutations causing autosomal dominant hypercholesterolemia [368], subsequent GWAS studies identified genetic variants associated with low-density lipoproteins (LDL) cholesterol levels and coronary heart disease risk [369]. *PCSK9* controls LDL levels targeting LDL receptors hence being associated with changes in LDL levels in plasma [370]. After promising clinical trials results showing the efficacy of *PCSK9* inhibition to control treat hypercholesterolemia and reduce the risk of cardiovascular events [371], Alirocumab, a *PCSK9* inhibitor, was approved by the FDA to treat hypercholesterolemia [372].

In PD, the largest GWAS of European ancestry patients to date defined a total of 90 independent variants at 78 loci [51]. All risk variants together have been found to have a high predictive capability for disease diagnosis alone (AUC: 70%) based on machine-learning model training and optimization [373]. However, the risk variants uncovered so far explain just a fraction (16-36%) of the total heritability component of idiopathic PD estimated to be 22% [51]. Recent work expanding the largest PD genetic study to more ancestry diverse populations, including PD samples from East Asian,

Latino and African individuals, identified 12 novel loci [82]. Despite the unquestionable progress uncovering the genetic architecture of PD, these well powered case-control GWAS meta-analyses [4] have not contributed to the identification of novel disease modifying treatments. An alternative approach is the identification of targets for disease-modifying treatments based on phenotypic progression and severity GWASs that can identify genetic loci associated with prognosis. Together with downstream functional annotation to establish the nearest genes and their expression patterns [285], this strategy might be particularly powerful in pinpointing actionable targets for disease-modifying treatments.

In recent years, several progression and severity genetic studies have been conducted in PD [108,109,114,374–376] using different genetic quality control and modelling strategies. These have successfully identified genetic markers associated with different PD outcomes, such as motor and cognitive performance. Here, I introduce long-gwas, a Nextflow pipeline that makes longitudinal and severity disease-specific GWAS accessible and scalable, and reproducible by decreasing the introduction of user systematic errors. Here, I used long-gwas to conduct a large-scale proof of concept disease severity analysis across multiple phenotypic outcomes and ancestry groups. I describe putative loci for hyposmia, an established phenotypic marker of α-synuclein, pathology, as well as other potential novel genetic markers that capture diverse aspects of the PD symptomatology.

b) Methods

i) Long-gwas

The development of long-gwas has been part of a collaboration with Michael Ta at the NIH. My participation can be found on GitHub in the commit history (https://github.com/michael-ta/longitudinal-GWAS-pipeline/commits/main/) and the pull requests from developmental branch I created in the process of improving and adding new features to the tool (https://github.com/michael-ta/longitudinal-GWAS-pipeline/pulls?q=is%3Apr+is%3Aclosed)

Here I summarise in brief, my main contributions to the software development:

- Rewrite of the software on DSL version 2

- Developing a modularised version of long-gwas
- Developing configuration files for cloud and HPC based execution support
- Enabling of parallel executions of multiple GWAS outcomes
- Dealing with outcome missingness time points on the fly during pipeline execution
- Enhancement of Manhattan and QQ plots
- Enhancement of results processing (merge all splits into one readable file)
- Improvement during runtime to only load modules needed for analysis (if GLM analysis is needed, then I do not load other modules which speeds up the tool and use less resources during runtime)
- Several bugs fixes for example wrong SNPs ordering between chromosome splits before splits merging in plink, or overcome memory limitations when merging chromosome files in plink
- Remove deprecated arguments no longer used in the new version
- Yaml file to more efficiently write parameters
- Enhance docker image to incorporate missing dependencies
- Add Nextflow scripts and tree structure needed to run lon-gwas from GitHub without the need of manually download the workflow
- Update web based documentation supervising a research assistant in the lab
- Efficiently use the bin folder in the GitHub remote to allocate the custom scripts I use in long-gwas that enable no longer needing to hard code those scripts in the docker image, lighting up the image and making deployment faster

Long-gwas is a workflow developed using Nextflow domain-specific language version 2 (DSL2). Nextflow dataflow programming is inspired by the Unix philosophy, in the sense that it is based on the use of pipes, and one can chain multiple simple operations together. One Nextflow structure called channels enables multiple tasks to communicate by piping the output of a task to the input of a downstream task. Likewise, parallelization inherently happens based on how process outputs are channelled into other processes, avoiding the use of complex parallelization definitions [377,378]. ***Long-gwas*** exploits the parallelization capabilities of Nextflow to speed up the process of running time-consuming end-to-end (longitudinal) GWAS analyses. It ensures workflow portability and reproducibility of results based on the containerization of the tool using the Docker software platform. In addition, long-gwas

configuration supports Google Cloud Batch execution, making scalability possible when users intend to run long-gwas on heavy genomic data batches. The GitHub integration of long-gwas enables users to constantly track software changes.

Long-gwas covers all the steps that are necessary to perform a (longitudinal) Genome-Wide Association Analysis. Based on our experience performing this type of analysis, we have integrated the tools that are required to achieve the different goals at each step of the workflow. In addition, we automatically deal with common pitfalls when performing this type of analysis.

Long-gwas is intentionally developed so that anyone with basic understanding of the command line can perform their analyses. Nextflow, and for instance long-gwas is supported in the main operating systems (Linux, Mac OS, and Windows through WSL2). Whereas we have hard coded on the long-gwas configuration file the amount of resources (CPUs and RAM) that tasks need to run, Long-gwas implements an efficient caching approach so that very time consuming sections only need to be run once when variations of a task need to be executed for the same input genetic data. For instance, when working with input data for a given study, if the number of genetic variants in the input is very high, the operation of loading the data and performing initial genetic data QC is time consuming. As this step is highly generalisable across any GWAS, we save the arranged output of this process on a cache directory so that any subsequent run will skip this first step. Therefore, we can rapidly re-execute parts of the workflow of interest. Each step is encapsulated on a container via Docker or Singularity. Because we acknowledge many users might not have access to an HPC or a cloud based platform, we provide a Nextflow configuration in long-gwas to support local (HPC or personal desktop) or Cloud-based executions of the workflow. This configuration file can be further customised by the user to exploit the job schedulers and cloud platforms that Nextflow supports for the workflow deployment.

i.i) Inputs and outputs

Long-gwas inputs are handled through a yaml file. We provide a thorough description of each long-gwas input file and argument on the web-based documentation page [379].

We can group the workflow inputs in four main types:

Data files. Chromosome level genetic VCF input files, a covariates file, and a phenotype file are the data files needed to run a long GWAS following multiple steps.

Quality control arguments. Even though the data QC is fully automated in long-gwas, we grant users full control of parameters to make decisions on data quality at both the sample and the genetic level. For instance, users can specify the r-squared threshold to filter out genetic variants with low imputation confidence, the minor allele frequency and count of variants to retain for analysis, the kinship parameter to decide the accepted relatedness between each pair of samples from the input, the ancestry from which we want to keep individuals from the input data to account for population structure, and the assembly of the input genetic data to infer whether genetic data liftover is necessary or not. Finally, we also provide a variable to enable users to decide how to create the chunks of genetic files at the chromosome level. Based on this variable, chromosome level data is split into chunks of N genetic variants enabling Nextflow to create a parallelization backend to run the genetic quality control.

GWAS model parameters. In a generalisable way across all the models, we allow users to specify the covariates, phenotypes, as well as the ability to perform a separate GWAS for each study code based on a study grouping variable. Long-gwas supports the running of one model at a time. We provide boolean (True or False) variables to specify whether to use the gallop powered linear mixed effect (LMM) model, the Cox Proportional Hazard (CPH) model, or the Generalised Linear Model (GLM) model. More specifically, for the LMM models, which are meant to be hierarchical and grouped at the individual level, we provide a time variable to specify the time point each record was taken for each sample ID. Similarly, for the CPH model this time variable is used to specify the time for individuals to reach the outcome. We also provide a variable to customise the GWAS results output name.

Cache and results directory. The *dataset* argument is used in long-gwas to create a folder in which we save the multiple outputs of the long-gwas workflow, as well as a cache directory where we save intermediary output files that could be reused in other analyses with the same input genetic data files.

Once the long-gwas workflow is completed, a results directory is generated that contains the following **output** files:

- Diagnostic plots (Q-Q plot, PCs scree plot, 2-D PCs plots).

- GWAS results (Manhattan plot, and a tsv file with all GWAS SNP-level results).
- Operation logs (log file for merging operations after preprocessing step and log file for the PC inference).

i.ii) Genetic association models

Long-gwas is intentionally developed to support three main types of statistical models to perform GWAS: GLM, LMM powered by GALLOP and CPH. More information on each type of the model is available **in Chapter 2 – Methods**.

i.iii) Workflow description

Long-gwas can be divided into three main sub workflows summarised in **Figure 20**.

The **preprocessing** step is the first sub workflow run using the long-gwas tool, in which data is prepared for the downstream analysis. Genetic data is processed to retain highly confidently imputed genotypes. We left-align and normalise indels, and also split multiallelic sites into biallelic records using bcftools. We generate plink2 genetic binary files on the output. Long-gwas standardises genetic data into build hg38. Therefore, if input files are provided in hg19 or other build, long-gwas performs the liftover of genetic files using the LiftOver tool from UCSC [380], In addition, we ensure consistent representation of variants in the input VCFs, by normalising genetic variants using vt normalise [381] and the FASTQ hg38 reference panel as reference for the alignment. Lastly, genetic data is processed to remove singletons, duplicate variants, as well as to keep only those genotypes with high call rates. We accelerate this time-consuming first step by splitting the input genetic data at the chromosome level in multiple chunks of a defined number of SNPs, each based on the *chunk_flags* long-gwas argument.

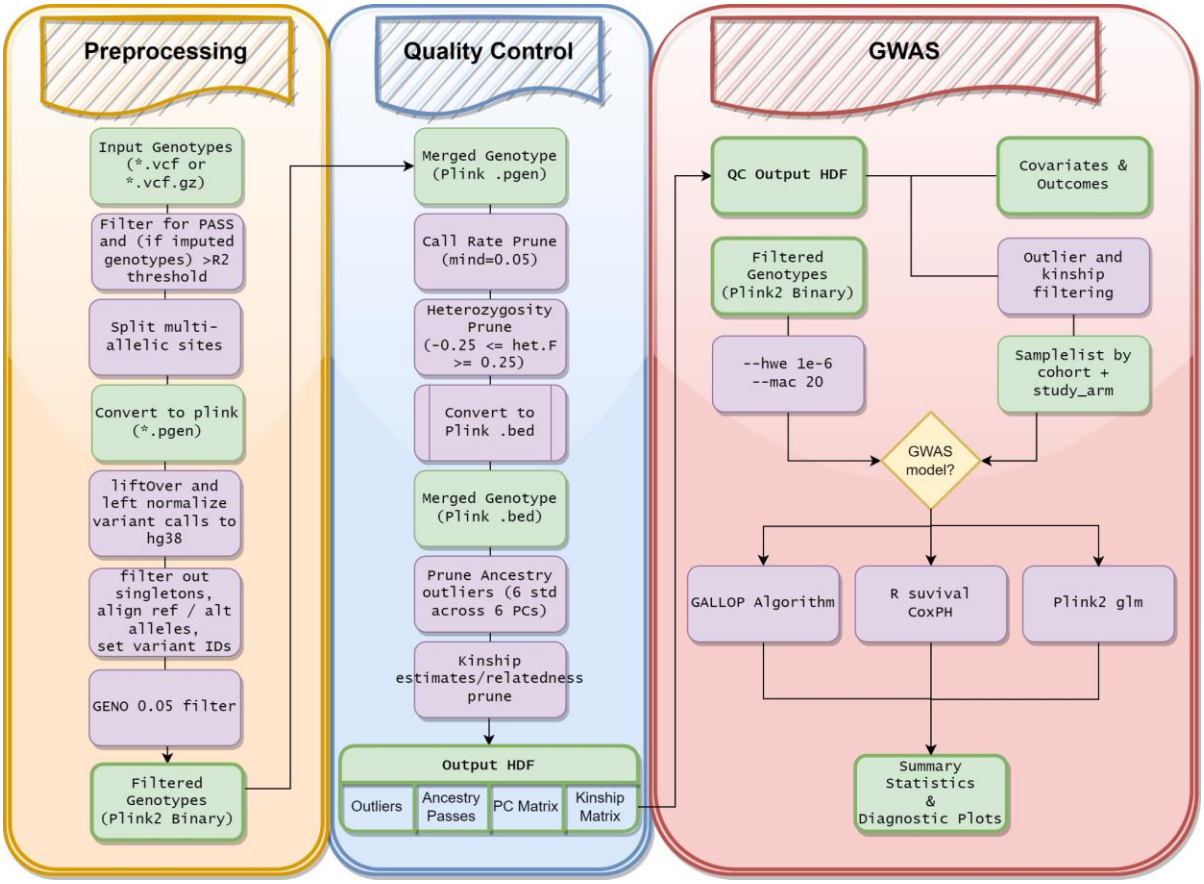
Once the QC is complete for each data subset, we merge the chunks back into one chromosome file, and those into one unique genetic file containing all genetic variants for all chromosomes of all samples in the study. We generate bed and pgen output file formats using plink2 software.

The **second step** of the workflow is to perform **quality contro**l that guarantees we only retain high quality samples and genetic data for analysis. For that, samples with high genotyping missingness rate are removed, heterozygosity and ancestry outliers are pruned and one individual of each pair of related individuals at the first degree

relative level are removed. Genetic variants that are not in Hardy-Weinberg equilibrium are excluded, as well as variants with a minimum allele count <20 (rare variants exclusion). For interpretability, we provide descriptive plots highlighting the data processing during QC. For instance, we provide a scree plot showing the number of PCs that were included to decide which samples were selected as ancestry outliers, as well as a 3D plot showing the PCs distribution against a reference panel. We provide the sample list that passed all QC steps on an h5 file.

The **third step** of the workflow is to run a **GWAS analysis** on the resulting QC data. The covariates and phenotype data are filtered to match samples that passed the QC stage using the h5 file from the QC step. In addition, new PCs are calculated on the final QC subset of data, and these are merged with the covariates file. Users can specify PCs as covariates, in which case the model will be fitted accordingly. For each outcome, we remove missing outcome data points that would make the model fit fail, and report any sample excluded for this reason. Currently, three types of analyses are available in long-gwas. GLMs can be deployed to evaluate the effect of variants either on disease risk (case-control GWAS) or on disease severity measured by quantitative cross-sectional clinical instruments. CPH models [382] can be used to determine the impact of genome-wide variants on survival based on a predefined outcome relevant to the disease under study. LMMs powered by GALLOP [194], in which we can evaluate the impact of genetics on both disease progression (slope-term) and severity (intercept-term), allowing for random sources of variation to account for unexplained heterogeneity at both disease presentation and progression. We exploit the Nextflow parallelization capabilities by splitting the data into multiple chunks and running a GWAS on each chunk in parallel. In addition, long-gwas can be run simultaneously on many outcomes, by parallelizing the GWAS analysis of all outcomes specified on the long-gwas *pheno_name* argument. Finally, we merge all data into one file per outcome, and we generate the Manhattan and diagnostic plots (QQ plots) for all results.

***Figure 20**. Long-gwas pipeline schematic overview.*



First we do a pre-processing step, followed by quality control. Finally, long-gwas performs a GWAS based on different settings on the QCed data subset.

ii) Study design and participants

In my initial work with the long-gwas tool I have used four data sources: the Unified cohorts from the Accelerating Medicines Partnership Program in Parkinson's Disease (AMP-PD Unified, (https://amp-pd.org/unified-cohorts), which gathers common clinical and genomic data in an harmonised format from eight different cohorts (BioFIND, HBS, LBD, LCC, PDBP, PPMI (genetically enriched and sporadic PD patients), STEADY-PDR, SURE-PD3) with whole genome sequencing; the federated cohorts provided by the Global Parkinson's Genetics Program (AMP-PD GP2), genotyped with NBA [155,347,383]; the Tracking Parkinson's cohort; and the Oxford Discovery cohorts [150,151]. Based on participant longitudinal data, we excluded any participant whose latest diagnosis was not PD. If longitudinal data was not available, we used the latest diagnostic information. We carried out clinical data quality control by excluding: PD patients without clinical or demographic records at baseline, patients

with unmatched clinical and genetic data and patients with long disease duration at baseline (>20 years). In addition, we found the case of negative values calculating disease duration at baseline (diagnosis before reported symptom onset). We excluded these from analysis.

For each cohort, we selected the following clinical assessments: MDS-UPDRS Part I (non-motor aspects or experiences of daily living), Part II (motor aspects or experiences of daily living), Part III (motor examination; recorded in the on-state for treated participants), HY stage, MoCA, MMSE, SEADL (a scale to assess the capabilities of people with impaired mobility), RBD questionnaire, University of Pennsylvania Smell Identification test (UPSIT), the Epworth Sleepiness Scale (ESS), and the Parkinson's Disease Questionnaire (PDQ-8). In addition, we accessed the α-synuclein amplification assay from AMP-PD PPMI study participants.

We matched clinical and genetic data in the remaining participants after QC, resulting in deeply phenotyped PD records for two ancestry groups: Ashkenazi Jewish and Europeans. Data availability for the rest of the ancestry groups was lower than 100 samples. We considered a genetic association study on less than 100 samples to be under-powered so we excluded these samples from the study.

iii) Genetic data quality control

To assign ancestry labels to patients that passed the sample quality control stage, we used GenoTools, a python module to perform data quality control and genetic analyses [17]. To do quality control based on the available genetic data, we used the ***long-gwas*** arguments to perform the automated quality control steps in the workflow, based on our input parameters. During the data preparation step, we excluded non-autosomal or singleton genetic variants. For Tracking Parkinson's and Oxford Discovery cohorts for which data was in hg19, we lifted over the data using long-gwas, by setting the assembly argument (**--assembly**=hg19). When using imputed genetic data, we filtered out variants with a R squared value <0.7 (**--rsthres**=0.7). In addition, we only kept genetic variants with minor allele frequency >5% (**--minor_allele_freq**=0.05), and variants with a minor allele count of at least 20 (**--minor_allele_ct**=20). We filtered one patient out of a pair of closely related individuals by generating a kinship square matrix and setting a kinship argument threshold (**--

kinship=0.177). To account for any possible population stratification, we ran the ancestry outlier detection long-gwas utility, which uses 1000 Genomes as a reference panel [220] . For instance, if the input genetic data was from an AJ subset inferred by ancestry tool, then we set up the ancestry argument to Ashkenazi Jewish (**--ancestry**=AJ) or European (**--ancestry**=EUR) ancestry. In addition, we derive genetic principal components (PCs) that we use to adjust the genetic model for population stratification.

On average, a total of 6,500,000 variants remained available for each independent ancestry cohort level data subset. For each genetic analysis on the subset of SNPs, we generated Q-Q plots for a graphical check, as well as estimated genomic inflation factors to guarantee the GWAS test-statistics were a unique function of genetic variability and were not inflated due to cryptic-relatedness and/or population stratification. To do so, we set the long-gwas arguments *mh_plot* to true (**--mh_plot**=true).

A further description of all the arguments we have used to conduct data quality control can be found on the long-gwas web-based documentation [384].

iv) Genome-wide disease severity model and meta-analysis

I performed multiple genome-wide association studies across the range of quantitative outcomes. I used clinical baseline measures (cross-sectional data points). Data availability for each outcome is described in detail in **Table 20**. I studied the impact of genetic variants on disease severity using GLMs, with sex, cohort, age at baseline (AAB) and the first three genetic PCs as covariates.

Table 20*. Overview of data availability for each clinical outcome across ancestry groups in the four PD data sources*.

	AMP-PD Federated (GP2)		**AMP-PD Unified**		**OPDC**		**PROBAND**		**N total**	
Ancestry	EU	AJ	EU	AJ	EU	AJ	EU	AJ	EU	AJ
MDS-UPDRS I	-	-	2128	345	787	4	1619	10	4534	359
MDS-UPDRS II	-	-	2127	345	787	4	1622	10	4536	359
MDS-UPDRS III	1052	141	2170	344	786	4	1589	9	5597	497

	AMP-PD Federated (GP2)		AMP-PD Unified		OPDC		PROBAND		N total	
HY	1470	165	2763	421	797	4	1699	11	6729	601
MOCA			1669	244	788	4	1578	10	4065	258
MMSE	1497	249	-	-	792	4	-	-	2289	253
SEADL	425	-	2167	398	-	-	1694	11	4286	411
UPSIT	-	-	1575	**320**	-	-	1089	8	2664	328
RBD	-	-	1306	335	-	-	1600	10	2906	345
ESS	-	-	-	-		-	1636	10	1636	10
PDQ8	-	-	-	-	-	-	1634	10	1634	10

I used METAL software (version released on 25/03/2011) to meta-analyse results of multiple GWASs at the cohort level for matched ancestry groups. I applied a fixed-effect model based on the sum of the β coefficients for each SNP *i* and study *j*, weighted by the inverse of the variance of the estimated effect of the *j*th variant in the *i*th study (1 / [Var(*βij*)]). Upon meta-analysis, I applied a genomic control correction to the cohort-specific summary statistics by computing the inflation of the test statistic, and then applying the genomic control correction to the standard errors. Similarly, I used the METAL software fixed-effect model to perform a meta-analysis across ancestries to investigate the homogeneous allelic effects between ancestry groups for a targeted clinical outcome, the UPSIT Olfactory Test.

To ensure the consistency of allelic effects across the multiple genetic studies included in the meta-analyses, I filtered out heterogeneous variants based on the Cochran's statistic (test of heterogeneity of allelic effects) and I2 (a quantification of the extent of heterogeneity in allelic effects across GWASs). SNP level estimates were excluded if the *P*-value for the Cochran's Q-test for heterogeneity was <0.05 and the *I*2 statistic was ≤80%**.** In addition, I filtered meta-analysis results to only include genetic variants that were available for at least 40% of the total genetic markers.

v) Proteome and transcriptome differential abundance and expression analysis

To find transcriptomic and proteomic based biomarkers based on *LRRK2* G2019S and *GBA1* N370S status, I accessed blood whole-transcriptome counts and Data-

Independent Acquisition mass spectrometry-based ("untargeted") proteomics from cerebrospinal fluid (CSF) Untargeted protein abundance measures from AMP-PD release v3. A more detailed overview on the transcriptomic data preparation is available at (https://amp-pd.org/transcriptomics-data#workflows). A more detailed overview on Non Targeted Proteomic data preparation is available at (https://amp-pd.org/data/untargeted-proteomics-data).

For our differential expression transcriptome analysis, I used the limma R package. I transformed count features to log2-counts per million (logCPM), estimated the mean-variance relationship and used this to compute appropriate observational-level weights. I then computed a linear model fit for each gene adjusting by sex, the plate number, age and neutrophils and lymphocytes percentages. Finally, I computed moderated t-statistics, moderated F-statistic, and log-odds of differential expression by empirical Bayes moderation of the standard errors towards a common value.

For the analysis of differential abundant protein analysis, I used a custom function in Python. I transformed protein abundance to log2-counts, and removed the variance on measures driven by age and sex. Then, I ran a t-test to get the significance of the mean difference between two groups (mutation carriers versus non carriers).

vi) Functional annotation of genetic association results

To annotate results from genetic studies, I used the Functional Mapping and Annotation (FUMA) v1.3.8 web platform. FUMA defines genomic risk loci based on linkage disequilibrium (LD) blocks composed of independent significant SNPs at Rsq > 0.6. To define LD blocks, I set up a distance of 500kb between different LD blocks edges. For each LD block, a subset of lead SNPs is defined by finding the independent significant SNPs that are independent of each other at Rsq 0.1. Then, these SNPs are merged into LD blocks separated 250Kb from one another. Each locus is represented by the top lead SNP (minimum P-value) in the locus. To calculate the Rsq, I used the 1000 Genome Reference panel with all ancestries.

Independent significant SNPs are mapped to genes based on a window size of 10kb using ANNOVAR. To know the functional consequences of SNPs using ANNOVAR in FUMA. I also queried several eQTL sources to map SNPs to genes which likely affect

expression of those genes. I only used cis-eQTL data (up to 1Mb) from eQTLGen, PsychEncode, and GTEx V8 data sources [269,385,386].

Finally, I determined the tissue specificity of our PD phenotypes meta-analysis based on an enrichment analysis against all differentially expressed genes by performing a two-sided t-test for any label against all others for all GTEx V8 tissues. Then based on the sign of t-statistics, up-regulated and down-regulated DEGs were also computed. To test input genes derived from the meta-analysis summary statistics against each DEG set for each tissue, the hypergeometric test was used. Significant enrichment at Bonferroni corrected P-value are coloured in red.

c) Results

i) Summary of clinical and demographic data available for analysis

In this analysis, I included 8,458 European and 963 Ashkenazi Jewish PD cases (**Table 21)**. The GP2 European cohort had the longest mean disease duration from diagnosis. GP2 European cases were also younger at baseline on average. For AJ ancestry cases, average disease duration was also longer in the GP2 cohort. AAB was similar among the AJ cases from the different data sources. The male:female ratio was very similar across the studies for the European ancestry samples. However, the proportion of males among AJ cases varied between 58% and 100% in the different cohorts

Mean motor and cognitive scores at study entry, as measured by MDS-UPDRS part III and MoCA respectively, were similar across cohorts. Interestingly, the Oxford discovery cohort showed the lowest proportion of patients in a more advanced disease stage, as measured by the Hoehn and Yahr scale, as well as the highest average score of MDS-UPDRS part III. UPSIT and RBD assessments are used to quantify the olfactory impairment and REM sleep behaviour disorder, and they were available for the AMP-PD Unified and Tracking Parkinson's cohorts. There were no differences in the average scores for these two assessments across cohorts. I did not find clear differences in the average outcomes capturing PD non-motor features or the motor and cognitive states between European and AJ groups.

***Table 21**. Summary of clinical and demographic features at baseline across European and Ashkenazi Jewish groups*

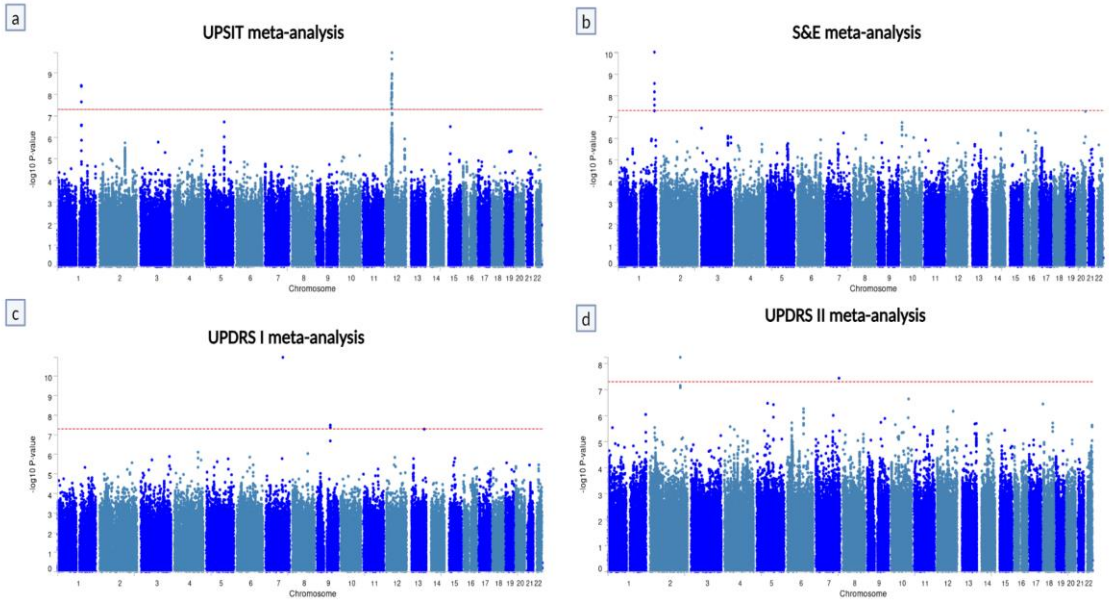
	AMP-PD Federated (GP2)		AMP-PD Unified		OPDC		PROBAND	
Ancestry	EU	AJ	EU	AJ	EU	AJ	EU	AJ
N	3492	429	2470	519	797	4	1699	11
Age at baseline	63·1 (21·0 to 93·0)	66.3 (40·0 to 91·0)	64·2 (30·0 to 90·0)	66.85 (32·0 to 90·0)	67·3 (32·2 to 90·5)	67.9 (58·2 to 81·6)	67·5 (31·1 to 90·4)	64.7 (49·5 to 75·6)
Disease duration,	4·4 (0 to 19·0)	6.1 (0·0 to 19·0)	3·6 (0 to 19·0)	4·8 (0 to 19·0)	1·3 (0 to 3·5)	0.7 (0·3 to 1·2)	1·2 (0 to 3·4)	1.5 (0·1 to 3·0
Male sex	2134 (63%)	306 (71%)	1545 (63%)	302 (58%)	513 (64%)	4 (100%)	1104 (65%)	7 (64%)
Female sex	1307 (37%)	123 (29%)	925 (37%)	217 (42%)	284 (36%)	0 (0%)	595 (35%)	4 (36%)
MDS-UPDRS I	-	-	7·7 (0·0 to 37·0)	8.6 (0·0 to 37·0)	8·7 (0·0 to 33·0)	8.2 (5·0 to 12·0)	9·2 (0·0 to 34·0)	11.2 (4·0 to 21·0)
MDS-UPDRS II	-	-	8·2 (0·0 to 46·0)	8.7 (0·0 to 39·0)	8·8 (0·0 to 33·0)	6.0 (4·0 to 10·0)	9·8 (0·0 to 48·0)	7.9 (1·0 to 20·0)
MDS-UPDRS III	21·1 (0·0 to 100·0)	20·7 (2·0 to 64·0)	23·2 (0·0 to 83·0)	24.4 (0·0 to 83·0)	26·6 (5·0 to 77·0)	24.5 (14·0 to 35·0)	22·9 (1·0 to 76·0)	13.6 (3·0 to 26·0)
Hoehn and Yahr stage, 3–5	544 (16%)	65 (14%)	409 (17%)	73 (14%)	65 (8%)	0 (0%)	331 (19%)	0 (0%)
MOCA total	-	-	26·1 (3·0 to 30·0)	25.7 (8·0 to 30·0)	24·5 (12·0 to 30·0)	23.7 (20·0 to 27·0)	24·9 (9·0 to 30·0)	26.7 (22·0 to 30·0)
MMSE total	30·0 (0·0 to 30·0)	28.9 (21·0 to 30·0)	-	-	27·4 (18·0 to 30·0)	27.0 (24·0 to 30·0)	-	-
SEADL	91·5 (60·0 to 100·0)	91.2 (70·0 to 100·0)	89·1 (0·0 to 100·0)	87.5 (10·0 to 100·0)	-	-	88·2 (20·0 to 100·0)	92.2 (80·0 to 100·0
UPSIT total	-	-	21·4 (0·0 to 40·0)	22.6 (0·0 to 40·0	-	-	19·7 (3·0 to 37·0)	21.3 (15·0 to 31·0
RBD total	-	-	4·4 (0·0 to 13·0)	4.0 (0·0 to 13·0	-	-	4·7 (1·0 to 13·0)	4.1 (2·0 to 7·0
ESS total	-	-	-	-	-	-	6·8 (0·0 to 24·0)	5.3 (0·0 to 12·0

ii) Large-scale disease severity meta-analysis across multiple PD clinical outcomes

I conducted a large-scale European and Ashkenazi Jewish genetic association analysis across all Parkinson's clinical outcomes available capturing motor features (MDS-UPDRS part II, MDS-UPDRS part III), cognitive features (MMSE, MoCA), disability (SEADL), disease severity (H&Y), and non-motor features (MDS-UPDRS part I, UPSIT, RBD) using GLM for baseline data. Subsequently, I meta-analysed genetic variant summary statistics for each clinical outcome. The meta-analysis was performed for each ancestry group independently. The sample size for each GWAS depended on data availability across clinical outcomes (**Table 20**). I did not detect significant genomic inflation for any of the genetic-association studies.

For the European ancestry meta-analyses, I found statistically significant genetic associations with non-motor features (MDS-UPDRS part I, UPSIT), motor features (MDS-UPDRS part II), and disability (SEADL) (**Figure 21**) (**Table 22**). All significant associations were intragenic.

***Figure 21**. Manhattan plots of the GWAS meta-analyses with significant associations.*



Meta-analysis for a) Upsit score b) SEADL, c) UPDRS-I, d) UPDRS-II. The x-axis represents the chromosome, and the position of each variant in the meta-analysis. The y-axis shows the two-sided P-value in the -log10 scale. Genome-wide significance is set at a P-value of 5e-08, and is represented by the red line on the Manhattan plots.

***Table 22**. Table of the lead SNP for each significant LD block part of the meta-analysis, including variants with at least 30% availability across the multiple cohorts*

rsID	chr	A1	MAF	beta	se	P-value	Gene	func	outcome
rs6702348	1	G	0.012	-11.76	1.81	9.25e-11	*GPR137B*	intronic	SEADL
rs34637584	12	T	0.09	4.96	0.64	2.08e-10	*LRRK2*	exonic	UPSIT
rs76763715	1	C	0.002	-4.49	0.76	3.89e-09	*GBA1*	exonic	UPSIT
rs142137167	2	G	0.008	6.52	1.12	5.72e-09	*HECW2*	intronic	UPDRS II
rs181145947	9	T	0.005	6.16	1.11	3.20e-08	*AGTPBP1*	intronic	UPDRS I
rs11764231	7	G	0.005	7.34	1.33	3.65e-08	*WDR86*	intronic	UPDRS II

For each lead SNP, I provide which is the closest gene. Because all lead SNPs were intragenic, I mapped each SNP falling on each gene boundary. P-value, two-sided P-value of association from meta-analysis.

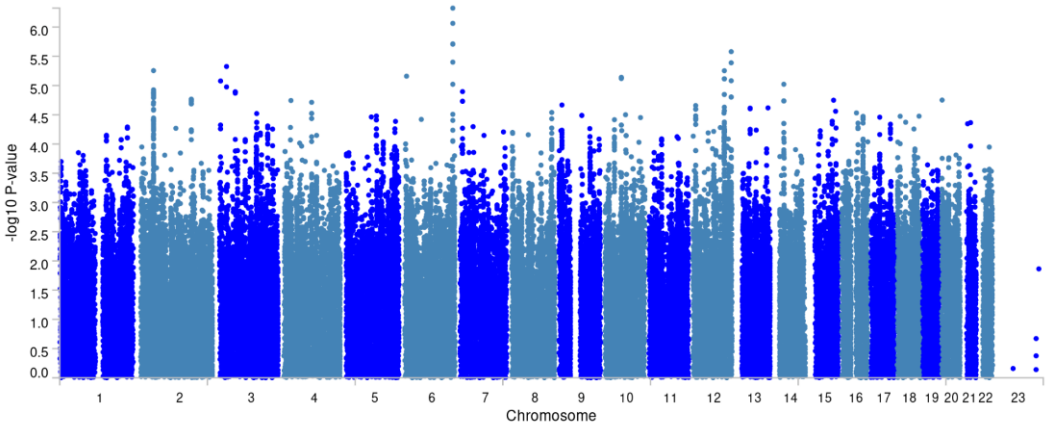
In addition, I have made available a list of all the lead SNPs that reached nominal significance on each locus (P-value < 1e-6) and the genes that they were mapped onto for each individual clinical outcome **Table 23**.

***Table 23**. All nominal and significant associations from multi GLM GWAS.*

chr	A1	MAF	beta	se	gwasP	nearestGene	dist	func	study
5	G	0.11	2.26	0.46	1.90E-07	*AC008565.1*	83226	intergenic	UPSIT
1	G	0.08	0.11	0.02	4.98E-07	*ADH5P2*	80069	intergenic	HY
2	T	0.27	0.08	0.02	2.64E-07	*ATIC*	0	intronic	HY
4	G	0.03	0.18	0.04	5.92E-07	*SLC9B2*	0	intronic	HY
10	T	0.01	-3.24	0.60	5.07E-08	*RPL39P25*	10649	intergenic	MOCA
13	A	0.01	-3.77	0.71	9.75E-08	*XPO4*	0	intronic	MOCA
20	G	0.09	-3.53	0.65	5.47E-08	*NTSR1*	0	intronic	SCHWAB
13	A	0.02	3.53	0.65	5.16E-08	*NALCN*	0	intronic	UPDRS_I
2	G	0.01	6.53	1.12	5.72E-09	*HECW2*	0	intronic	UPDRS_II
3	G	0.13	1.72	0.33	1.52E-07	*TSC22D2*	10385	intergenic	UPDRS_III
3	C	0.22	1.36	0.27	4.28E-07	*TSC22D2*	14022	intergenic	UPDRS_III
10	A	0.01	8.54	1.72	7.22E-07	*ZNF438*	16990	intergenic	UPDRS_III
16	G	0.23	-1.37	0.26	1.55E-07	*TEKT5*	0	intronic	UPDRS_III
20	A	0.45	1.11	0.22	7.64E-07	*RP11-137F15.1*	16380	intergenic	UPDRS_III

The number of participants of Ashkenazi Jewish ancestry was sufficiently large in the AMP-PD Unified and GP2 cohorts to perform a large-scale GWAS. However, for the Ashkenazi Jewish samples, I only performed a meta-analysis for the Hoehn and Yahr and MDS-UPDRS part III GWASs due to clinical data availability (**Table 20**). I did not find any statistically significant association in the AJ HY and MDS-UPDRS meta-analyses. Interestingly, I found a nominally significant LD block for the Hoehn and Yahr disease severity meta-analysis (**Figure 22**). The lead SNP (**rs510791**; β = 0.25; SE = 0.05; P-value = 4.74e-07), is an intronic variant at the ***PACRG*** gene, a gene next to the *PARKIN* gene, which is associated with autosomal recessive juvenile PD. *PACRG* and *PARKIN* are co-regulated in multiple tissues and share a bi-directional promoter. In addition, the *PARKIN* co-regulated protein is a component of Lewy bodies in PD patients [387,388]. This variant was not associated with the HY stage in Europeans (**rs510791**; β = 0.0036; SE = 0.0123 ; P-value = 0.7728), which suggests that some genetic determinants of disease severity might be ancestry-specific.

Figure 22*. AJ Hoehn and Yahr meta-analysis*

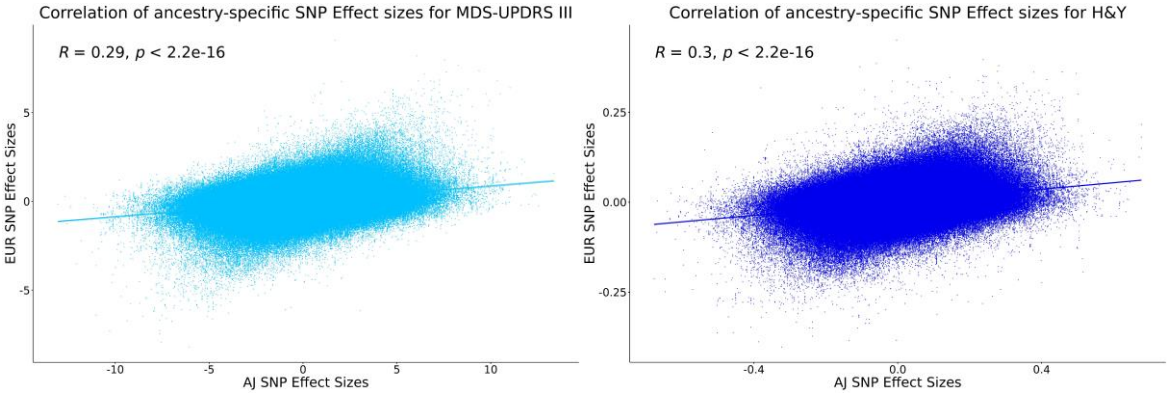


The x-axis represents the chromosome, and the position of each variant in the meta-analysis. The y-axis shows the two-sided P-value in the -log10 scale.

Finally, to estimate to what extent genetic determinants might contribute to disease severity outcomes in PD, we derived Pearson correlation coefficients for the SNP effect sizes between the AJ and EUR ancestry meta-analyses of the Hoehn and Yahr and the MDS-UPDRS part III outcomes, which were available for both ancestry groups (**Figure 23**). Overall, the correlation results suggest that the genetic makeup that contributes to PD severity as measured by multiple clinical outcomes, is ancestry-

specific. However, the correlation result across ancestries suggests that homogeneous effects exist.

***Figure 23**. Pearson correlation of Effect sizes between AJ and EUR MDS-UPDRS III (left) and H&Y (right) meta-analyses.*

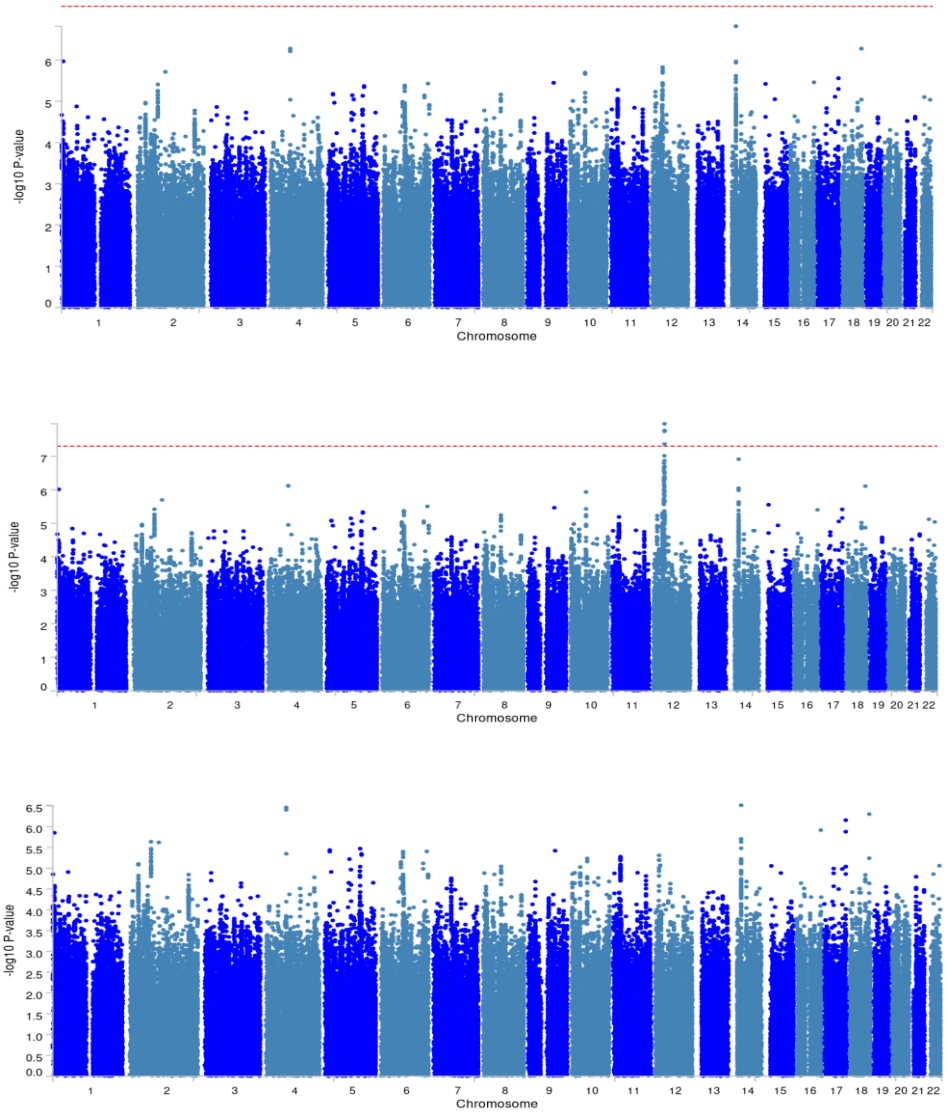


iii) *LRRK2* G2019S and *GBA1* N370S are the main genetic determinants of the olfactory impairment that arise in PD patients of European ancestry

The two significant LD blocks in the disease severity meta-analysis for the olfaction UPSIT assessment corresponded to the *GBA1* locus on chromosome 1 and the *LRRK2* locus on chromosome 12, respectively (**Table 22**). The lead variant at the *GBA1* locus (rs76763715, also known as **N370S**), was associated with a decrease in the average UPSIT score (β = -4.49 , SE = 0.76 , P-value = 3.89e-09). The lead SNP at the *LRRK2* locus (rs184460887) was associated with an increase in the average UPSIT score β = 4.96; SE = 0.77; P-value = 1.06e-10). Similarly, the second most significant association (rs34637584, also known as **G2019S**) was associated with an increase in the average UPSIT score (β = 4.07 , SE = 0.64 , P-value = 2.08e-10).

To further understand the pathway associated with the smell impairment involving *LRRK2* and *GBA1*, I adjusted the UPSIT total score GWAS, adding G2019S, N370S and G2019S or N370S mutation carrier status as mediators in the genome-wide scale multiple regression model on the meta-analysis results from the European-based UPSIT meta-analysis. I did not find any novel genetic association with the UPSIT outcome (**Figure 24**).

***Figure 24**. Association between genetic variants at the LRRK2 locus and Hoehn and Yahr outcome.*



a) Manhattan without any covariate b) Manhattan with GBA1 N370S as covariate c) Manhattan with LRRK2 G209S as covariate.

N307S remained significantly associated after adjusting on G2019S, as well as G2019S remained significant after adjusting on N307S. However, the *LRRK2* locus spans more than 1Mb away from the top lead SNP, and the variance captured by this locus on the UPSIT score seems to be more complex than just by the effect of one single pathogenic mutation. For instance, on our G2019S conditional GWAS, the top lead SNP in chromosome 12 (rs185993818), remained nominally associated with the UPSIT outcome (β = 3.5 ; SE = 0.94; P-value = 2e-4). Finally, adjusting the UPSIT GWAS on both N370S and G2019S removed all the significance from all variants at both the *GBA1* and the large *LRRK2* loci (**Table 24**).

***Table 24**. G2019S and N370S, conditional GWAS.*

SNP	CHR	Position	Condition	MAF	Effect	SE	P-value
rs34637584	**12**	40340400	**N370S**	**0.09**	**3.20**	**0.68**	**2.52E-06**
rs76763715	1	155235843	N370S	0.04	-8.00	4.00	4.00E-02
rs185993818	**12**	41691672	**N370S**	**0.06**	**3.93**	**0.81**	**1.37E-06**
rs34637584	12	40340400	G2019S	0.09	7.14	2.82	1.00E-02
rs76763715	**1**	155235843	**G2019S**	**0.04**	**-3.30**	**0.81**	**4.59E-05**
rs185993818	**12**	41691672	**G2019S**	**0.06**	**3.49**	**0.94**	**2.00E-04**
rs34637584	12	40340400	G2019S and N370S	0.09	7.00	2.80	1.00E-02
rs76763715	1	155235843	G2019S and N370S	0.04	-7.56	3.90	5.00E-02
rs185993818	**12**	41691672	**G2019S and N370S**	**0.06**	**2.03**	**0.95**	**2.00E-03**

I applied the same principle to explore the effect of adding *GBA1* and *LRRK2* mutation carrier status as mediators in the association of genetics with the remainder of the baseline outcomes (**Table 20**). Interestingly, I found a significant association between Hoehn and Yahr stage and the SNPs at the *LRRK2* locus, and this association was not uncovered on the simpler model without *GBA1* N370S mutation carrier status as a mediator. The lead SNP was rs991584002 (β = 0.31; SE = 0.05; P-value = 1.1E-8). GS019S was also present on this LD block and nominally associated (β = 0.26; SE = 0.05; P-value = 4.87E-5).

In addition, I explored all independent SNPs associated with the UPSIT score at the nominal significance level (P-value 1e-6) and annotated them using the FUMA web platform. In brief, I mapped all independent SNPs to HUGO symbols based on positional distance. Moreover, based on LD, I inferred missing SNPs on my data from the 1000 Genomes reference panel, as I expect that SNPs in high LD with tagging SNPs, to also have an inflation in the test statistics. This provides additional information as one independent SNP in LD with the tag SNP that was not present in the input data for the genetic association study, might be mapped to a different gene, giving us further insights into the functional implication of the locus of interest (**Table 25**).

***Table 25**. Nominal significant association with the baseline UPSIT score*

HUGO	chr	nIndSigSNPs	minGwasP	posMapSNPs	posMapMaxCADD
GBA1	1	1	3.89E-09	1	23.7
FDPS	1	1	4.26E-09	1	2.417
RUSC1	1	1	4.26E-09	1	2.417
MSTO1	1	1	2.84E-07	1	14.66
UBQLN4	1	1	2.84E-07	0	0
SMG5	1	1	2.84E-07	0	0
TMEM79	1	1	2.84E-07	0	0
IL6R	1	1	NA	1	0.342
SHE	1	1	NA	1	1.782
TDRD10	1	1	NA	1	1.782
KCNN3	1	1	NA	5	14.05
PYGO2	1	1	NA	1	18.51
SHC1	1	1	NA	1	18.51
CKS1B	1	1	NA	1	18.51
ADAM15	1	1	NA	1	4.595
EFNA4	1	1	NA	1	4.595
EFNA3	1	1	NA	1	4.595
ASH1L	1	1	NA	4	9.391
GON4L	1	1	NA	2	2.504
SYT11	1	1	NA	1	7.134
RIT1	1	1	NA	1	7.134
SCN1A	2	2	1.79E-06	38	10.76
SCN9A	2	1	1.79E-06	0	0
TTC21B	2	2	3.05E-06	0	0
GALNT3	2	1	5.43E-06	0	0
SCN7A	2	1	1.47E-05	0	0
LSAMP	3	1	1.66E-06	2	1.323
FBXW11	5	1	5.54E-06	0	0
PDZRN4	12	4	1.33E-09	11	13.31
ABCD2	12	1	4.09E-09	1	0.213
CNTN1	12	3	1.18E-08	70	12.3

HUGO	chr	nIndSigSNPs	minGwasP	posMapSNPs	posMapMaxCADD
LRRK2	12	2	4.27E-08	3	6.718
GXYLT1	12	3	7.58E-08	1	1.031
C12orf40	12	3	3.89E-07	5	1.904
SLC2A13	12	3	3.89E-07	6	1.904
MUC19	12	2	4.29E-07	6	10.41
KIF21A	12	3	1.07E-06	2	10.56
DHX37	12	1	1.18E-06	36	10.63
BRI3BP	12	1	1.18E-06	0	0
CPNE8	12	4	1.23E-06	5	10.28
TMEM117	12	1	1.29E-06	7	11.66
ADAMTS20	12	1	NA	2	2.058
IRAK4	12	1	NA	1	0.699
NELL2	12	1	NA	5	4.172
DBX2	12	1	NA	1	0.419
TEX101	19	1	4.39E-06	0	0

iv) Validation of the *LRRK2* and the *GBA1* associations with the UPSIT score in Tracking Parkinson's

I found high variability in the *GBA1* N370S and *LRRK2* G2019S MAFs between European cases from AMP-PD Unified and Tracking Parkinson's cohorts (4% vs 0.3% for N370S and 6% vs 0.6% for G2019S, respectively). Therefore, these variants were excluded during the long-gwas quality control framework of the TPD cohort.

To validate our findings in the Tracking Parkinson's data source, I accessed Sanger sequencing patient-level genetic data and characterised each patient based on G2019S and N370S mutation carrier status, for a total of 2000 PD patients. I adjusted a GLM model on sex, standardised age at diagnosis, and the first 3 PCs, as well as the two pathogenic variants separately. Promisingly, I found **G2019S** approaching significance in the association model against the UPSIT total score and the directionality of the effect was consistent with that from the UPSIT European meta-analysis (**β = 6.07, SE = 3.25, P-value = 0.052)**. I did not find a significant association between **N370S** and the UPSIT score, but I found the directionality of the effect to be

consistent with the findings from the AMP-PD Unified European cohort (**β = -1.88, SE = 2.17, P-value = 0.4**). Interestingly, I found a significant association between *GBA1* N370S status and Hoehn and Yahr stage and MDS-UPDRS III total score (**Table 26**).

***Table 26**. GBA1 N370S GLM summary statistics across multiple outcomes from Tracking Parkinson's*

Variable	Estimate	std.error	statistic	p.value
HY	0.57	0.20	2.92	0.0035
UPDRSIII	11.41	4.29	2.66	0.0079
BFI	2.80	1.58	1.77	0.0772
ESS	2.38	1.60	1.48	0.1386
GASTRO	-0.26	0.19	-1.34	0.1819
RBD	1.29	1.10	1.18	0.2392
Leeds dep	0.96	1.05	0.91	0.3619
UPSIT	-1.88	2.17	-0.87	0.3871
UPDRSII	1.75	2.15	0.81	0.4157
UPDRSI	-1.18	1.76	-0.67	0.5029
UPDRSIV	-0.24	0.61	-0.39	0.6967
PDQ8	0.61	1.59	0.39	0.6998
NMSS	-2.44	10.67	-0.23	0.819
Leeds anx	0.22	1.17	0.19	0.8518
PDSS	1.33	7.73	0.17	0.864
LEDD	8.62	67.28	0.13	0.898
MOCA adj	0.12	1.07	0.12	0.9075
MOCA	-0.02	1.10	-0.02	0.9873

BFI=Brief Fatigue Inventory; **ESS**=Epworth Sleepiness Scale; **GASTRO**=Gastrointestinal symptoms; **HY**=Hoehn and Yahr; **LEDD** = Levodopa Equivalent Daily Dose; **Leeds dep**=Leeds scale to assess depression; **Leeds anx**=Leeds scale to assess anxiety; **MOCA**=Montreal Cognitive Assessment; **MOCA adj**=Montreal Cognitive Assessment adjusted score; **NMSS**=Non-Motor Symptoms Scale for Parkinson's Disease; **PDQ8**=Parkinson's Disease Questionnaire-8; **PDSS**=Parkinson's Sleep Scale; **RBD**=REM Sleep Behavior Disorder; **UPDRSI**=UPDRS scale part I; **UPDRSII**=UPDRS scale part II; **UPDRSIII**=UPDRS scale part III; **UPDRSIV**=UPDRS scale part IV; **UPSIT**=University of Pennsylvania Smell Identification Test

I also found *LRRK2* G2019S mutation status to be associated with a higher MDS-UPDRS IV total score (**Table 27**).

***Table 27**. LRRK2 G2019S GLM summary statistics across multiple outcomes from Tracking Parkinson's*

Variable	Estimate	std.error	statistic	p.value
UPDRSIV	1.64	0.78	2.11	0.0346
UPSIT	6.07	3.25	1.87	0.0518
PDQ8	2.48	1.95	1.28	0.2021
UPDRSIII	6.81	5.42	1.26	0.2092
BFI	2.77	2.51	1.11	0.2682
GASTRO	-0.26	0.24	-1.08	0.2791
HY	0.20	0.25	0.79	0.4315
ESS	1.48	2.03	0.73	0.4661
Leeds dep	0.88	1.28	0.68	0.4957
Leeds anx	0.87	1.43	0.61	0.5427
UPDRSI	1.08	2.16	0.50	0.6166
NMSS	4.38	11.72	0.37	0.7086
RBD	-0.39	1.39	-0.28	0.7778
LEDD	-18.05	86.87	-0.21	0.8354
MOCA adj	0.18	1.44	0.12	0.9016
PDSS	-0.73	9.48	-0.08	0.9383
UPDRSII	0.15	2.64	0.06	0.9559
MOCA	-0.02	1.47	-0.02	0.9868

BFI=Brief Fatigue Inventory; **ESS**=Epworth Sleepiness Scale; **GASTRO**=Gastrointestinal symptoms; **HY**=Hoehn and Yahr; **LEDD** = Levodopa Equivalent Daily Dose; **Leeds dep**=Leeds scale to assess depression; **Leeds anx**=Leeds scale to assess anxiety; **MOCA**=Montreal Cognitive Assessment; **MOCA adj**=Montreal Cognitive Assessment adjusted score; **NMSS**=Non-Motor Symptoms Scale for Parkinson's Disease; **PDQ8**=Parkinson's Disease Questionnaire-8; **PDSS**=Parkinson's Sleep Scale; **RBD**=REM Sleep Behavior Disorder; **UPDRSI**=UPDRS scale part I; **UPDRSII**=UPDRS scale part II; **UPDRSIII**=UPDRS scale part III; **UPDRSIV**=UPDRS scale part IV; **UPSIT**=University of Pennsylvania Smell Identification Test

v) Elucidating the molecular implication of the olfactory impairment in PD.

It has recently been described that hyposmic PD patients were predominantly positive for the α-synuclein seed amplification assay (SAA) [389]. Interestingly, the rate of α-synuclein SAA positivity was shown to be decreased among *LRRK2* mutation carriers, and this was reduced even further in *LRRK2* mutation carriers without olfactory impairment [389]. This is consistent with the finding that PD associated with *LRRK2*

pathogenic variants can present without synucleinopathy at autopsy. However, in *LRRK2* mutation carriers with an α-synuclein SAA positive result, the proportion of hyposmia was high (75%). Similarly, more than 90% of *GBA1* PD mutation carriers with a positive α-synuclein assay were found to have hyposmia. Altogether, this clearly suggests a central role of *LRRK2* and *GBA1* in the hyposmia manifestation in PD patients, which might lead to the subsequent α-synuclein pathology as measured by α-synuclein SA.

Our results from the UPSIT meta-analysis support a role for *LRRK2* and *GBA1* in determining olfactory performance in PD. Previously, it has been reported that α-synuclein overexpression could result in α-synuclein aggregation [4]. I did not find any nominal significant association to be a cis-eQTL for *SNCA* expression. In this section I explored whether G2019 and N370S status explain differences in the expression or accumulation of α-synuclein, which could be a marker of α-synuclein aggregation and pathology. This would provide further support in the *LRRK2* and *GBA1* mutations role controlling α-synuclein aggregation manifesting with hyposmia.

One major strength of the AMP-PD unified cohorts is the availability of matched clinical and multi-omics data. Likewise, I could assess the hypothesis of N370S and G2019S PD pathogenic mutations leading to a prominent *SNCA* overexpression which could lead to aggregation. In addition, I explored whether there could be any other biomarker contributing to pathology as a consequence of any of the *GBA1* and *LRRK2* pathogenic mutations. I used data from AMP-PD CSF Untargeted protein measures.

I matched samples with transcriptomics and genetics and stratified the data based on G2019S and N370S mutation carriers leading to a total of 1120 PD patients available for the differential expression analysis. Out of the 1120 PD patients with matched genetic, clinical and whole-transcriptome data, 75 were *LRRK2* G2019 mutation carriers, and 135 were *GBA1* N370S mutation carriers. I found the expression of *SNCA* and *SNCA-AS1* to be nominally significant in the differential expression analysis based on G2019 status (*SNCA* logFC = 0.2; P-value = 0.03) (*SNCA-AS1* logFC = 0.3; P-value = 0.03). However, neither *SNCA* or its antisense form differential expression reached significance after applying Bonferroni correction. In addition, I found three genes to be differentially expressed based on G2019S status (**Table 28**) (**Figure 25**).

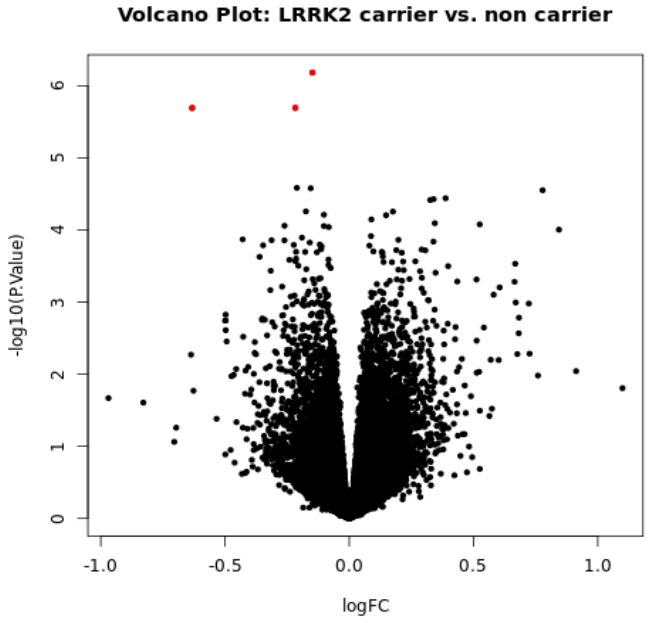
I did not find any significant differential expression based on N370S status. *SNCA* did not reach nominal significance (*SNCA* logFC = -0.21; P-value = 0.1).

***Table 28**. Differential expression significant results and SNCA nominal significant results*

gene_name	gene_type	chr	logFC	t	P.Value	adj.P.Val	AvgExpr_ Control	AvgExpr_Case
AC022150.4	sense_intronic	19	-0.14	-5	6.58E-07	1.70E-02	3.22	3.05
AC090630.1	lincRNA	12	-0.63	-4.78	2.04E-06	1.70E-02	-2.05	-2.42
ZC3H11B	processed_pseudogene	1	-0.22	-4.78	2.03E-06	1.80E-02	0.21	0.07
SNCA-AS1	antisense	4	0.27	2.1	3.60E-02	5.30E-01	-2.13	-2
SNCA	protein_coding	4	0.2	2.19	2.90E-02	5.10E-01	9.92	10.2

LogFC = the mutation carriers vs non-carriers protein fold change in logarithmic scale; t = test-statistic of the fold change; P.value = Significance of the test-statistic; Adj.P.Val = P.value adjusted by Bonferroni correction. AvgExpr_Control = Average protein expression in the non-carrier group; AvgExpr_Case = Average protein expression in the carriers group.

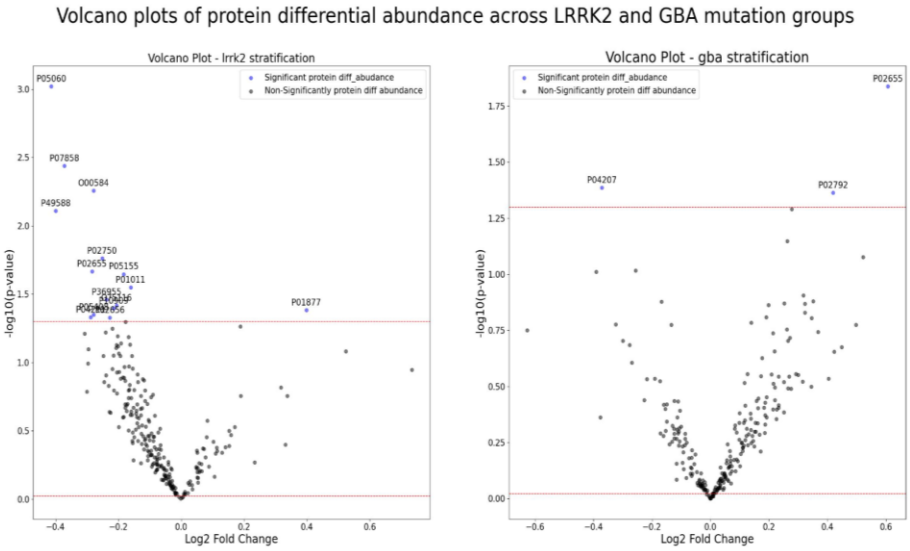
***Figure 25**. Volcano plot of differentially expressed genes based on G2019S status*



The X axis represents the log2 fold change between G2019S carriers and non-carriers groups. The y-axis represents the -log10(P-value). Each dot is the fold change of the normalised measure of a transcript.

Similarly, I matched samples with proteomics and genetics and stratified the data based on *LRRK2* and *GBA1* mutation carriers leading to a total of 357 PD patients available for the differential abundance analysis. Of these, 27 were *LRRK2* G2019 mutation carriers, and 9 were *GBA1* N370S mutation carriers. I did not find any patient carrying both G2019S and N370S mutations. α-synuclein abundance was not available among the proteins with abundance records. I found proteins whose abundance was significantly different between the mutation carriers and non-carriers groups (**Figure 26**).

Figure 26*. Volcano plot of differential abundant proteins between LRRK2 G2019S (left) and GBA1 N370S (right) mutation carriers versus non-carriers*



The red dashed line represents the significance threshold of P-value = 0.05 in -log10 scale. The X axis represents the log2 fold change between carriers and non carriers groups. The y-axis represents the -log10(P-value). Each dot is the fold change of the normalised measure of a protein.

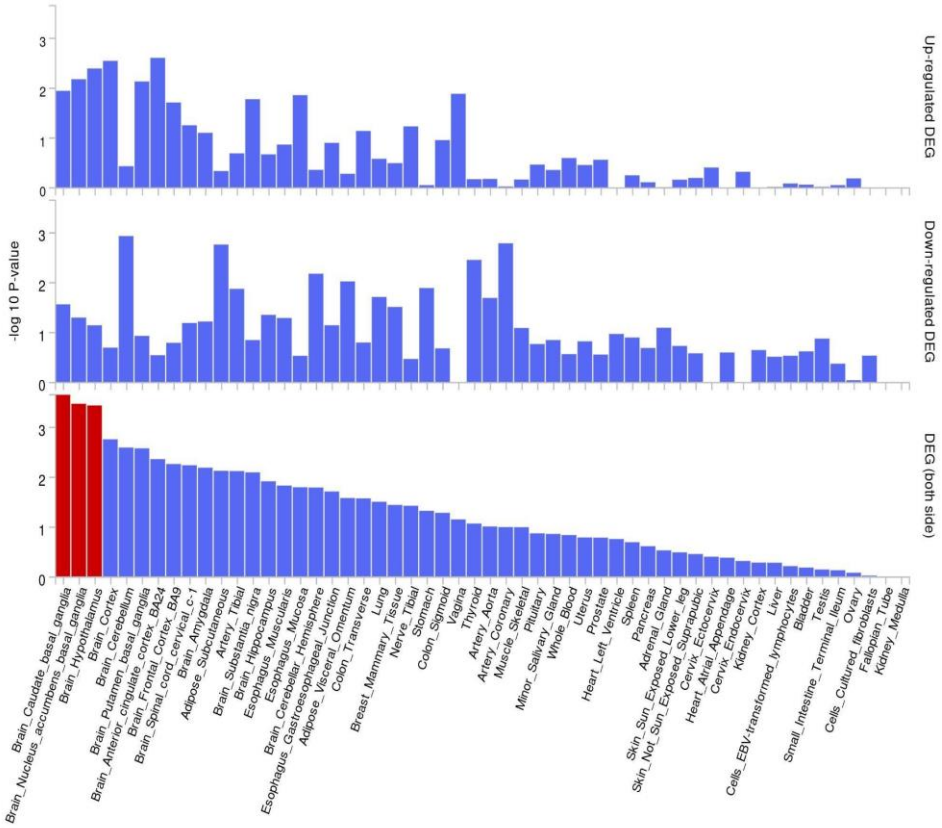
Interestingly, when stratifying based on *GBA1* status, we found the most differentially abundant protein to be P02655 (ApoC II), a protein that belongs to a protein cluster together with ApoE and ApoC-I. Because they share a protein cluster, they also share a generalisable function. Based on *LRRK2* stratification, we found P05060 (Secretogranin-1) as the most significantly differentially abundant protein. We have summarised UniProt codes with the corresponding HUGO symbols and a description of the protein function in **Table 29**.

***Table 29**. Differentially abundant proteins.*

UniProt	Grouping	coef	std err	z	p_value	0.025	0.975
P02655	*GBA1*	-4.8329	1.837	-2.631	0.009	-8.433	-1.232
P02749	*GBA1*	-4.6593	2.165	-2.152	0.031	-8.903	-0.416
P02792	*GBA1*	-3.3194	1.683	-1.972	0.049	-6.619	-0.02
P01011	*GBA1*	7.0553	3.625	1.946	0.052	-0.05	14.161
P19827	*GBA1*	-2.46	1.283	-1.918	0.055	-4.974	0.054
P07858	*LRRK2*	3.6779	1.346	2.732	0.006	1.039	6.317
P05060	*LRRK2*	3.83	1.39E+00	2.752	0.006	1.103	6.559
P02655	*LRRK2*	3.50	1.41E+00	2.475	0.013	0.729	6.269
P05090	*LRRK2*	4.66	1.97E+00	2.371	0.018	0.808	8.513
Q8NE71	*LRRK2*	2.37	1.03E+00	2.296	0.022	0.347	4.399
P05155	*LRRK2*	5.19	2.28E+00	2.275	0.023	0.72	9.668
P02656	*LRRK2*	3.17	1.51E+00	2.097	0.036	0.207	6.127
P00738	*LRRK2*	-1.38	6.57E-01	-2.094	0.036	-2.664	-0.088
P49588	*LRRK2*	2.27	1.11E+00	2.039	0.041	0.088	4.456
O00584	*LRRK2*	3.31	1.66E+00	1.993	0.046	0.055	6.563

Knowing the tissue and the cell type burden of the PD olfaction impairment meta-analysis lead SNPs gives us new insights into how disease develops and progresses based on the hypothetical central role of *LRRK2*, *GBA1* and related proteins in α-synuclein deposition. I accessed FUMA, and conducted enrichment analyses on the UPSIT severity GWAS against all 54 tissues represented in GTEx. I found an enrichment of either down-regulated or up-regulated genes in basal ganglia (brain caudate and the nucleus accumbens), as well as a significant enrichment in the hypothalamus (**Figure 27**).

***Figure 3**. Tissue specificity analysis based on the differentially expressed genes (DEG) for each tissue from the GTEx Consortium*



DEGs were pre-calculated based on a two-sided t-test for gene expression values of any one of the tissues against all others. P-values < 0.05 after Bonferroni correction and an absolute log fold change > 0.58 were defined as differentially expressed genes in a given tissue compared to the rest. Apart from the two-sided DEG analysis, the analysis of the test-statistics was used to derive an up-regulated DEG and down-regulated measure of specific trend of differential enrichment. Finally, a hypergeometric test was used to test the input gene set nominated from MAGMA gene-set analysis against each of the tissue level DEG sets. From top to bottom, Up regulated DEG, Down regulated DEG, Both sides DEG.

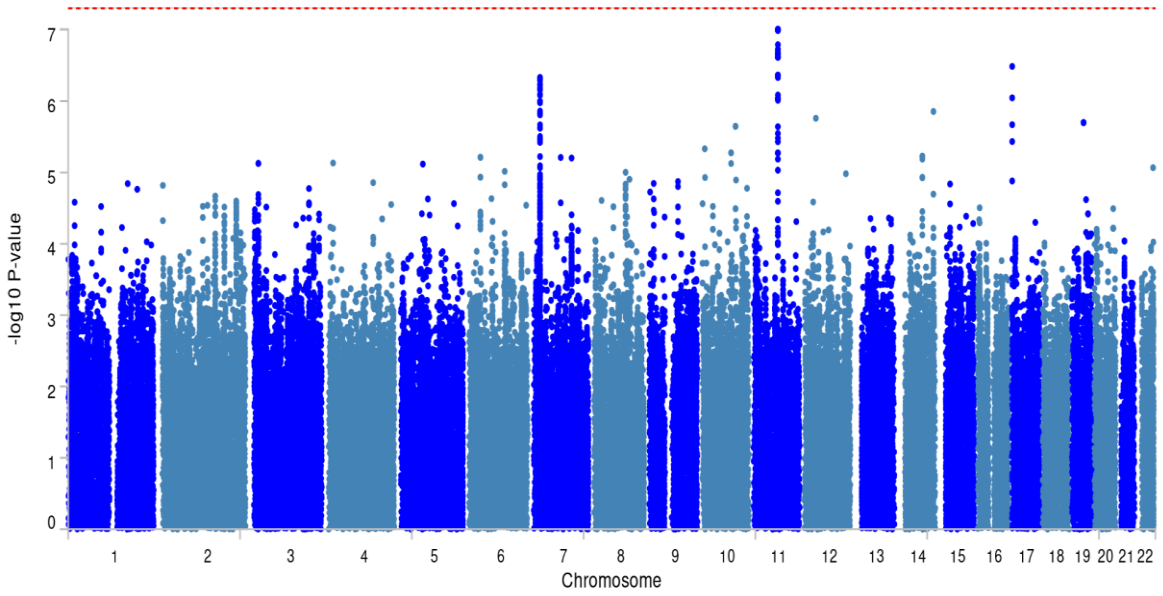
vi) Nominating genetic determinants of SAA independent of *LRRK2* and *GBA1*

Following up on the hypothesis that *LRRK2* and *GBA1* mutations have a central role in controlling α-synuclein pathology, whose clinical manifestations include hyposmia, we conducted a GWAS on a-syn-SAA to investigate if we could further implicate *GBA1*

and *LRRK2* on α-synuclein deposition. In addition, we were interested in knowing if there are any other genetic determinants aside from the *LRRK2*/*GBA1-related* autophagy-lysosomal pathway).

We accessed PPMI PD samples available from the AMP-PD Unified cohort with available SAA data and that passed long-gwas quality control steps. We conducted a logistic regression GWAS using SAA binary status as the outcome. We did not find support for *LRRK2* or *GBA1* variants associated with the α-synuclein pathology as measured by SAA (**Figure 28**). Of note, N370S, G2019S, and the G2019S-independent top lead SNP at the *LRRK2* locus, did not reach genome-wide significance (**N370S** log(OR) =-0.79 SE = 0.39; P-value = 0.04); (**G2019S** log(OR) = 0.90; SE = 0.30; P-value = 2e-3); (**rs185993818** OR = 0.39 SE = 0.32; P-value = 0.23).

***Figure 28**. Manhattan plot of the SAA GWAS.*



The x-axis represents the chromosome, and the position of each variant in the meta-analysis. The y-axis shows the two-sided P-value in the -log10 scale. Genome-wide significance is set at a P-value of 5e-08, and is represented by the red line on the Manhattan plots.

Instead, I found three LD blocks approaching genome-wide significance in chromosomes 7, 11 and 17. I have annotated the lead SNP on each locus in **Table 30**. I cross-checked the association of these SNPs in the UPSIT meta-analysis. None of the variants reached nominal significance (P < 1e-6). The variant at the *CRK* locus

was nominally associated at the significance threshold of P < 0.05. (*CRK* locus ;rs117985867 β = 1.04 ; SE = 0.42 ; P-value = 0.01), (*LOC124902707* locus; rs28370535 β = 0.63; SE =0.39 ; P-value = 0.11), and (*HDAC9* locus rs111978 β = 0.11; SE = 0.24 ; P-value = 0.64).

***Table 30**. Lead SNP for each LD block approaching nominal significance in the SAA GWAS.*

rsID	chr	pos	A1	MAF	beta	se	gwasP	nearestGene	func
rs111978	7	18271954	A	0.28	0.91	0.18	4.67E-07	*HDAC9*	intronic
rs28370535	11	71388977	G	0.08	1.35	0.25	9.84E-08	*LOC124902707*	ncRNA_intronic
rs117985867	17	1453350	G	0.04	1.36	0.27	3.26E-07	*CRK*	intronic

For each lead SNP, I provide which is the mapped Gene based on distance to it. Because all lead SNPs were intragenic, I mapped each SNP falling on each gene boundary. P-value, two-sided P-value of association from meta-analysis; nearestGene, the closest gene to the top lead SNP on each genomic risk locus; fun, the ANNOVAR annotated function of the top lead SNP.

In addition, I adjusted the SAA GWAS on *LRRK2*, *GBA1*, and *LRRK2* + *GBA1* mutation carriers status. I did not find a drop in significance on any of these three loci. This suggests that the three genes found to associate with SAA outcome are independent of the *LRRK2* and *GBA1* driven α-synuclein accumulation.

Finally, I also adjusted the UPSIT genetic study previously described based on SAA status. Of note, for this analysis, I could only run a GWAS and not a meta-analysis, since SAA status was only available for the AMP-PD cohort. Interestingly, adjusting the UPSIT outcome on SAA status led to a loss of significance at the *LRRK2* locus (rs185993818 β =3.03 SE = 1.04869; P-value =0.004) (G2019S β = 2.92; SE = 0.90; P-value = 0.001) locus but not the *GBA1* locus (N370S β = -4.03; SE = 0.96; P-value = 3.26e-05).

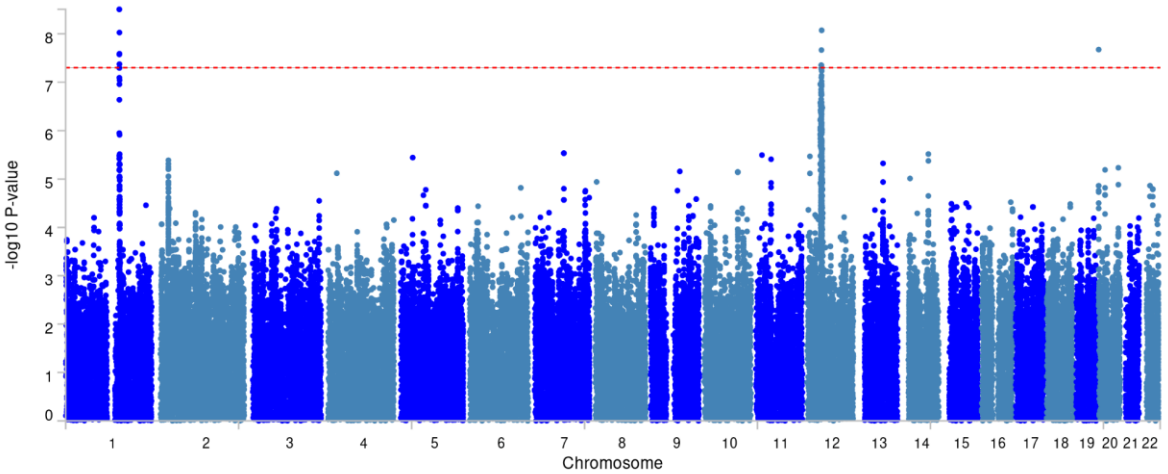
vii) Multi-ancestry analysis of the olfactory impairment reveals novel genetic markers

Recent work is shedding light into ancestry-specific risk factors of PD as well as gains in power to uncover novel associations based on multi-ancestry meta-analysis studies [82,390]. I wanted to explore whether performing a meta-analysis between summary

statistics of different ancestry groups, with a focus on uncovering homogeneous genetic effects, would increase the power to detect novel significant associations. Previously, I found the correlation between the effect sizes of the H&Y and MDS-UPDRS III meta-analyses in AJ and EUR to be modest (**Figure 23**). Therefore, I hypothesised that a fixed-effect meta-analysis would lead to an increase in power to uncover homogeneous genetic variants, primarily as a result of an increase of the final sample size and shared disease severity risk factors between AJ and EUR PD ancestry groups.

For this analysis, I was primarily interested in the UPSIT score, following up on results from Europeans previously described. I accessed the AJ UPSIT GWAS performed among AMP-PD data (**Table 20**). As expected, I also found the *LRRK2* and *GBA1* loci to be significant in the UPSIT GWAS in the AJ subset (**Figure 29**). Interestingly, I found the *GBA1* locus to have the most significant associations. The lead SNP at the *GBA1* locus was rs76763715 (β = -5.78; SE = 0.95; P-value = 3.11e-9). The most significant association at the *LRRK2* locus was rs184460887 (β = 4.85; SE = 0.82; P-value = 8.44e-09). G2019S was among the top hits (β = 5.01; SE = 0.89; P-value = 4.43e-08). In addition, I found another single intragenic variant reaching genome-wide significance at the ***RBCK1*** locus in chromosome 20 (lead SNP rs6051899; Effect = 3.75; SE = 0.65, P-value = 2.11e-8).

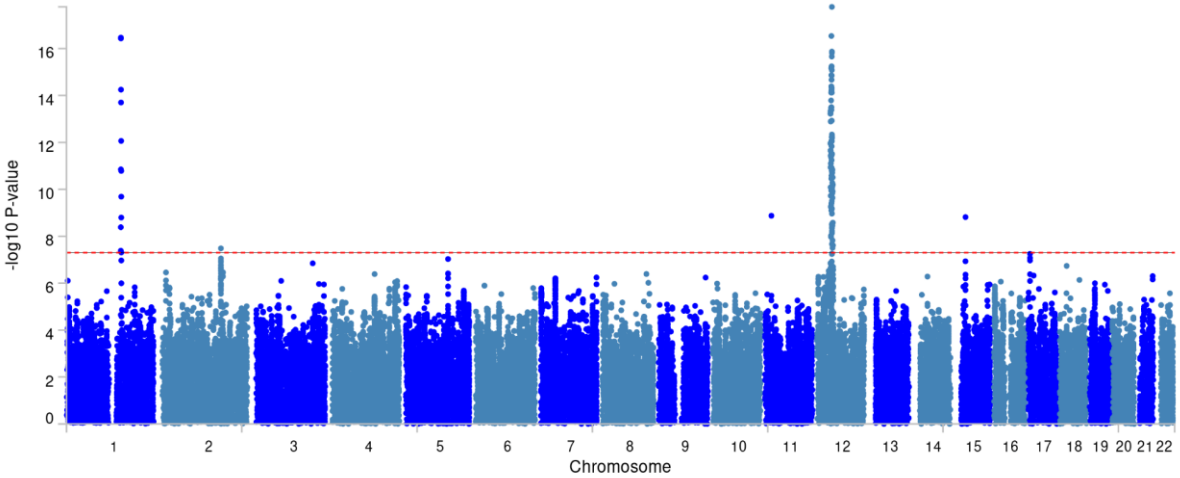
Figure 29. *AJ UPSIT AMP-PD GWAS.*



The x-axis represents the chromosome, and the position of each variant in the meta-analysis. The y-axis shows the two-sided P-value in the -log10 scale. Genome-wide significance is set at a P-value of 5e-08, and is represented by the red line on the Manhattan plots.

Then, I performed a multi-ancestry meta-analysis for the UPSIT results from the AJ and EUR clusters. I found **5** independent disease severity loci (**Figure 30**). I annotated the meta-analysis genomic risk loci using FUMA. I found a large inflation in test-statistics at the *GBA1* (top lead SNP = rs76763715 , Effect = -4.9, SE = 0.6, P-value = 3.45-17) and the *LRRK2* (top lead SNP = rs3463758; Effect = 4.96; SE = 0.64, P-value = 2.92e-17) loci, similar to the results of the UPSIT ancestry-specific GWASs. In addition, I found novel independent loci reaching genome-wide significance at the ***SERGEF***, ***SCN1A***, ***OTUD7A*** loci (**Table 31**).

Figure 30*. Manhattan plot of the UPSIT score multi-ancestry meta-analysis (EUR and AJ).*



The x-axis represents the chromosome, and the position of each variant in the meta-analysis. The y-axis shows the two-sided P-value in the -log10 scale. Genome-wide significance is set at a P-value of 5e-08, and is represented by the red line on the Manhattan plots.

Table 31*. Table of the lead SNP for each significant LD block multi-ancestry meta-analysis*

rsID	chr	pos	A1	MAF	beta	se	P	nearestGene	func
rs34637584	12	40340400	T	0.09	4.38	0.52	1.92e-17	*LRRK2*	exonic
rs76763715	1	155235843	T	0.05	-4.99	0.58	3.43e-17	*GBA1*	exonic
rs147669178	11	17914781	G	0.04	-3.92	0.65	1.34e-09	*SERGEF*	intronic
rs146931292	15	31720353	G	0.12	2.49	0.41	1.52e-09	*OTUD7A*	intronic
rs1960242	2	166990047	G	0.29	-1.55	0.28	3.29e-08	*SCN1A ; SCN1A-AS1*	intronic

Based on ANNOVAR annotations, most of the SNPs in LD with the independent significant SNPs at the nominated loci were intergenic (87%) or intronic (7%). A significant proportion of SNPs was also falling on intronic regions of non-coding RNAs (2%) (**Figure 31**). **Table 32** summarises all lead SNPs with a P-value <1e-6 and the gene they are mapped onto.

***Figure 31**. Proportion of the type of variants nominated by the EUR + AJ meta-analysis*

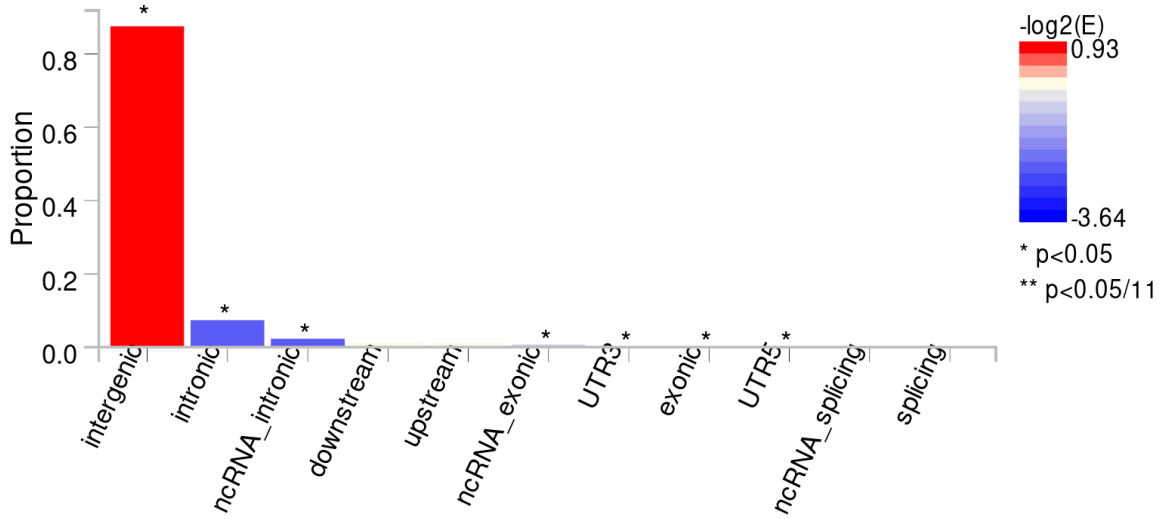


***Table 32**. Nominal association multi-ancestry meta-analysis UPSIT.*

rsID	chr	A1	MAF	beta	se	gwasP	nearestGene	dist	func
rs1078801	17	A	0.15	-1.95	0.35	5.6e-08	*ZMYND15*	0	intronic
rs4383741	5	A	0.11	2.24	0.42	9.3e-08	*CTD-2334D19.1*	83226	intergenic
rs308702	3	C	0.13	1.98	0.37	1.4e-07	*RP11-774I5.1*	38963	intergenic
rs10048227	18	T	0.11	2.08	0.40	1.8e-07	*RP11-863N1.4*	8354	intergenic
rs117811597	12	C	0.11	2.54	0.49	3e-07	*YARS2*	4500	intergenic
rs13034296	2	A	0.05	3.14	0.61	3.4e-07	*AC104801.1*	0	ncRNA_intronic
rs116676008	2	A	0.05	-3.11	0.61	3.5e-07	*SNORA51*	1814	intergenic
rs116984338	8	C	0.03	3.51	0.69	4.0e-07	*LINC00964*	0	ncRNA_intronic
rs9884655	4	C	0.5	-1.30	0.25	4.1e-07	*RP11-236P13.1*	19526	intergenic
rs2323561	17	T	0.03	4.08	0.81	4.7e-07	*CDRT8*	26333	intergenic

rsID	chr	A1	MAF	beta	se	gwasP	nearestGene	dist	func
rs73239719	21	G	0.04	-3.23	0.64	5.0e-07	*ITGB2*	0	intronic
rs8011258	14	C	0.44	-1.31	0.26	5.2e-07	*RP11-662J14.2*	51529	intergenic
rs71534222	12	C	0.04	-3.32	0.66	5.3e-07	*PRMT8*	0	intronic
rs73061198	12	G	0.07	2.43	0.49	5.4e-07	*RERGL*	105162	intergenic
rs11761517	7	C	0.44	1.29	0.26	5.7e-07	*ESYT2*	0	intronic
rs12340454	9	G	0.02	4.57	0.92	5.7e-07	*LAMC3*	30679	intergenic
rs255016	7	A	0.42	1.29	0.26	6.2e-07	*AC005022.1*	56932	intergenic
rs62093526	18	T	0.05	-3.06	0.62	7.e-07	*RP11-108P20.4*	0	ncRNA_intronic
rs112805272	1	G	0.14	-1.80	0.37	7.8e-07	*TP73*	0	intronic
rs116123005	3	T	0.03	3.85	0.78	8.0e-07	*FOXP1*	0	intronic
rs6552740	4	T	0.35	-1.35	0.27	8.1e-07	*RP11-616K6.1*	4127	intergenic
rs60050831	16	C	0.05	3.11	0.63	8.6e-07	*ACSM5*	1968	intergenic
rs4931112	12	G	0.12	1.95	0.40	9.0e-07	*RP11-977P2.1*	49205	intergenic
rs78902372	4	C	0.16	-1.74	0.36	9.1e-07	*RP11-84H6.1*	65668	intergenic
rs74452013	8	T	0.03	-3.58	0.73	9.6e-07	*ASAP1*	51277	intergenic

d) Discussion

In this study, we have introduced long-gwas, a tool that democratises genetic studies by automating all the steps involved in the integration of clinical and genetic data, from data pre-processing, to accurate quality control, and then GWAS using different settings. Here, we demonstrate the use of the long-gwas workflow to study the severity of PD, as a proof of concept. To do so, we have explored a wide-range of clinical outcomes in two different ancestry groups, Ashkenazi Jewish and Europeans.

The use of long-gwas has allowed us to identify two major genetic determinants for hyposmia, as measured by the UPSIT scale. Our results from the UPSIT meta-analysis revealed that *LRRK2* G2019S and *GBA1* N370S modulate olfactory performance in PD. *LRRK2* G2019S was found to be associated with better olfaction, whereas *GBA1* N370S was associated with worse olfaction. These results were consistent in the two ancestry groups separately. Moreover, our results are in agreement with a recent study [389] which showed strong association between

hyposmia and positive α-synuclein SAA, and that the α-synuclein signature differs between *LRRK2* and *GBA1* mutation carriers. Our results provide additional support for the role of genetic variation at the *LRRK2* and *GBA1* loci in determining smell outcomes in PD, which is significant as smell loss is a potential surrogate for α-synuclein pathology.

To further understand the role of *LRRK2* and *GBA1* variation in α-synuclein pathology, we conducted a GWAS on the SAA binary phenotype (positive or negative). We did not find genome-wide significance support for G2019S and N370S to be associated with the SAA status. However, it is worth mentioning that N370S and G2019S reached nominal significance (P-value < 0.05). G2019S was associated with higher odds of being SAA positive. N370S was associated with higher odds of being SAA negative. These associations were in opposite directions to what I was expecting. Moreover, we explored conditioning the UPSIT genetic association study on the SAA status. We found the association between the two top independent variants at the *LRRK2* locus to be lost, whereas the association the *GBA1 N370S* and the UPSIT score remained nominally associated (P-value < 1e-5), with consistent effects (N370S patients having a worse average olfactory performance). This suggests that *LRRK2* genetic variability might be associated with the subsequent development of α-synuclein pathology, whereas the pathological influence of *GBA1*, which is also associated with UPSIT performance, might be independent than that from the α-synuclein deposition and accumulation pathological implications in PD. This suggests that within the PD-SAA positive population LRRK2 does not affect hyposmia, whereas *GBA1* status continues to be associated with more severe hyposmia. In fact, when we performed a transcriptome wide differentially expression analysis, based on *LRRK2* G2019S mutation carrier stratification, we did find the expression levels of *SNCA* were nominally different between G2019S mutation carriers versus non carriers (P-value = 0.03) (not significant after applying Bonferroni correction). This nominal association was not seen when we stratified based on N370S status.

Here, we have also proven the importance of performing ancestry-based analyses in PD. We found an LD block in the AJ meta-analysis to be nominally associated with H&Y score. This LD block tagged the *PACRG* gene, a gene next to the *PARKIN* gene that is associated with autosomal recessive juvenile PD. This LD block was not present in the H&Y EUR-specific meta-analysis. In addition, we have shown based on a multi-

ancestry meta-analysis that there are ancestry-shared genetic determinants of PD phenotypes, and that by gathering summary statistics from those, we can gain further insights into the genetic determinants of the phenotype. For instance, based on an multi-ancestry fixed-effects IVW meta-analysis in the UPSIT phenotype, we uncover three novel LD blocks at the *SERGED*, *SCN1A*, *OTUD7A* loci that were missed on the ancestry-specific genetic studies due to a lack of power. At each of these loci the lead SNP is more common in the AJ population.

This study has some limitations. We have performed a large meta-analysis of PD phenotypes across data sources. During post meta-analysis processing, we kept all the SNP-level summary statistics that were present in at least 40% of the total data available. This implies that for some SNPs, the association could arise from only one data source out of the 4 data sources part of the large meta-analysis. We highlight three reasons why we decided to do the analysis this way. Some cohorts such as AMP-PD and GP2 gather data from multiple sources in a way that a significant association from any of those comes as a result of a shared effect and direction of an hypothetical SNP on an outcome. Those data sources have a large number of samples available. Second, not all the data has been whole-genome sequenced which leads to a mismatch of variants present between the genetic datasets. This time we did not want to discard genetic associations just because we did not have enough power to impute one haplotype in 1 out of the 4 data sources, for example, which we think is very conservative for a discovery-based analysis. The final reason is that observational studies might have some inherent selection-bias based on study design which could be masking true associations, particularly where we do not have large sample sizes of deeply phenotyped data. Therefore, we decided to perform a meta-analysis in which the subsequent quality control involved keeping only those variants that were present for at least 40% of all data points, as well as applying heterogeneity tests. Another limitation is that, we have only been able to gather deep phenotype data from Europeans and Ashkenazi Jewish individual. This has prevented us from using meta-analysis approaches that efficiently capture the heterogeneity in allelic effects that is correlated with ancestries, therefore being limited to only being able explore ancestry-shared genetic variations and accounting for between groups heterogeneity.

Finally, the severity of PD, as measured by a variety of clinical assessments, is only one aspect of PD, and more generally, of any complex disorder with genetic influence.

How disease progresses, and when a certain phenotype is reached are two other major questions of interest that can be assessed from a genetic point of view. Long-gwas automates these two other types of analyses and we envision it will facilitate discoveries of genetic factors of progression, as well as guarantee reproducibility across sites performing similar analysis on different data repositories. Therefore, we hope long-gwas becomes a useful workflow for people to better understand the genetic implications of progressive and worsening trends in complex genetic disorders, which we believe will give new insights into direct actionable mechanisms to test and develop disease-modifying treatments.

6) PD progression cell type enrichment analysis

a) Introduction

GWAS is a valuable method that has enabled the genetic characterization of many diseases and traits, leading to novel successful genetic-based therapeutics [240]. One premise for conducting GWAS is that they make key contributions to a refined biological understanding of heritable diseases and traits. However, functional annotation of genetic association studies is a major challenge [391]. The number of genetic studies with successful functional annotation are very few [250]. Most genetic associations fall in non-coding regions [252], which have proven challenging to annotate, therefore interpret. Some studies suggest that these non-coding disease risk loci are enriched for cis-regulatory elements (CREs) [252], so it is plausible to think that they might be associated with phenotypes through the control in gene expression.

Genetic variation and regulation does not generalise across tissues nor cell types [256,260]. Farh and colleagues reported that genetic variants in the non-coding genome were enriched for promoters and enhancers. In addition, they found the enrichment to be cell-specific [253]. A major open question when interpreting GWAS is to know the cell type and tissues in which fine-mapped variants and their nominated affected genes are active. In recent years, large consortium studies have generated data to enable an understanding of the heterogeneity around regulatory hallmarks and to determine how those differentially influence disease. The GTEx project was set up to explore how genetic variation influences the transcriptome levels across human tissues. The latest GTEx consortium data analysis, derived a catalogue of genetic regulatory variants that control gene expression and splicing events in *cis* (within 1Mb span of the genetic variant position) and trans (distal regulations within or between chromosomes) across 49 tissues. This effort led to an atlas of tissue specific regulatory effects [386]. These genetic regulatory atlases are a new avenue to characterise at a greater resolution variation of complex diseases and traits [392]. However, studying genetic regulatory events from bulk tissues limits the functional interpretation due to the lack of knowledge of cellular specificity, that is bulk RNA analysis includes multiple cell types and states. Further efforts using GTEx data were able to nominate cell type

clusters within bulk tissues and derive cell-type interaction quantitative trait loci. The cellular context can be further characterised by mapping these cell type interactions QTLs to genetic variants regulating expression and splicing, helping to define how genetic regulation events happen in a cell type specific manner [393]. In addition, single-cell (sc) and single-nuclei (sn) methods for RNA sequencing have helped in understanding cell-type specific phenotypes. They allow to profile gene expression in specific cells [394–396].

Integrating gene expression data from tissue and single cells and overlaying it with genome-wide knowledge of disease and trait risk variants provide valuable insight as to what cell types and tissues relate to specific variants of interest. In the past decade, there have been successful studies performing cell and tissue enrichment analyses [285,397–399], some with a special focus on PD [285,400,401]. Strikingly, Bryois and colleagues found cholinergic and monoaminergic neurons, enteric neurons and oligodendrocytes to be the primary cell types involved in PD aetiology based on cell type specific data from the entire nervous system and the Nalls' and colleagues PD risk GWAS [51,285,402].

While progress is being made understanding the cell specific alterations in idiopathic PD, we know little about the cellular basis of disease progression. which may be distinct from PD risk. In addition, the state of the art methods to perform cell type enrichment analyses are based on partitioned heritability LDSC and MAGMA [115,283]. Despite their robust statistical power to nominate genes in close proximity to the target gene part of a gene set to be tested for enrichment in tissues and cell types, the influence of CREs in the non-coding genome might be partially missed. Other existing approaches that take into account distal regulatory information such eQTLs to nominate genes from SNP-level data might be more accurate.

In this chapter I explore the advanced annotation of progression GWAS signals defined earlier in the thesis with the growing catalogue of tissue and cell specific gene expression datasets. As the amount of snRNA and scRNA-seq data increases over the next 5 years, decoding GWAS will become an increasingly tractable problem. I also propose a novel framework using Transcriptome Wide Mendelian Randomization TWMR that efficiently adds causal interpretations of GWASs based on distal regulatory information eQTL studies.

b) Methods

The code I developed to run a type of cell type enrichment analysis is available at (https://github.com/AMCalejandro/celltype_twmr). The README explains each workflow step with a reference to the specific notebook to run the analysis.

i) GWAS data and quality control

For this analysis, I used the PD progression GWAS carried out in our lab, as well as large PD risk and AAO GWASs [51,105,109,374,375,403]. I also used non PD-related GWASs as control datasets to validate the sensitivity of our pipeline to nominate expected associations between traits and certain cell types. I accessed the meta-analysis GWAS for amyotrophic lateral sclerosis (ALS), schizophrenia, coronary artery disease (CAD), height and body mass index (BMI) [366,404–407]. I first applied control and harmonised all the GWAS progression summary statistics using the MungeSumStats Bioconductor package [408]. I removed any GWAS SNPs that are strand-ambiguous or non-biallelic, and we stored them in a standard format .

ii) Cell type and tissue expression datasets

I accessed several cell type specific expression datasets. A superset of mice brain scRNA-seq data from the cortex, hippocampus, hypothalamus and midbrain of independent studies from the Karolinska Institutet (KI) but that was generated with identical methods, making the ensemble possible [398]. Zeisel mouse hippocampus and cortex scRNA-seq data [396]. SnRNA-seq from the middle temporal gyrus of the human cortex from Allen Institute of Brain Sciences (AIBS) [394]. Blue Lake adult human frontal cortex snRNA-seq data [395]. I only worked with those genes that have 1:1 orthologs with the human species. All these cell type datasets were pre-harmonised and standardised beforehand, having matrices related to the gene's mean expression, specificity (division of each gene expression in a given cell type and the total expression of that gene across all tissues), as well as specificity quantiles and deciles. I accessed harmonised cell type datasets through MAGMA.Celltyping R package [282].

I also accessed GTEx bulk RNA-seq tissue-level data. I followed the same data preparation procedure that Bryois and colleagues highlighted in their analysis [285]. I

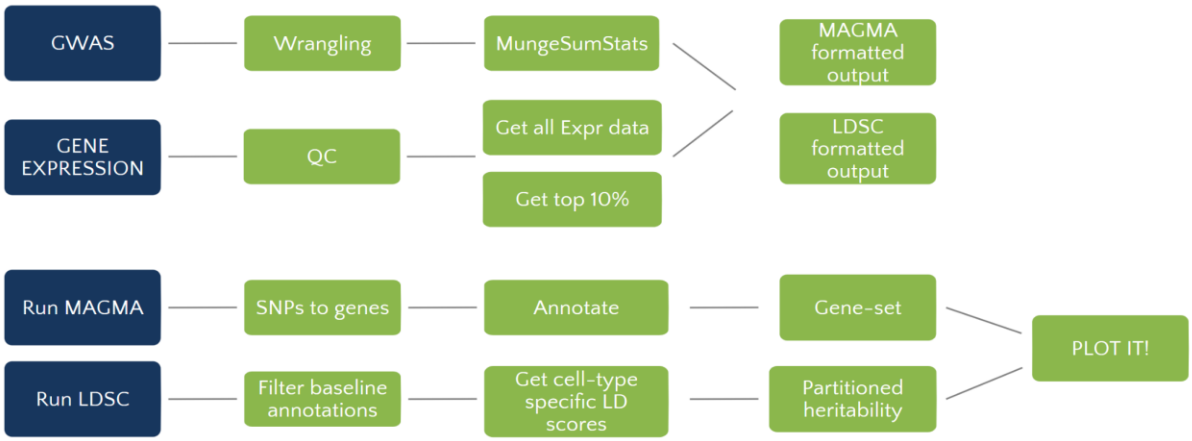
used GTEx pre-computed tissue median expression. I removed tissues sampled in less than 100 individuals. I averaged the expression of tissues by organ (with the exception of brain tissues), resulting in a gene expression profile for 37 tissues. I removed genes without 1:1 human mice orthologs. I scaled gene expression to 1 million UMIs or transcript per million (TPM) for each cell type and tissue. I calculated gene expression specificity based on the division of each gene expression in a given cell type and the total expression of that gene across all tissues. Then, I derived the 10% most specific genes across tissues.

iii) S-LDSC and MAGMA

To perform cell type enrichment of PD progression traits, I used MAGMA and S-LDSC methods as described in **Chapter 2 - Methods**. In brief, MAGMA uses a multiple regression approach to assess the association between the top 10% cell and tissue type specific gene markers and the gene-level P-values converted to Z-scores from GWAS data. With S-LDSC, I tested if the 10% most specific genes of each cell type were enriched in heritability for the PD progression traits, based on the specificity measures available for each cell. I computed LD scores for each cell type and tissue. S-LDSC computes the proportion of SNP heritability associated with our cell type taking into account all other baseline functional annotations in the baseline model. I used the coefficient z-score P-value to assess the association of the cell type with a trait.

I set up a workflow around MAGMA and S-LDSC strategies to perform a cell-type enrichment analysis using GWAS and cell type and tissue datasets highlighted previously (**Figure 32**).

***Figure 32**. Workflow description to perform cell type enrichment analyses based on MAGMA and LDSC across a range of input GWASs.*



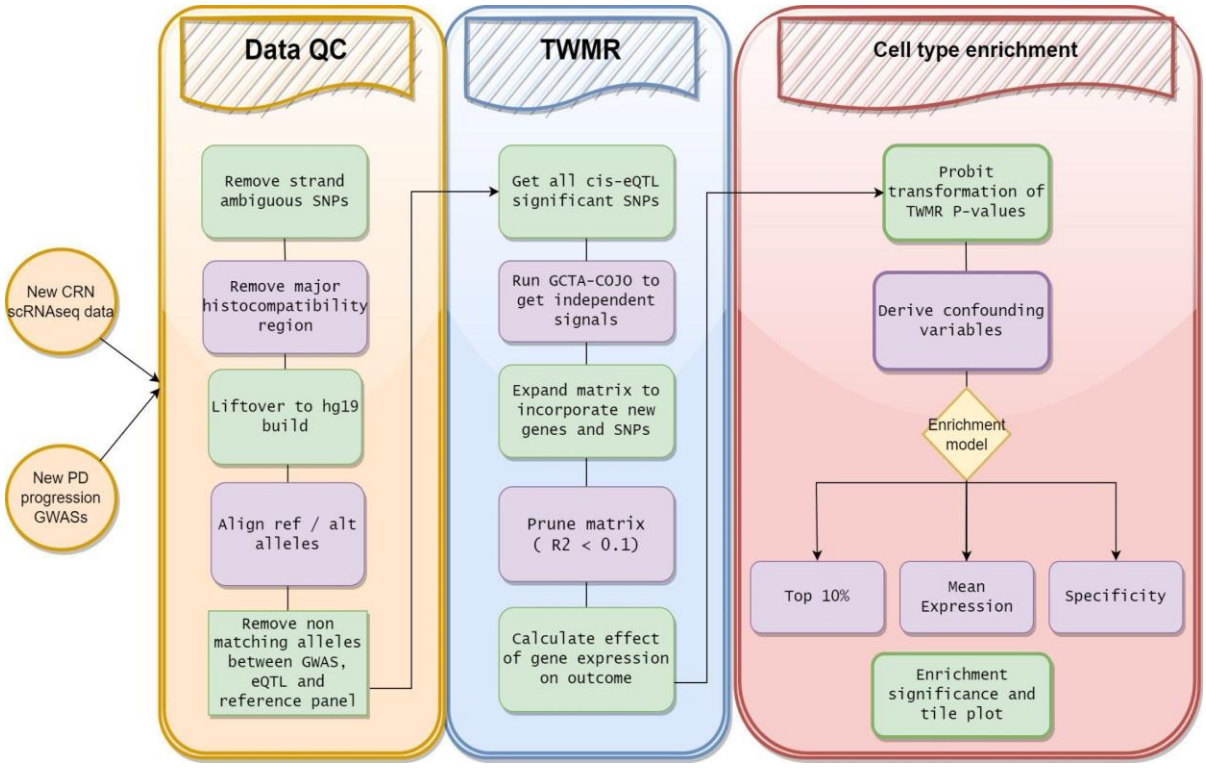
iv) Transcriptome-wide Mendelian Randomization (TWMR)

TWMR is a method with a Mendelian Randomization framework (explained in Methods) that incorporates information from cis-regulatory elements (<1MB from lead SNP) based on eQTLs to measure the causal effect of gene expression on complex traits by using multiple genetic variants as instrumental variables. The two main strengths of this approach is the aggregation of multiple SNPs together, which increases the statistical power as opposed to MR-based SNP approaches which have, by GWAS definition, small effects on phenotypes. In addition, TWMR effectively accounts for horizontal pleiotropy by adding on the multivariable MR framework the mediators as exposures to a given variant that exhibits horizontal pleiotropy. Likewise, bias is mitigated through the joint estimation of the causal effects of all exposures on an outcome [409].

I adopted TWMR inferences and added it to a novel framework to perform cell type enrichment analyses (**Figure 33**). I applied a first step of data QC that led to data subsets of shared SNPs between GWAS, eQTL datasets, and a reference panel. Then, TWMR is performed as described in the referenced methods. At the per-gene level, I define all significant cis-eQTLs to then perform GCTA-COJO to derive all independent significant eQTLs for single gene expression as my quantitative trait. Then, I expand the resulting eQTL matrix to add extra exposures that enable us to efficiently account for horizontal pleiotropy, as well add the extra instrumental variables

a new exposure might have from the cis-eQTL significant SNPs input. Subsequently, I perform data pruning (Rsq < 0.1) to avoid multicollinearity issues and then run TWMR. Based on TWMR estimates, I can apply cell type enrichment analysis based on a multiple regression model. I use a Z-score (probit transformed P-values for the TWMR causal inference) as my outcome, and gene expression estimates from cell and tissue expression datasets as the regressor. I treat the regressors in three different ways, which are the mean expression of genes in a cell type, specificity of a gene expression on a cell type, and a binary indicator capturing the top 10% most differentially expressed genes on a cell type. I account for confounders in the regression model (number of SNPs incorporated in the TWMR framework, number of genes included for the causal inference, and gene size).

***Figure 33**. Workflow description to perform cell type enrichment analyses based on TWMR.*



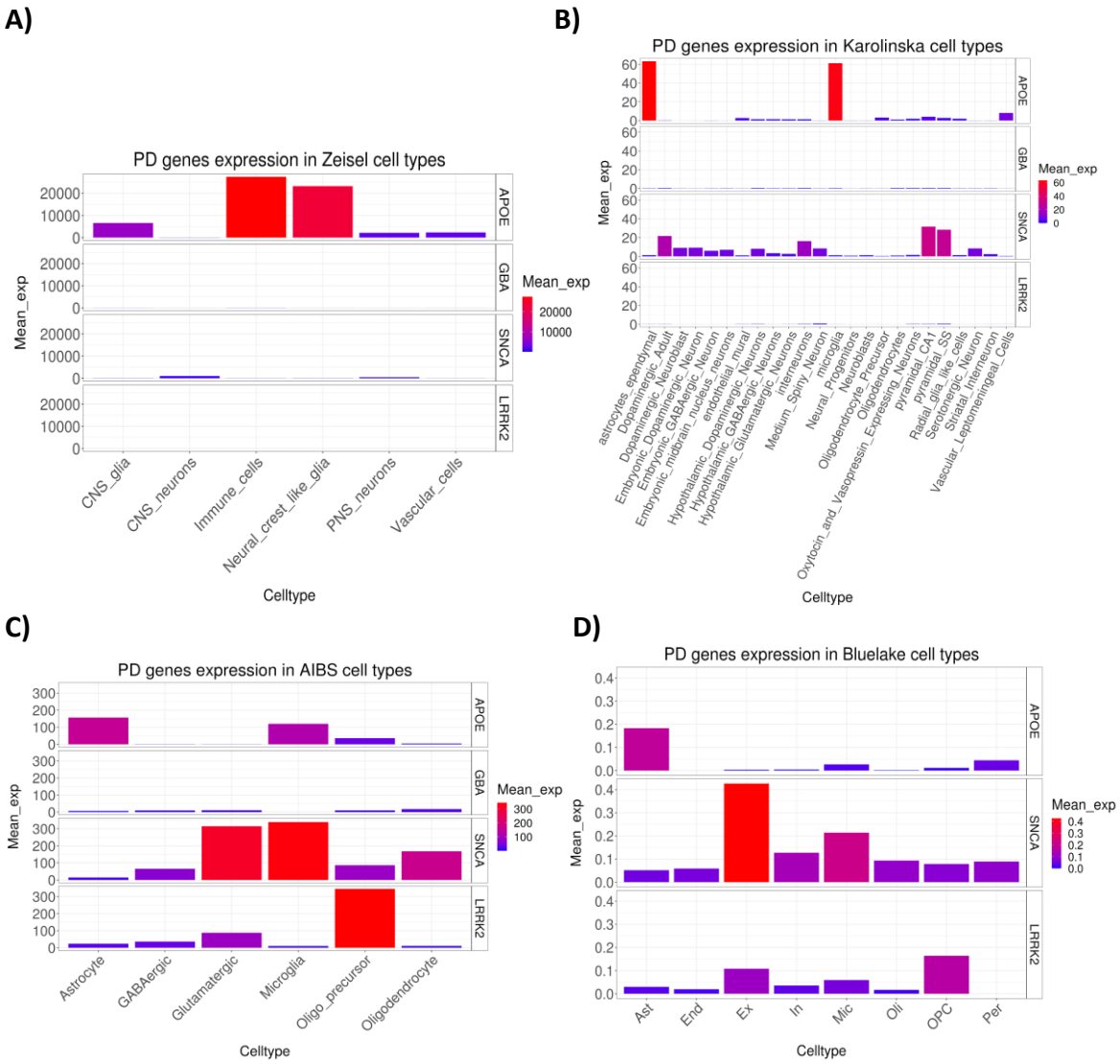
I used a custom multiple regression model to assess cell type specific genes mean expression, specificity, and top 10% differential expression against GWAS Z-scores as model outcome.

c) Results

i) Single nuclei and cell RNA-seq datasets highlight variability in expression between cell types

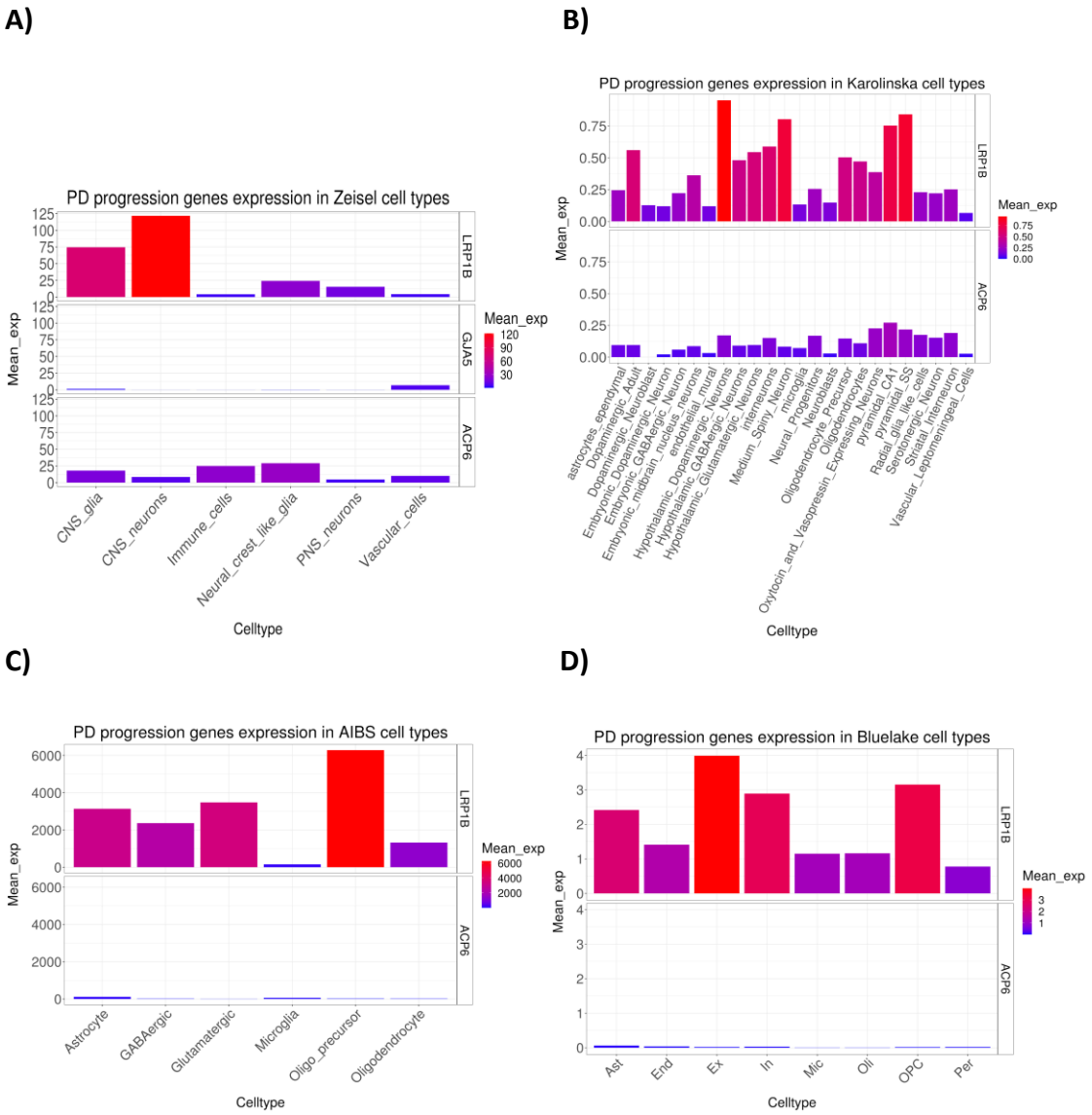
I accessed Zeisel scRNA-seq mouse cortex and hippocampus, single-cell RNA seq superset of the entire nervous system from multiple KI sites, AIBS snRNA-seq, and Blue Lake snRNA-seq datasets as described in **Methods**. I explored the cell type specific data for the mean expression of candidate PD risk and progression genes. I plotted the mean expression across cell types and cell type datasets. I found variability in the average expression of 4 well established PD risk and progression genes (*SNCA*, *GBA1*, *APOE*, *LRRK2*) (**Figure 34**). I found *APOE* gene expression to be predominant in astrocytes and nervous system immune cells (microglia). *SNCA* was expressed across most brain cell types from the KI superset, without a clear predominant expression on a specific cell type. This was not consistent with the mean expression observed in cell types from the AIBS scRNA-set dataset, in which the *SNCA* expression was predominant in microglia and Glutamatergic neurons. *LRRK2* and *GBA1* expression was close to 0 in all cell types and data sources. For the nominated PD progression genes, I found *LRP1B* expression to be predominant in central nervous system neurons based on Zeisel mouse data. *LRP1B* was found to be expressed in oligodendrocyte precursor cells and excitatory neurons based on the snRNA-seq adult human cortex datasets. *ACP6* was expressed across different cell types of the Zeisel mice dataset (**Figure 35**).

***Figure 34**. Average expression of Parkinson's disease* candidate *genes*



I assessed the average expression of PD candidate genes on A) Zeisel mouse cell types B) KI mouse cell types C) AIBS human cortex cell types and D) Blue Lake human cortex cell types.

***Figure 35**. Average expression of Parkinson's disease progression genes.*



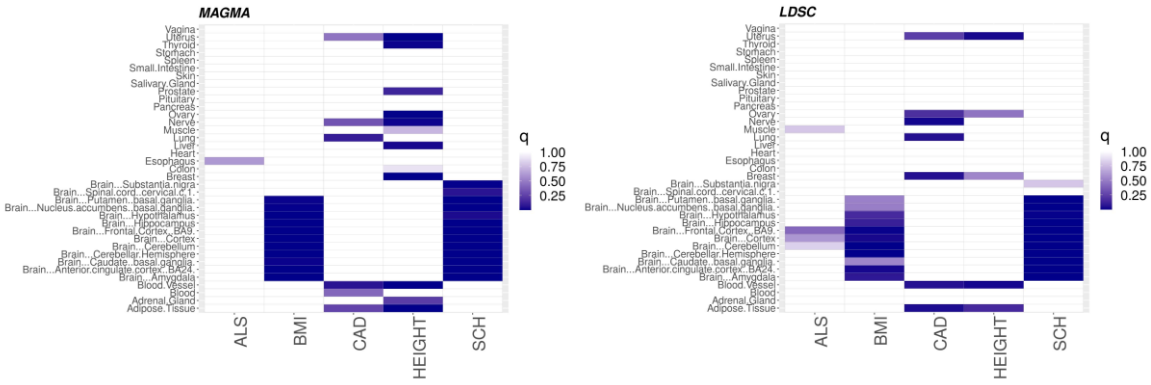
I assessed the average expression of PD candidate genes on A) Zeisel mouse cell types B) KI mouse cell types C) AIBS human cortex cell types and D) Blue Lake human cortex cell types.

ii) Cell type enrichment workflow validation on GTEx tissues

This analysis involved two major steps, which are gene expression and GWAS data pre-processing and then running cell type enrichment analysis (**Figure 32**). As a proof of concept, I validated the workflow using GTEx data and control GWASs to assess whether the workflow was accurate to detect the expected trait-tissue enrichments (**Figure 36**). I found results from MAGMA and S-LDSC to be consistent with one another with some exceptions. Interestingly, I found BMI GWAS signals to be enriched in brain tissues only, consistent with previous findings [410]. CAD GWAS signals were

found to be mostly enriched in blood and blood vessel tissues, but not in heart tissues. I found a high enrichment of schizophrenia traits across most of the brain tissue. I could not find the same enrichment across brain tissues for ALS. ALS was not enriched for any of the tissues in GTEx. This lack of tissue enrichment for the ALS GWAS could be a limitation of the quality of the ALS meta-analysis or power limitation given its smaller sample size. Overall, I concluded the workflow I set up enabled us to detect expected cell type enrichments for the GWAS diseases and traits we tested here.

Figure 36. *Tile plot showing the cell type enrichment results based on A) Magma and B) LDSC.*

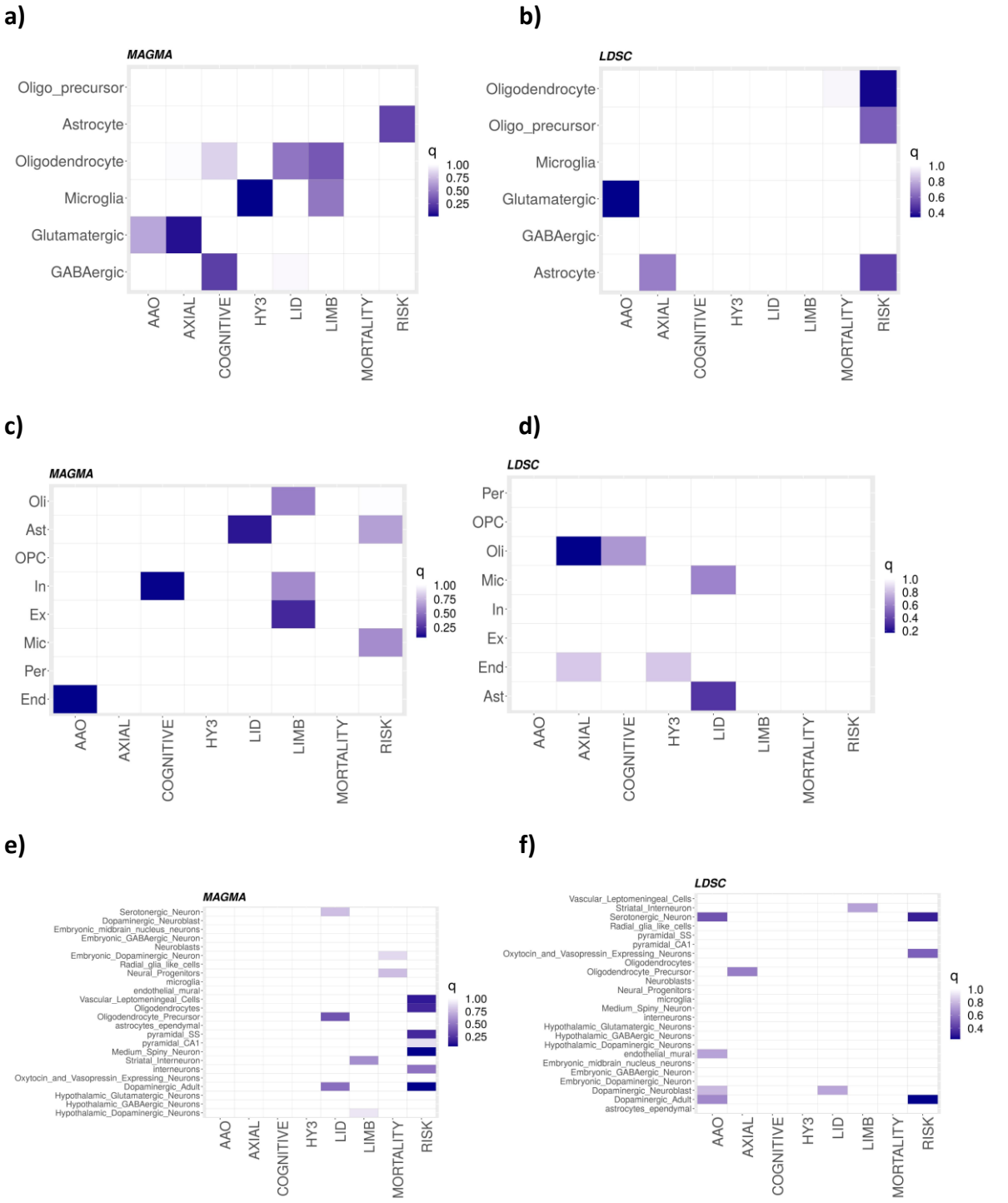


The X-axis labels correspond to the GWASs. The y-axis labels are the GTEx tissues assessed for enrichment. The colour represents the strength of the association based on the Bonferroni corrected P-value (q). The stronger the colour, the stronger the association.

iii) Assessing the cell and tissue enrichment of PD traits and PD risk

Once I validated the sensitivity of the two methods to detect enrichment over expected tissues from the GTEx consortium, I expanded the analysis to assess the enrichment related to PD traits. I explored the enrichment of PD risk, PD AAO and PD progression traits and cell types from AIBS human cortex snRNA-seq (**Figure 37 a, b**), Blue Lake human cortex snRNA-seq (**Figure 37 c, d**), and Karolinska Institutet (KI) (**Figure 37 e,f**).

***Figure 37**. Tile plot showing the cell type enrichment results based on MAGMA (Figure a,c, e) and LDSC (Figure b, d, f).*



The X-axis labels correspond to the GWASs. The y-axis labels are the cell types assessed for enrichment. Figures A and B include the enrichment analysis on the AIBS cell types. Figures C and D include the enrichment analysis on the Blue Lake cell types. Figures E and E include the enrichment analysis on the KI cell types. The colour represents the strength of the association based on the Bonferroni corrected P-value (q). The stronger the colour, the stronger the association.

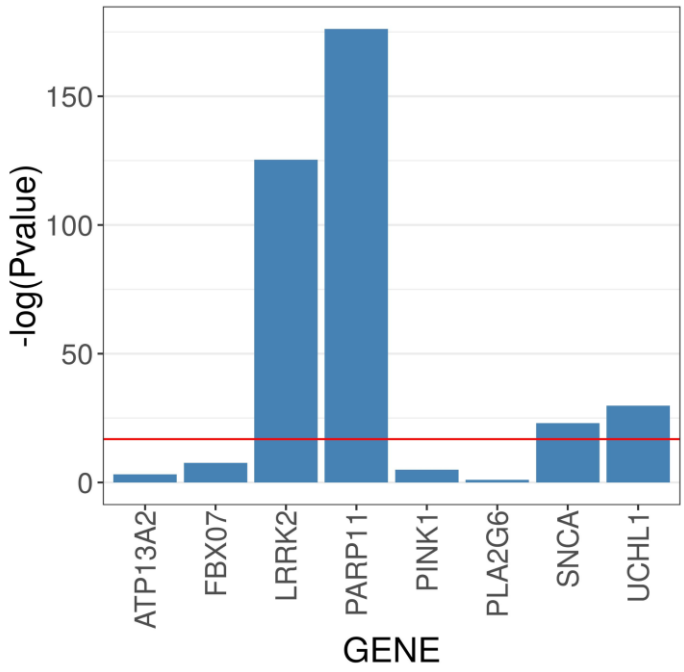
I found a significant association between HY3 stage and microglia (Bonferroni corrected P-value: q = 0.04), based on results from MAGMA analysis and cell types from AIBS snRNA-seq (**Figure 37 a**). I could not replicate this significant association based on the S-LDSC approach (**Figure 37 b**). Interestingly, this association between microglia and H&Y stage was preserved in cell types from different samples from the Blue Lake human cortex snRNA-seq using MAGMA (**Figure 37 c**). I found the same situation in which the association was not replicated with the S-LDSC method (**Figure 37 d).** I also found a nominal significant enrichment between Glutamatergic and GABAergic cell types and the LiD GWAS (**Figure 37 a**). Even though this enrichment did not reach significance after Bonferroni correction, previous studies suggest that disrupted inputs into the basal ganglia from the GABAergic and glutamatergic pathways, may be involved in the occurrence of LiD [411]. I found a nominal association (q = 0.05) between PD risk GWAS and Dopaminergic Adult and Medium Spiny Neurons (MSNs) from the mouse entire nervous system KI superset, using the MAGMA method. This nominal association between Dopaminergic neurons and PD risk was expected, and has been previously reported [285]. In addition, MSNs have been previously reported to suffer a severe lack of Dopamine which lead to compensatory and dysregulatory changes [412]. Interestingly, the association between Dopaminergic Adult neurons prevailed based on the S-LDSC method on the same KI superset (**Figure 37 f**). The association was not significant after Bonferroni correction (q = 0.1). Finally, I also found a nominal association between oligodendrocytes and the PD axial GWAS (**Figure 37 d)** (q = 0.09).

v) TWMR cell type enrichment on GTEx tissues

One major limitation of this analysis of cell type enrichment of PD risk and progression traits is the limited sample size of the PD progression GWASs. Another limitation comes from the two main approaches I used, MAGMA and S-LDSC, which depend on the position of the SNPs or the LD structure of a tagging SNP. Therefore, they impose a limitation based on the distance from genes and length of LD blocks used to generate with gene-level summary statistics and LD-Scores respectively. Neither of these methods can accurately account for influences of distal regulatory elements. As an alternative I adopted TWMR and incorporated it into a framework to perform cell type enrichment analyses.

I first generated TWMR inferences for the PD risk GWAS, which is an estimate of the gene's causal effect on the outcome, and a test-statistics of the causal association. I checked whether TWMR successfully nominated PD candidate genes as PD risk causal genes. I used the list of genes from Blauwendraat and colleagues [4]. Eight genes were available in our TWMR results. Promisingly, I found the test-statistics of the TWMR inferences to be significant in four out of the eight genes we tested (**Figure 38**).

***Figure 38**. Histogram showing the PD candidate genes (X-axis), and the strength of the association (y-axis).*



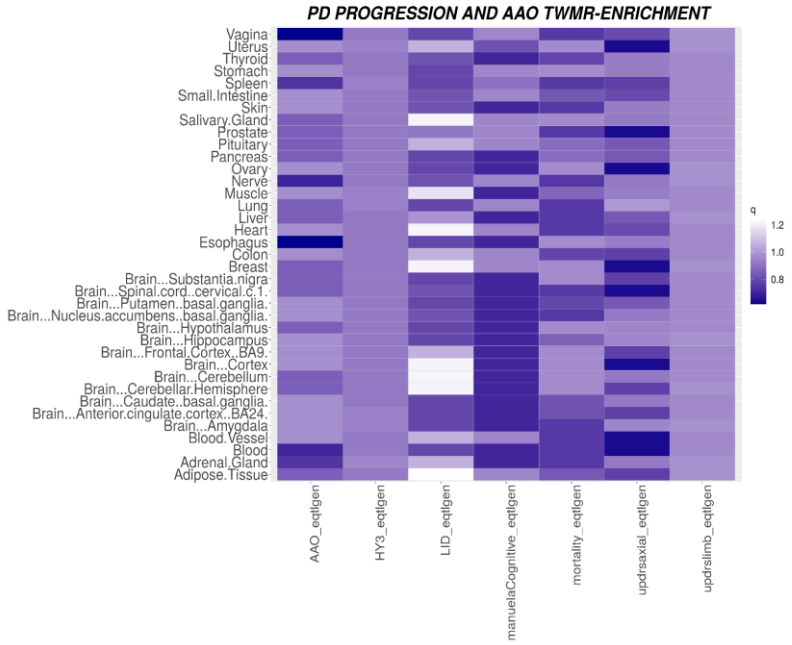
I then performed cell type enrichment analysis against all GTEx tissues based on 3 distinct measures (Specificity, mean expression, and top 10% most differentially expressed genes). I found a significant enrichment after Bonferroni correction ($q < 0.05$) between PD risk and several basal ganglia regions (Putamen, Nucleus accumbens, and caudate), and Hypothalamus, anterior cingulate cortex, and Amygdala (**Figure 39**). Comparing the three measures available to test against enrichment, I found the specificity measure of genes on cell types, being the most accurate to use as the regressor in the enrichment analyses.

***Figure 39**. Tile plot showing the cell type enrichment results based on TWMR against the PD risk GWAS.*

The x-axis labels are the three different measures to record gene expression that were used to regress against PD risk GWAS. (mean_exp =mean expression of a gene; top10 = top 10% differentially expressed genes; specificity = gene expression specificity to a cell type). The y-axis labels correspond to the tissues from GTEx and that were used to perform enrichment on the PD risk GWAS. The colour represents the strength of the association based on the Bonferroni corrected P-value (q). The stronger the colour, the stronger the association.

Subsequently, I performed an enrichment analysis using the specificity measure against all the PD progression traits (**Figure 40**). Unfortunately, I did not find any significant enrichment. I could not find a clear significant enrichment between PD progression GWASs and brain tissues as I found in the previous section. I hypothesised this is due to the limited power based on the sample size the progression GWASs had.

Figure 40. *Tile plot showing the cell type enrichment results based on TWMR against all PD progression GWASs and PD AAO GWAS.*



The x-axis labels highlight the GWAS. The y-axis labels correspond to the tissues from GTEx. The colour represents the strength of the association based on the Bonferroni corrected P-value (q). The stronger the colour, the stronger the association.

d) Conclusion

In this analysis, I expanded the latest cell type enrichment analysis on PD risk GWAS to map specific PD progression traits to cell types. I used MAGMA and S-LDSC, which are the state of the art algorithms to perform cell type enrichment. I found a significant association between HY progression and microglia. I was able to find this association replicated in two different datasets of Frontal Cortex from adult human brains. In addition, I found a nominal association between LiD progression and Glutamatergic and GABAergic neurons, which have been previously reported in literature as possible pathways contributing to LiD [411].

In addition, I developed a novel framework using TWMR to perform cell type enrichment analysis that efficiently incorporates distal regulation to perform causal inferences of genetic variants in the different PD progression genetic studies, through changes in gene expression [409]. I found the approach well powered to detect expected enrichment of PD risk GWAS on brain tissues such as Hypothalamus and

basal ganglia regions. However, I did not find any enrichment against the PD progression GWASs, possibly due to a lack of significant associations from the PD progression GWASs due to their limited sample size as a result of the study design and data availability.

I envision this multivariable MR framework to become a core approach to map the genetic risk of PD progression at the cellular level. As we increase the size of PD progression GWASs, we will also gain new insight into the brain cells whose genetic variants are altered contributing to the progressive decline observed in PD.

7) Conclusions and future directions

Nominating novel genetically defined targets that can ameliorate the progressive decline of PD is a worthwhile endeavour since existing PD therapies are based on symptomatic treatments and we lack disease modifying therapies.

Summary and new insights derived from work

In this research, I have explored different statistical approaches to define how genetics contributes to the progression and severity of PD as measured by clinical instruments that capture different aspects of the disorder. To assess the impact of genetics on disease progression I have used two statistical models, LMM and CPH. LMMs take into account repeated quantitative measures within groups, as well as unexplained random variability in the outcome, to then estimate the effect of genetic variants on disease progression and severity. CPH models allow us to measure the relation between genetic variants and the time to reach a certain outcome. For quantitative outcomes such as MDS-UPDRS, we found the use of LMMs more powerful since this statistical framework can use the high variability between and within groups to retrieve population average estimates based on the time-varying trends observed in the data and to define how much variability is explained on independent variables included in the model. For ordinal measures, that are unable to capture much variability over time due to the limited space of the variable, I found the use of CPHs ideal. We can record the time to reach a certain value of the ordinal variable that records the degree of dyskinesias, and then estimate the effect of variables such as genetic variation on the time to reach that outcome.

To measure how genetics influence disease severity at baseline, I have used GLM. In this scenario, we do not take into account the patients' progression trends to explain heterogeneity in disease presentation by genetics. I consider this type of modelling a useful approach for several reasons. Cross-sectional data is richer than longitudinal data, sometimes not recorded in studies, with drop out of the most severely affected individuals. Findings from GLM GWAS, enable the formulation of progression hypotheses that can be explored based on data-driven approaches. Significant genetic variants from GLM GWAS add another source of information that can be used during longitudinal model design. For instance, we could add prior knowledge of

variants associated with an outcome, as mediator variables in our model to increase the statistical power to assess genetic variants associated with progression.

Based on these statistical models, I have investigated how genetics influences progression and severity of PD motor aspects as measured by limb and axial motor outcomes derived from the MDS-UPDRS part III. I have looked at how genetics contribute to the development of dyskinesias in PD. Finally, I have also undertaken a large-scale meta-analysis in PD exploring the effect of genetic variants on the disease presentation at baseline as measured by a wide range of clinical outcomes.

Performing these analyses involves multiple steps such as data pre-processing and quality control, efficient model fitting, development of results and diagnostic plots. During my PhD, I have been involved in the development of long-gwas, a Nextflow pipeline that automates the multiple steps involved in running longitudinal, cross-sectional, and survival GWASs, as well as all the data preparation and quality control needed beforehand.

Understanding GWAS nominated loci is still a challenge for several reasons. Most of the loci that are usually nominated in association studies fall in non-coding regions [252]. This complicates the identification of the causal gene associated with the trait and the causal mechanism. In addition, the LD structure of the genome confounds efforts to select the causal variant, which adds another difficulty in understanding disease biology. During my thesis I have explored multiple approaches to interpret loci nominated from GWAS. I have used several fine-mapping approaches to nominate consensus SNPs (those nominated to be causal from at least 2 out of 4 fine-mapping tools). I have performed GCTA-COJO to find the number of independent significant SNPs. I have done colocalization analyses to determine the mechanism through which causal genes might be influencing the outcome. In addition, I have accessed functional annotation datasets (ENCODE, Roadmap, FANTOM5, Brain cell types epigenetic markers) to characterise the transcription activity, chromatin state, the presence of transcription factor binding sites, the presence of distal enhancer-promoter chromatin loops. This has helped us to propose genes and mechanisms during the GWAS interpretation phase.

As a result of my research, I have nominated genes that are significantly associated with motor progression, the development of dyskinesias, olfaction, and activities of

daily living. In the large meta-analysis of MDS-UPDRS part III motor outcomes, we found one haplotype block at the *GJA5* locus that was significantly associated with the axial motor progression. Further exploration of the GWAS significant signals in eQTL databases suggests that the GWAS hits may control the expression of *ACP6*, an enzyme that regulates lipid metabolism in mitochondria [299]. Based on a separate meta-analysis assessing the association of genetic variants with motor severity, we identified *MAD1L1* and *SOX9* as candidate genes associated with PD axial motor severity. As a proof of the importance of coupling GWAS findings with external functional annotation datasets, the nomination of *SOX9* is a good example. The SNP I linked to *SOX9* was found at a long non-coding locus. When I integrated the locus-specific summary statistics with PLAC-seq data from Brain cell type specific epigenetic marks, we were able to find a long range promoter-enhancer chromatin loop, suggesting this SNP could be associated with MDS-UPDRS part III axial severity through changes in *SOX9* expression.

I found significant associations with the time-to-develop LiD at the *LRP8*, *LINC02353* and *XYLT1* loci. In addition, based on a candidate gene analysis, exploring genetic variants reported to be associated with LiD risk in my large GWAS meta-analysis, I found that genetic variability in *BDNF* and *ANKK2*, were nominally associated with LiD. I did not replicate any other variant associated with LiD risk. In addition, based on a colocalization analysis, looking at all genes within ±1Mb from all GWAS variants with P-value< 1e-7 revealed a second independent causal association in chromosome 1 between LiD and *DNAJB4* gene expression. Conditional analysis further confirmed that both regions were in LD, hence both *LRP8* and *DNAJB4* were independently associated with the time-to-LiD. I was not able to efficiently resolve the non-coding region associated with LiD, and further research should be focused on understanding variability in this locus and the impact on the time-to-LiD. Functional annotation is limited by the availability of cell specific expression data, although this is rapidly increasing.

I also studied the impact of genetics on the severity of Parkinson's disease in a large-scale analysis, as measured by multiple clinical assessments that capture different aspects of the condition (motor performance, cognition, overall disability, and other non-motor features such as olfaction). I undertook a large-scale multi-ancestry analysis across many clinical outcomes and identified two major genetic determinants

for hyposmia, as measured by the UPSIT scale. My results from the UPSIT meta-analysis revealed that *LRRK2* G2019S and *GBA1* N370S modulate the olfactory performance in PD. *LRRK2* G2019S was found to be associated with better olfaction, whereas *GBA1* N370S was associated with worse olfaction. These results were consistent in the two ancestry groups separately. Finally, based on a multi-ancestry meta-analysis, I was able to increase the power and find novel associations at the *SERGED*, *SCN1A*, *OTUD7A* loci.

It is worth noting that I did not find any genetic variants at the *SNCA* locus associated with PD progression, in any of my genetic association studies, which could support the Braak' progression staging through α-synuclein pathological inclusions. Based on findings from the large-scale disease severity GWAS, an hypothesis could be that other genetic factors such as genetic variants at the *LRRK2* and *GBA1* loci have a primary role in disease aetiology through the modulation of the autophagy-lysosomal pathway. Such pathway disruption could then lead to subsequent α-synuclein pathology and characteristic progression pattern as opposed to point mutations in *SNCA* having an impact in disease aetiology and progression. This hypothesis agrees with findings from the α-synuclein SAA analysis, in which they found differences in the rate of α-synuclein SAA positive in *LRRK2* and *GBA1* mutation carriers with hyposmia [389].

In my study, I found *GBA1* and *LRRK2* mutations to be significantly associated with hyposmia based on a GWAS analysis. Braak staging begins at the olfactory bulb. *GBA1* and *LRRK2* mutations occurring at the olfactory bulb could then cause the α-synuclein accumulation and aggregation and therefore be the trigger to the subsequent spread. It is worth mentioning these results were based on the UPSIT olfactory test cross-sectional data. Further analysis investigating how olfaction progresses according to *LRRK2* and *GBA1* status overlying multi-omic data will shed new light into *LRRK2* and *GBA1* implication in progression. I did not find genetic variability in the *APOE* locus associated with the cognitive disease severity as measured by MoCA and MMSE. Other studies have found *APOE* ε4 allele to be strongly associated with cognitive progression [109,116].

A separate but related question is to understand the relevant cell types for nominated GWAS risk variants. Regulatory elements, including promoters, enhancers, and

silencers, are key components of the non-coding genome. They play a pivotal role in determining how genes are expressed in different tissues and cell types. Importantly, many of these elements exhibit cell type- or state-specific activity, underscoring their role in cellular differentiation and function [413]. A cell's phenotype is influenced by the epigenome, DNA accessibility, and chromatin state. As a result of this unique regulatory control, a cell's transcriptome also has a unique signature. As an example, previous data-driven research has been able to find cell type specific differential gene expression markers across main human cell types [414]. As we previously mentioned, some efforts have focused on determining the cell type enrichment of PD risk GWAS using state-of-the-art tools [285,400]. However, knowing the cell type specificity of PD progression GWASs remains a major open question that can be of additional value in understanding how the underlying neuropathology progresses in brain tissues in relation to clinical progression.

During my PhD, I have undertaken a cell type enrichment analysis on all PD progression GWASs that we have undertaken in the lab. I found a significant enrichment between microglia and Hoehn and Yahr state, a measure of disease severity, which suggests genetic variability related to health decline in PD patients might evolve as a result of an impaired immune system. I was able to find this association in two separate adult human brain datasets (AIBS, BlueLake). I also found other associations approaching Bonferroni corrected significance such as the enrichment of my LiD progression GWAS and impairment in Glutamatergic and GABAergic cells, cells previously suggested to influence the occurrence of the condition [411]. These findings need further validation and investigation, as it can be knowledge that must be taken into during drug design and testing of disease-modifying. In addition, I developed a novel framework around TWMR to infer causal associations from GWAS, based on a multivariable MR framework that incorporates information about distal regulation, in theory more efficiently than previous state of art methods, MAGMA and S-LDSC. This work will need further development for its wider use, including optimization to nominate true positives and reduce the background noise, and the development of an R package. However, I found promising results so far, nominating PD candidate genes using a TWMR approach, including *SNCA* and *LRRK2*, as well as promising results from a TWMR-based PD risk GWAS cell type

enrichment analysis nominating and several basal ganglia regions (Putamen, Nucleus accumbens, and caudate), and Hypothalamus, anterior cingulate cortex.

Limitations and future work

Here I faced some limitations that I would like to highlight for the consideration of future research in PD progression. From a statistical point of view, it has been suggested that prognosis GWAS, those only involving individuals which have a condition, might partially suffer from collider bias, as a result of selection bias, in which non correlated causes of the disease appear correlated when only including cases affected by the condition. This scenario occurs as a result of unaccounted confounding between disease incidence and the outcome for the condition of interest, in which causes of the incidence will spuriously correlate with the condition assessed [415]. For instance, among PD cases selected according to their status (PD with a certain disease duration, being treated or not, carrying *GBA1* and *LRRK2* PD causing alleles or not), some SNPs can show spurious associations with PD prognosis or severity. A plausible explanation in this scenario is that in our PD subset, there may well be other factors that associate with the outcome, in which case, the SNP of interest might be significantly associated with the PD clinical outcome at least partially due to collider bias. Recently, methods such as 'Slope-Hunter' have been proposed for adjustment of collider bias in the so-called prognosis GWASs [415]. Further studies involving PD severity and progression should make an effort to always account for collider bias in their pipelines as well as validate existing PD progression GWAS results to be free of such selection bias.

All my research on the association of genetics with PD prognosis and severity has been limited to common SNPs (MAF > 1%). With prognosis and severity GWAS, we do not have enough power to assess how rarer variants can contribute to the heterogeneity of PD progression with the current sample size [240]. Undertaking a genome-wide burden analysis in PD patients with fast versus slow progression might help to nominate certain genes harbouring rare variants associated with the progression of PD. Similarly, assessing the burden of rare variants in cis-regulatory elements, might help to understand the impact of regulatory elements such as enhancers or repressors that have a role in PD progression. These variations may be more pronounced in specific populations and could be overlooked or entirely missed

in others due to naturally occurring, population-based differences in allele frequencies [416]. Therefore adding more ancestry diverse groups when exploring rare variants in relation to PD progression might increase the power to detect significant associations. Similarly this study is limited to SNPs, therefore other types of mutations such as chromosomal mutations that lead to chromosomes being duplicated or lost are not considered in this analysis. A good example is copy number variants (CNVs), with implication in human diseases and evolution [417]. The identification of causal variants influencing PD risk can be facilitated by employing state-of-the-art high-throughput long-read sequencing technologies. Causal variants are not confined to SNPs but can also encompass more complex genomic variations, such as repeat expansions or structural variants. These variations may be easily overlooked in short-read sequencing and can be technologically challenging to genotype due to repetitive sequences or high GC content. PD studies specifically investigating non-SNP variations are beginning to emerge [418].

Another limitation to highlight is the sample size achieved gathering data from multiple sources in this study. Even if we performed some of the more powered PD progression studies in terms of the sample size we gathered for several PD traits, we are far from the sample sizes achieved by large studies assessing the genetic risk of traits and diseases [51,365,366]. GWAS sample size is another major determinant of power to uncover significant association genome-wide [240]. Initiatives such as GP2 hold the promise of gathering a deeply phenotyped and harmonised dataset of thousands of PD cases with a multitude of clinical assessments and longitudinal data available. Such effort, if it overcomes the many difficulties that data gathering and curation to high standard imposes, might lead to unprecedented discoveries explaining PD severity and progression. And their translation to disease-modifying therapies.

Another limitation is that I have entirely focused on clinical assessments to assess PD progression. However wearable technology enables recording measures for patients more reliably and continuously. Such data is starting to become available for some biobanks such as in the UK Biobank. Using more novel deep learning approaches such as Long Short Term Memory (LSTM) models might provide additional value and increase the power to analyse this type of wealthy data with many time points available longitudinally at the patient level, which me able to better capture treatment responses and fluctuations in the progression patterns.

In the study of PD motor progression, I relied on LMM as I found this model that assumes a linear trend to fit our data best. However, other studies of longitudinal studies of longer duration (8-year follow up) found nonlinear mixed-effect models to fit UPDRS serial measures better than LMM. Several reasons such as the inclusion of more advanced PD participants and long-term effect of patients under treatment could explain the differing results. However, as we approach larger-scale longitudinal analyses including outcome for which we are unfamiliar what progression patterns they might follow, as well as we include data from sources with notable differences of exclusion and inclusion criteria (for example, disease duration restrictions), it is important that careful investigation on the progression trajectories of the multiple clinical outcomes we study is performed so that it can help during the model-decision making, and detect unanticipated nonlinear progressive trends.

Replication across genome-wide analyses is important in the identification of “definite” progression variants. This has proven to be difficult as outlined in the previous studies carried out by Liu and colleagues and Real and colleagues I highlighted in the Methods. Similarly, my results have not been in agreement with previous analyses of PD motor progression or dyskinesias. There are several potential reasons that might explain lack of replication. This could relate to differences in coverage from the genotyping arrays used across different studies. This might lead to a low imputation power in key loci which would lead to, for example, dropping the haplotypes that are associated with the PD phenotype studied. Second reason could be due to variability in these phenotypes, characterised by daily fluctuations, measurement inaccuracies, the possible influence of environmental factors specific from each region, and the challenges of standardising clinical assessments across various research sites and study methodologies, which could lead to the surge of false positives. Another plausible reason could be systematic errors during the analysis process, which might lead to collider or confounding bias on GWASs results when assessing the association between a variant and a PD progression trait, which could lead to inflated test-statistics and therefore false positives and negatives. Even if I anticipate a decrease in the non replication issue with the use of long-gwas, further efforts should be focused on understanding the lack of replication to understand potential novel confounders we can account for as the field moves forward.

To date, in the PD genetics field, most studies have focused on European ancestry populations [419], and more recently, Asian descent [51,83]. Incorporating ancestral diversity into PD genetics research holds crucial significance for enhancing various aspects of PD healthcare. Firstly, the exclusion of non-European populations in research may result in an underestimation of specific genetic risk factors to those populations, which could serve as valuable markers for early disease detection and risk assessment. Furthermore, including diverse populations can be of value to validate PD risk factors from European populations or define ancestry specific novel genetic factors. An illustrative example of such distinctions is evident in the largest GWAS conducted in the Asian population, where no associations between the PD phenotype and *GBA1* or *MAPT* variants were observed, and variability in those loci where among the top hits in the European ancestry GWAS [83]. The incorporation of ancestral diversity into PD research is vital for advancing our comprehension of the disease's biology and pathogenesis. This understanding, in turn, facilitates the customisation of preventive measures and therapeutic interventions. As Nalls and colleagues have anticipated, the inclusion of ancestry-diverse groups will allow us to increase the genetic data granularity, which will improve the interpretation of GWAS signals and improve the applicability and usefulness of PD genetics studies [83]. During my PhD, I have analysed the impact of genetics on disease severity of PD patients of AJ descent. However, this still holds a limitation to highlight for several reasons: Sample sizes available for study are much lower than the numbers available for EUR PD patients. This study was limited to the impact of AJ genetic on severity and overlooked the impact on progression due to low availability of time-series data for the clinical outcomes part of studies.

Exploring nominated loci from GWAS involves dealing with diverse molecular pathways contributing to the phenotype of interest [420]. Within these nominated loci, detecting the causal variant can be challenging, often obscured by other non-causal alleles falling within the same haplotype block due to the underlying LD structure. In this context, the refinement of genotyping approaches and the development and implementation of novel bioinformatics tools are crucial. Additionally, methods that overlay functional annotation resources such as DNA methylation or histone modification of regulatory elements, as well as the formation of chromatin loops, with GWASs provide insights into the putative epigenetic signatures of GWAS-nominated

loci. Further research to decrease the uncertainty from fine-mapping tools, understand the cell type basis of genetic association studies, will boost GWAS discoveries and easy transferability from research to disease-modifying therapies testing and development. In addition, keeping up to date with advances in the deep learning genomics will help GWAS decoding with variant prioritisation, interpretability of the functional implications of non-coding genome nominated variants, as well as accurate cell type agnostic predictions of the impact of PD specific genetic architecture [421–425].

Post-GWAS analyses are primarily directed towards identifying molecular pathways and promising targets for biomarkers and drug development. The process involves the discovery and validation of potential findings in independent cohorts, allowing the nomination of pathways for further assessment in cell lines and animal models or the construction of networks. Additionally, novel datasets for PD genetics research are becoming publicly available resources for the research community. An example is the Foundational Data Initiative for Parkinson's Disease (FOUNDIN-PD) [426], an international, collaborative, and multi-year project. It aims to generate a multi-layered molecular dataset using a large cohort of 95 induced pluripotent stem cell (iPSC) lines at multiple time points during differentiation to dopaminergic (DA) neurons. Frameworks for GWAS decoding should work around these single cell PD specific powerful datasets. Similarly, the use of deeply phenotype cohorts with matched multi-omic data provide us with a unique opportunity to explore nominated genetic variants and the impact on broad gene expression and translation. This enables us to nominate potential biomarkers of genetic markers of progression as well as to decode non-coding genetic association from GWASs.

8) References

1. Parkinson J. An essay on the shaking palsy. 1817. J Neuropsychiatry Clin Neurosci [Internet]. 2002 Mar;14(2). Available from: http://dx.doi.org/10.1176/JNP.14.2.223

2. Pringsheim T, Jette N, Frolkis A, Steeves TDL. The prevalence of Parkinson's disease: A systematic review and meta-analysis. Mov Disord. 2014 Nov 1;29(13):1583–90.

3. Greenland JC, Williams-Gray CH, Barker RA. The clinical heterogeneity of Parkinson's disease and its therapeutic implications. Eur J Neurosci. 2019 Feb;49(3):328–38.

4. Blauwendraat C, Nalls MA, Singleton AB. The genetic architecture of Parkinson's disease. Lancet Neurol. 2020 Feb;19(2):170–8.

5. GBD 2016 Neurology Collaborators. Global, regional, and national burden of neurological disorders, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol. 2019 May;18(5):459–80.

6. Bloem BR, Okun MS, Klein C. Parkinson's disease. Lancet. 2021 Jun 12;397(10291):2284–303.

7. Collier TJ, Kanaan NM, Kordower JH. Ageing as a primary risk factor for Parkinson's disease: evidence from studies of non-human primates. Nat Rev Neurosci. 2011 Jun;12(6):359–66.

8. Hindle JV. Ageing, neurodegeneration and Parkinson's disease. Age Ageing. 2010 Mar;39(2):156–61.

9. Ben-Shlomo Y, Darweesh S, Llibre-Guerra J, Marras C, San Luciano M, Tanner C. The epidemiology of Parkinson's disease. Lancet. 2024 Jan 20;403(10423):283–92.

10. Tansey MG, Wallings RL, Houser MC, Herrick MK, Keating CE, Joers V. Inflammation and immune dysfunction in Parkinson disease. Nat Rev Immunol. 2022 Nov;22(11):657–73.

11. Wooten GF, Currie LJ, Bovbjerg VE, Lee JK, Patrie J. Are men at greater risk for Parkinson's disease than women? J Neurol Neurosurg Psychiatry. 2004 Apr;75(4):637–9.

12. Dotchin C, Msuya O, Kissima J, Massawe J, Mhina A, Moshy A, et al. The prevalence of Parkinson's disease in rural Tanzania. Mov Disord. 2008 Aug 15;23(11):1567–672.

13. Morens DM, Davis JW, Grandinetti A, Ross GW, Popper JS, White LR. Epidemiologic observations on Parkinson's disease: incidence and mortality in a prospective study of middle-aged men. Neurology. 1996 Apr;46(4):1044–50.

14. Jankovic J. Parkinson's disease: clinical features and diagnosis. J Neurol Neurosurg Psychiatry. 2008;79(4):368–76.

15. Chaudhuri KR, Healy DG, Schapira AHV. Non-motor symptoms of Parkinson's disease: diagnosis and management. Lancet Neurol. 2006 Mar;5(3):235–45.

16. Schenck CH, Bundlie SR, Mahowald MW. Delayed emergence of a parkinsonian disorder in 38% of 29 older men initially diagnosed with idiopathic rapid eye movement

sleep behavior disorder. 1996.

17. Olson EJ, Boeve BF, Silber MH. Rapid eye movement sleep behaviour disorder: demographic, clinical and laboratory findings in 93 cases. Brain. 2000;123 (Pt 2)(2):331–9.

18. Aarsland D, Andersen K, Larsen JP, Lolk A, Kragh-Sørensen P. Prevalence and characteristics of dementia in Parkinson disease: an 8-year prospective study. Arch Neurol. 2003 Mar 1;60(3):387–92.

19. Morris HR, Spillantini MG, Sue CM, Williams-Gray CH. The pathogenesis of Parkinson's disease. Lancet. 2024 Jan 20;403(10423):293–304.

20. Postuma RB, Berg D, Stern M, Poewe W, Olanow CW, Oertel W, et al. MDS clinical diagnostic criteria for Parkinson's disease. Mov Disord. 2015 Oct;30(12):1591–601.

21. Kalia LV, Lang AE. Parkinson's disease. Lancet. 2015 Aug 29;386(9996):896–912.

22. Dickson DW, Braak H, Duda JE, Duyckaerts C, Gasser T, Halliday GM, et al. Neuropathological assessment of Parkinson's disease: refining the diagnostic criteria. Lancet Neurol. 2009 Dec;8(12):1150–7.

23. Goedert M, Spillantini MG, Del Tredici K, Braak H. 100 years of Lewy pathology. Nat Rev Neurol. 2013 Jan;9(1):13–24.

24. Tansey MG, Goldberg MS. Neuroinflammation in Parkinson's disease: its role in neuronal death and implications for therapeutic intervention. Neurobiol Dis. 2010 Mar;37(3):510–8.

25. Stefanis L. α-Synuclein in Parkinson's disease. Cold Spring Harb Perspect Med [Internet]. 2012;2(2). Available from: https://pubmed.ncbi.nlm.nih.gov/22355802/

26. Langston JW. The MPTP Story. J Parkinsons Dis. 2017;7(s1):S11–9.

27. Li W, Fu Y, Halliday GM, Sue CM. PARK Genes Link Mitochondrial Dysfunction and Alpha-Synuclein Pathology in Sporadic Parkinson's Disease. Frontiers in Cell and Developmental Biology [Internet]. 2021;9. Available from: https://www.frontiersin.org/articles/10.3389/fcell.2021.612476

28. Wang XL, Feng ST, Wang YT, Yuan YH, Li ZP, Chen NH, et al. Mitophagy, a Form of Selective Autophagy, Plays an Essential Role in Mitochondrial Dynamics of Parkinson's Disease. Cell Mol Neurobiol. 2022 Jul;42(5):1321–39.

29. Hely MA, Morris JG, Reid WG, O'Sullivan DJ, Williamson PM, Rail D, et al. The Sydney Multicentre Study of Parkinson's disease: a randomised, prospective five year study comparing low dose bromocriptine with low dose levodopa-carbidopa. J Neurol Neurosurg Psychiatry. 1994 Aug;57(8):903–10.

30. Hely MA, Morris JGL, Reid WGJ, Trafficante R. Sydney Multicenter Study of Parkinson's disease: non-L-dopa-responsive problems dominate at 15 years. Mov Disord. 2005 Feb;20(2):190–9.

31. Hely MA, Reid WGJ, Adena MA, Halliday GM, Morris JGL. The Sydney multicenter study of Parkinson's disease: the inevitability of dementia at 20 years. Mov Disord. 2008 Apr 30;23(6):837–44.

32. Variable expression of Parkinson's disease [Internet]. Neurology. [cited 2024 Jun 20].

Available from: https://www.neurology.org/doi/10.1212/WNL.40.10.1529

33. Graham JM, Sagar HJ. A data-driven approach to the study of heterogeneity in idiopathic Parkinson's disease: identification of three distinct subtypes. Mov Disord. 1999 Jan;14(1):10–20.

34. Dujardin K, Defebvre L, Duhamel A, Lecouffe P, Rogelet P, Steinling M, et al. Cognitive and SPECT characteristics predict progression of Parkinson's disease in newly diagnosed patients. J Neurol. 2004 Nov;251(11):1383–92.

35. Lewis SJG, Foltynie T, Blackwell AD, Robbins TW, Owen AM, Barker RA. Heterogeneity of Parkinson's disease in the early clinical stages using a data driven approach. J Neurol Neurosurg Psychiatry. 2005 Mar;76(3):343–8.

36. Schrag A, Quinn NP, Ben-Shlomo Y. Heterogeneity of Parkinson's disease. J Neurol Neurosurg Psychiatry. 2006 Feb;77(2):275–6.

37. Reijnders JSAM, Ehrt U, Lousberg R, Aarsland D, Leentjens AFG. The association between motor subtypes and psychopathology in Parkinson's disease. Parkinsonism Relat Disord. 2009 Jun;15(5):379–82.

38. Fereshtehnejad SM, Zeighami Y, Dagher A, Postuma RB. Clinical criteria for subtyping Parkinson's disease: biomarkers and longitudinal progression. Brain. 2017 Jul 1;140(7):1959–76.

39. Lawton M, Ben-Shlomo Y, May MT, Baig F, Barber TR, Klein JC, et al. Developing and validating Parkinson's disease subtypes and their motor and cognitive progression. J Neurol Neurosurg Psychiatry. 2018 Dec;89(12):1279–87.

40. Marras C, Lang A. Parkinson's disease subtypes: lost in translation? J Neurol Neurosurg Psychiatry. 2013;84(4):409–15.

41. Fereshtehnejad SM, Postuma RB. Subtypes of Parkinson's disease: What do they tell us about disease progression? Curr Neurol Neurosci Rep. 2017 Apr;17(4):34.

42. Fearnley JM, Lees AJ. Ageing and Parkinson's disease: substantia nigra regional selectivity. Brain. 1991 Oct;114 (Pt 5):2283–301.

43. Marras C, Beck JC, Bower JH, Roberts E, Ritz B, Ross GW, et al. Prevalence of Parkinson's disease across North America. NPJ Parkinsons Dis. 2018 Jul 10;4:21.

44. Braak H, Rüb U, Jansen Steur ENH, Del Tredici K, de Vos RAI. Cognitive status correlates with neuropathologic stage in Parkinson disease. Neurology. 2005 Apr 26;64(8):1404–10.

45. Braak H, Del Tredici K, Rüb U, de Vos RAI, Jansen Steur ENH, Braak E. Staging of brain pathology related to sporadic Parkinson's disease. Neurobiol Aging. 2003 Mar-Apr;24(2):197–211.

46. Hurtig HI, Trojanowski JQ, Galvin J, Ewbank D, Schmidt ML, Lee VM, et al. Alpha-synuclein cortical Lewy bodies correlate with dementia in Parkinson's disease. Neurology. 2000 May 23;54(10):1916–21.

47. Apaydin H, Ahlskog JE, Parisi JE, Boeve BF, Dickson DW. Parkinson disease neuropathology: later-developing dementia and loss of the levodopa response. Arch Neurol. 2002 Jan;59(1):102–12.

48. Colosimo C, Hughes AJ, Kilford L, Lees AJ. Lewy body cortical involvement may not always predict dementia in Parkinson's disease. J Neurol Neurosurg Psychiatry. 2003 Jul;74(7):852–6.

49. Halliday G, Hely M, Reid W, Morris J. The progression of pathology in longitudinally followed patients with Parkinson's disease. Acta Neuropathol. 2008 Apr;115(4):409–15.

50. Goldman SM, Marek K, Ottman R, Meng C, Comyns K, Chan P, et al. Concordance for Parkinson's disease in twins: A 20-year update. Ann Neurol. 2019 Apr;85(4):600–5.

51. Nalls MA, Blauwendraat C, Vallerga CL, Heilbron K, Bandres-Ciga S, Chang D, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet Neurol. 2019 Dec 1;18(12):1091–102.

52. Polymeropoulos MH, Lavedan C, Leroy E, Ide SE, Dehejia A, Dutra A, et al. Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. Science. 1997 Jun 27;276(5321):2045–7.

53. Valente EM, Abou-Sleiman PM, Caputo V, Muqit MMK, Harvey K, Gispert S, et al. Hereditary early-onset Parkinson's disease caused by mutations in PINK1. Science. 2004 May 21;304(5674):1158–60.

54. Deng H, Wang P, Jankovic J. The genetics of Parkinson disease. Ageing Res Rev. 2018;42(September 2017):72–85.

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

56. Polymeropoulos MH, Higgins JJ, Golbe LI, Johnson WG, Ide SE, Di Iorio G, et al. Mapping of a gene for Parkinson's disease to chromosome 4q21-q23. Science. 1996 Nov 15;274(5290):1197–9.

57. Singleton AB, Farrer M, Johnson J, Singleton A, Hague S, Kachergus J, et al. α-Synuclein Locus Triplication Causes Parkinson's Disease. Science. 2003 Oct 31;302(5646):841.

58. Ibáñez P, Bonnet AM, Débarges B, Lohmann E, Tison F, Pollak P, et al. Causal relation between alpha-synuclein gene duplication and familial Parkinson's disease. Lancet. 2004;364(9440):1169–71.

59. Tu PH, Galvin JE, Baba M, Giasson B, Tomita T, Leight S, et al. Glial cytoplasmic inclusions in white matter oligodendrocytes of multiple system atrophy brains contain insoluble alpha-synuclein. Ann Neurol. 1998 Sep;44(3):415–22.

60. Goedert M. Alpha-synuclein and neurodegenerative diseases. Nat Rev Neurosci. 2001 Jul;2(7):492–501.

61. Poewe W, Seppi K, Tanner CM, Halliday GM, Brundin P, Volkmann J, et al. Parkinson disease. Nat Rev Dis Primers. 2017 Mar 23;3:17013.

62. Di Fonzo A, Rohé CF, Ferreira J, Chien HF, Vacca L, Stocchi F, et al. A frequent LRRK2 gene mutation associated with autosomal dominant Parkinson's disease. Lancet. 2005 Jan 29;365(9457):412–5.

63. Lesage S, Dürr A, Tazir M, Lohmann E, Leutenegger AL, Janin S, et al. LRRK2 G2019S as a cause of Parkinson's disease in North African Arabs. N Engl J Med. 2006 Jan

26;354(4):422–3.

64. Pchelina SN, Yakimovskii AF, Ivanova ON, Emelianov AK, Zakharchuk AH, Schwarzman AL. G2019S LRRK2 mutation in familial and sporadic Parkinson's disease in Russia. Mov Disord. 2006 Dec;21(12):2234–6.

65. Ozelius Laurie J., Senthil Geetha, Saunders-Pullman Rachel, Ohmann Erin, Deligtisch Amanda, Tagliati Michele, et al. LRRK2 G2019S as a Cause of Parkinson's Disease in Ashkenazi Jews. N Engl J Med. 354(4):424–5.

66. Bouhouche A, Tibar H, Ben El Haj R, El Bayad K, Razine R, Tazrout S, et al. LRRK2 G2019S Mutation: Prevalence and Clinical Features in Moroccans with Parkinson's Disease. Parkinsons Dis. 2017 Mar 30;2017:2412486.

67. Kluss JH, Mamais A, Cookson MR. LRRK2 links genetic and sporadic Parkinson's disease. Biochem Soc Trans. 2019 Apr 30;47(2):651–61.

68. Nichols WC, Pankratz N, Hernandez D, Paisán-Ruíz C, Jain S, Halter CA, et al. Genetic screening for a single common LRRK2 mutation in familial Parkinson's disease. Lancet. 2005 Jan 29;365(9457):410–2.

69. Lee AJ, Wang Y, Alcalay RN, Mejia-Santana H, Saunders-Pullman R, Bressman S, et al. Penetrance estimate of LRRK2 p.G2019S Mutation in Individuals of Non-Ashkenazi Jewish Ancestry. Mov Disord. 2017 Oct 1;32(10):1432.

70. Kalia LV, Lang AE, Hazrati LN, Fujioka S, Wszolek ZK, Dickson DW, et al. Clinical correlations with Lewy body pathology in LRRK2-related Parkinson disease. JAMA Neurol. 2015 Jan;72(1):100–5.

71. Alessi DR, Sammler E. LRRK2 kinase in Parkinson's disease. Science. 2018 Apr 6;360(6384):36–7.

72. Nguyen APT, Tsika E, Kelly K, Levine N, Chen X, West AB, et al. Dopaminergic neurodegeneration induced by Parkinson's disease-linked G2019S LRRK2 is dependent on kinase and GTPase activity. Proc Natl Acad Sci U S A. 2020 Jul 21;117(29):17296–307.

73. Zimprich A, Biskup S, Leitner P, Lichtner P, Farrer M, Lincoln S, et al. Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. Neuron. 2004 Nov 18;44(4):601–7.

74. Liu W, Liu X 'nan, Li Y, Zhao J, Liu Z, Hu Z, et al. LRRK2 promotes the activation of NLRC4 inflammasome during Salmonella Typhimurium infection. J Exp Med. 2017 Oct 2;214(10):3051–66.

75. Matsumine H, Saito M, Shimoda-Matsubayashi S, Tanaka H, Ishikawa A, Nakagawa-Hattori Y, et al. Localization of a gene for an autosomal recessive form of juvenile Parkinsonism to chromosome 6q25.2-27. Am J Hum Genet. 1997 Mar;60(3):588–96.

76. Valente EM, Bentivoglio AR, Dixon PH, Ferraris A, Ialongo T, Frontali M, et al. Localization of a novel locus for autosomal recessive early-onset parkinsonism, PARK6, on human chromosome 1p35-p36. Am J Hum Genet. 2001 Apr;68(4):895–900.

77. Pickrell AM, Youle RJ. The roles of PINK1, parkin, and mitochondrial fidelity in Parkinson's disease. Neuron. 2015 Jan 21;85(2):257–73.

78. Langston JW, Ballard P, Tetrud JW, Irwin I. Chronic Parkinsonism in humans due to a

product of meperidine-analog synthesis. Science. 1983 Feb 25;219(4587):979–80.

79. Javitch JA, D'Amato RJ, Strittmatter SM, Snyder SH. Parkinsonism-inducing neurotoxin, N-methyl-4-phenyl-1,2,3,6 -tetrahydropyridine: uptake of the metabolite N-methyl-4-phenylpyridine by dopamine neurons explains selective toxicity. Proc Natl Acad Sci U S A. 1985 Apr;82(7):2173–7.

80. International Parkinson Disease Genomics Consortium, Nalls MA, Plagnol V, Hernandez DG, Sharma M, Sheerin UM, et al. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet. 2011 Feb 19;377(9766):641–9.

81. Simón-Sánchez J, Schulte C, Bras JM, Sharma M, Gibbs JR, Berg D, et al. Genome-wide association study reveals genetic risk underlying Parkinson's disease. Nat Genet. 2009 Dec;41(12):1308–12.

82. Kim JJ, Vitale D, Otani DV, Lian MM, Heilbron K, 23andMe Research Team, et al. Multi-ancestry genome-wide association meta-analysis of Parkinson's disease. Nat Genet. 2024 Jan;56(1):27–36.

83. Foo JN, Chew EGY, Chung SJ, Peng R, Blauwendraat C, Nalls MA, et al. Identification of Risk Loci for Parkinson Disease in Asians and Comparison of Risk Between Asians and Europeans: A Genome-Wide Association Study. JAMA Neurol. 2020 Jun 1;77(6):746–54.

84. Chang XL, Mao XY, Li HH, Zhang JH, Li NN, Burgunder JM, et al. Association of GWAS loci with PD in China. Am J Med Genet B Neuropsychiatr Genet. 2011 Apr;156B(3):334–9.

85. Alvarez Jerez P, Wild Crea PA, Ramos D, Gustavsson EK, Radefeldt M, Makarious MB, et al. African ancestry neurodegeneration risk variant disrupts an intronic branchpoint in GBA1. medRxiv [Internet]. 2024 Feb 24; Available from: https://www.medrxiv.org/content/10.1101/2024.02.20.24302827.abstract

86. Smith L, Mullin S, Schapira AHV. Insights into the structural biology of Gaucher disease. Exp Neurol. 2017 Dec 1;298(Pt B):180–90.

87. Malek N, Weil RS, Bresner C, Lawton MA, Grosset KA, Tan M, et al. Features of GBA-associated Parkinson's disease at presentation in the UK Tracking Parkinson's study. J Neurol Neurosurg Psychiatry. 2018 Jul 1;89(7):702–9.

88. Yu Z, Wang T, Xu J, Wang W, Wang G, Chen C, et al. Mutations in the glucocerebrosidase gene are responsible for Chinese patients with Parkinson's disease. J Hum Genet. 2015 Feb 1;60(2):85–90.

89. Aharon-Peretz J, Rosenbaum H, Gershoni-Baruch R. Mutations in the glucocerebrosidase gene and Parkinson's disease in Ashkenazi Jews. N Engl J Med. 2004 Nov 4;351(19):1972–7.

90. Sidransky E, Nalls MA, Aasly JO, Aharon-Peretz J, Annesi G, Barbosa ER, et al. Multi-center analysis of glucocerebrosidase mutations in Parkinson disease. N Engl J Med. 2009 Oct 22;361(17):1651.

91. Neumann J, Bras J, Deas E, O'sullivan SS, Parkkinen L, Lachmann RH, et al. Glucocerebrosidase mutations in clinical and pathologically proven Parkinson's disease. Brain. 2009 Jul;132(Pt 7):1783–94.

92. Setó-Salvia N, Pagonabarraga J, Houlden H, Pascual-Sedano B, Dols-Icardo O, Tucci A, et al. Glucocerebrosidase mutations confer a greater risk of dementia during Parkinson's disease course. Mov Disord. 2012 Mar 1;27(3):393–9.

93. Balestrino R, Tunesi S, Tesei S, Lopiano L, Zecchinelli AL, Goldwurm S. Penetrance of Glucocerebrosidase (GBA) Mutations in Parkinson's Disease: A Kin Cohort Study. Mov Disord. 2020 Nov;35(11):2111–4.

94. Blauwendraat C, Reed X, Krohn L, Heilbron K, Bandres-Ciga S, Tan M, et al. Genetic modifiers of risk and age at onset in GBA associated Parkinson's disease and Lewy body dementia. Brain. 2020 Jan 1;143(1):234–48.

95. Mazzulli JR, Xu YH, Sun Y, Knight AL, McLean PJ, Caldwell GA, et al. Gaucher disease glucocerebrosidase and α-synuclein form a bidirectional pathogenic loop in synucleinopathies. Cell. 2011 Jul 8;146(1):37–52.

96. Yap TL, Velayati A, Sidransky E, Lee JC. Membrane-bound α-synuclein interacts with glucocerebrosidase and inhibits enzyme activity. Mol Genet Metab. 2013 Jan;108(1):56–64.

97. Campêlo CL das C, Silva RH. Genetic Variants in SNCA and the Risk of Sporadic Parkinson's Disease and Clinical Outcomes: A Review. Parkinsons Dis. 2017 Jul 11;2017:4318416.

98. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009 Jun 9;106(23):9362–7.

99. Saunders-Pullman R, Mirelman A, Alcalay RN, Wang C, Ortega RA, Raymond D, et al. Progression in the LRRK2-Asssociated Parkinson Disease Population. JAMA Neurol. 2018 Mar 1;75(3):312–9.

100. Caminiti SP, Carli G, Avenali M, Blandini F, Perani D. Clinical and Dopamine Transporter Imaging Trajectories in a Cohort of Parkinson's Disease Patients with GBA Mutations. Mov Disord. 2022 Jan;37(1):106–18.

101. Pankratz N, Beecham GW, DeStefano AL, Dawson TM, Doheny KF, Factor SA, et al. Meta-analysis of Parkinson's disease: identification of a novel locus, RIT2. Ann Neurol. 2012 Mar;71(3):370–84.

102. Nalls MA, McLean CY, Rick J, Eberly S, Hutten SJ, Gwinn K, et al. Diagnosis of Parkinson's disease on the basis of clinical and genetic classification: a population-based modelling study. Lancet Neurol. 2015 Oct;14(10):1002–9.

103. Davis AA, Andruska KM, Benitez BA, Racette BA, Perlmutter JS, Cruchaga C. Variants in GBA, SNCA, and MAPT influence Parkinson disease risk, age at onset, and progression. Neurobiol Aging. 2016 Jan;37:209.e1–209.e7.

104. Iwaki H, Blauwendraat C, Leonard HL, Liu G, Maple-Grødem J, Corvol JC, et al. Genetic risk of Parkinson disease and progression: An analysis of 13 longitudinal cohorts. Neurology: Genetics. 2019;5(4):1–14.

105. Blauwendraat C, Heilbron K, Vallerga CL, Bandres-Ciga S, von Coelln R, Pihlstrøm L, et al. Parkinson's disease age at onset genome-wide association study: Defining heritability, genetic loci, and α-synuclein mechanisms. Mov Disord. 2019 Jun;34(6):866–75.

106. Chang D, Nalls MA, Hallgrímsdóttir IB, Hunkapiller J, van der Brug M, Cai F, et al. A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. Nat Genet. 2017 Oct;49(10):1511–6.

107. Grover S, Kumar Sreelatha AA, Pihlstrom L, Domenighetti C, Schulte C, Sugier PE, et al. Genome-wide Association and Meta-analysis of Age at Onset in Parkinson Disease: Evidence From the COURAGE-PD Consortium. Neurology. 2022 Aug 16;99(7):e698–710.

108. Liu G, Peng J, Liao Z, Locascio JJ, Corvol JC, Zhu F, et al. Genome-wide survival study identifies a novel synaptic locus and polygenic score for cognitive progression in Parkinson's disease. Nat Genet. 2021 Jun;53(6):787–93.

109. Real R, Martinez-Carrasco A, Reynolds RH, Lawton MA, Tan MMX, Shoai M, et al. Association between the LRP1B and APOE loci and the development of Parkinson's disease dementia. Brain. 2023 May 2;146(5):1873–87.

110. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science. 1993 Aug 13;261(5123):921–3.

111. Pankratz N, Byder L, Halter C, Rudolph A, Shults CW, Conneally PM, et al. Presence of an APOE4 allele results in significantly earlier onset of Parkinson's disease and a higher risk with dementia. Mov Disord. 2006 Jan;21(1):45–9.

112. Davis AA, Inman CE, Wargel ZM, Dube U, Freeberg BM, Galluppi A, et al. APOE genotype regulates pathology and disease progression in synucleinopathy. Sci Transl Med [Internet]. 2020 Feb 5;12(529). Available from: http://dx.doi.org/10.1126/scitranslmed.aay3069

113. Kim R, Park S, Yoo D, Jun JS, Jeon B. Impact of the apolipoprotein E ε4 allele on early Parkinson's disease progression. Parkinsonism Relat Disord. 2021 Feb;83:66–70.

114. Iwaki H, Blauwendraat C, Leonard HL, Kim JJ, Liu G, Maple-Grødem J, et al. Genomewide association study of Parkinson's disease clinical biomarkers in 12 longitudinal patients' cohorts. Mov Disord. 2019 Dec;34(12):1839–50.

115. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015 Apr;11(4):e1004219.

116. Tan MMX, Lawton MA, Jabbari E, Reynolds RH, Iwaki H, Blauwendraat C, et al. Genome-Wide Association Studies of Cognitive and Motor Progression in Parkinson's Disease. Mov Disord. 2021 Feb;36(2):424–33.

117. Chung SJ, Armasu SM, Biernacka JM, Anderson KJ, Lesnick TG, Rider DN, et al. Genomic determinants of motor and cognitive outcomes in Parkinson's disease. Parkinsonism Relat Disord. 2012 Aug;18(7):881–6.

118. Stoker TB, Camacho M, Winder-Rhodes S, Liu G, Scherzer CR, Foltynie T, et al. A common polymorphism in SNCA is associated with accelerated motor decline in GBA-Parkinson's disease. J Neurol Neurosurg Psychiatry. 2020 Jun;91(6):673–4.

119. Szwedo AA, Pedersen CC, Ushakova A, Forsgren L, Tysnes OB, Counsell CE, et al. Association of SNCA Parkinson's Disease Risk Polymorphisms With Disease Progression in Newly Diagnosed Patients. Front Neurol. 2020;11:620585.

120. Ritz B, Rhodes SL, Bordelon Y, Bronstein J. α-Synuclein genetic variants predict faster motor symptom progression in idiopathic Parkinson disease. PLoS One. 2012 May 15;7(5):e36199.

121. Mata IF, Leverenz JB, Weintraub D, Trojanowski JQ, Hurtig HI, Van Deerlin VM, et al. APOE, MAPT, and SNCA genes and cognitive performance in Parkinson disease. JAMA Neurol. 2014 Nov;71(11):1405–12.

122. Markopoulou K, Biernacka JM, Armasu SM, Anderson KJ, Ahlskog JE, Chase BA, et al. Does α-synuclein have a dual and opposing effect in preclinical vs. clinical Parkinson's disease? Parkinsonism Relat Disord. 2014 Jun;20(6):584–9; discussion 584.

123. Fagan ES, Pihlstrøm L. Genetic risk factors for cognitive decline in Parkinson's disease: a review of the literature. Eur J Neurol. 2017 Apr;24(4):561–e20.

124. Goris A, Williams-Gray CH, Clark GR, Foltynie T, Lewis SJG, Brown J, et al. Tau and alpha-synuclein in susceptibility to, and dementia in, Parkinson's disease. Ann Neurol. 2007 Aug;62(2):145–53.

125. Setó-Salvia N, Clarimón J, Pagonabarraga J, Pascual-Sedano B, Campolongo A, Combarros O, et al. Dementia risk in Parkinson disease: disentangling the role of MAPT haplotypes. Arch Neurol. 2011 Mar;68(3):359–64.

126. Papapetropoulos S, Farrer MJ, Stone JT, Milkovic NM, Ross OA, Calvo L, et al. Phenotypic associations of tau and ApoE in Parkinson's disease. Neurosci Lett. 2007 Mar 6;414(2):141–4.

127. Greely HT. The uneasy ethical and legal underpinnings of large-scale genomic biobanks. Annu Rev Genomics Hum Genet. 2007;8:343–64.

128. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012 May 2;13(6):395–405.

129. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015 Mar;12(3):e1001779.

130. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J Clin Epidemiol. 2016 Feb;70:214–23.

131. Nagai A, Hirata M, Kamatani Y, Muto K, Matsuda K, Kiyohara Y, et al. Overview of the BioBank Japan Project: Study design and profile. J Epidemiol. 2017 Mar;27(3S):S2–8.

132. All of Us Research Program Investigators, Denny JC, Rutter JL, Goldstein DB, Philippakis A, Smoller JW, et al. The "All of Us" Research Program. N Engl J Med. 2019 Aug 15;381(7):668–76.

133. Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner KM, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature. 2023 Jan;613(7944):508–18.

134. Kj R. Precision and validity in epidemiologic studies. Modern Epidemiology [Internet]. 1998 [cited 2024 Jan 19]; Available from: https://cir.nii.ac.jp/crid/1573950400304970496

135. Euser AM, Zoccali C, Jager KJ, Dekker FW. Cohort studies: prospective versus retrospective. Nephron Clin Pract. 2009 Aug 18;113(3):c214–7.

136. Song JW, Chung KC. Observational studies: cohort and case-control studies. Plast Reconstr Surg. 2010 Dec;126(6):2234–42.

137. Grimes DA, Schulz KF. Cohort studies: marching towards outcomes. Lancet. 2002 Jan 26;359(9303):341–5.

138. Williams-Gray CH, Mason SL, Evans JR, Foltynie T, Brayne C, Robbins TW, et al. The CamPaIGN study of Parkinson's disease: 10-year outlook in an incident population-based cohort. J Neurol Neurosurg Psychiatry. 2013 Nov;84(11):1258–64.

139. Heinzel S, Lerche S, Maetzler W, Berg D. Global, Yet Incomplete Overview of Cohort Studies in Parkinson's disease. J Parkinsons Dis. 2017;7(3):423–32.

140. Wang K, Zhang H, Kugathasan S, Annese V, Bradfield JP, Russell RK, et al. Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. Am J Hum Genet. 2009 Mar;84(3):399–405.

141. Moschen AR, Tilg H, Raine T. IL-12, IL-23 and IL-17 in IBD: immunobiology and therapeutic targeting. Nat Rev Gastroenterol Hepatol. 2019 Mar;16(3):185–96.

142. Bhide A, Shah PS, Acharya G. A simplified guide to randomized controlled trials. Acta Obstet Gynecol Scand. 2018 Apr;97(4):380–7.

143. Hariton E, Locascio JJ. Randomised controlled trials - the gold standard for effectiveness research: Study design: randomised controlled trials. BJOG. 2018 Dec;125(13):1716.

144. Cotzias GC, Papavasiliou PS, Gellene R. Modification of Parkinsonism — Chronic Treatment with L-Dopa. N Engl J Med. 1969 Feb 13;280(7):337–45.

145. Schwab RS, England AC, Poskanzer DC, Young RR. Amantadine in the Treatment of Parkinson's Disease. JAMA. 1969 May 19;208(7):1168–70.

146. Birkmayer W, Riederer P, Youdim MB, Linauer W. The potentiation of the anti akinetic effect after L-dopa treatment by an inhibitor of MAO-B, Deprenil. J Neural Transm. 1975;36(3-4):303–26.

147. Tolosa E, Martí MJ, Valldeoriola F, Molinuevo JL. History of levodopa and dopamine agonists in Parkinson's disease treatment. Neurology. 1998 Jun;50(6 Suppl 6):S2–10; discussion S44–8.

148. McFarthing K, Buff S, Rafaloff G, Dominey T, Wyse RK, Stott SRW. Parkinson's Disease Drug Therapies in the Clinical Trial Pipeline: 2020. J Parkinsons Dis. 2020;10(3):757–74.

149. Home [Internet]. [cited 2024 Jan 20]. Available from: https://amp-pd.org/

150. Malek N, Swallow DMA, Grosset KA, Lawton MA, Marrinan SL, Lehn AC, et al. Tracking Parkinson's: Study Design and Baseline Patient Data. J Parkinsons Dis. 2015 Nov 21;5(4):947.

151. Szewczyk-Krolikowski K, Tomlinson P, Nithi K, Wade-Martins R, Talbot K, Ben-Shlomo Y, et al. The influence of age and gender on motor and non-motor features of early Parkinson's disease: initial findings from the Oxford Parkinson Disease Center

(OPDC) discovery cohort. Parkinsonism Relat Disord. 2014 Jan;20(1):99–105.

152. Corvol JC, Artaud F, Cormier-Dequaire F, Rascol O, Durif F, Derkinderen P, et al. Longitudinal analysis of impulse control disorders in Parkinson disease. Neurology. 2018 Jul 17;91(3):e189–201.

153. Carroll CB, Webb D, Stevens KN, Vickery J, Eyre V, Ball S, et al. Simvastatin as a neuroprotective treatment for Parkinson's disease (PD STAT): protocol for a double-blind, randomised, placebo-controlled futility study. BMJ Open. 2019 Oct 7;9(10):e029740.

154. Stevens KN, Creanor S, Jeffery A, Whone A, Zajicek J, Foggo A, et al. Evaluation of Simvastatin as a Disease-Modifying Treatment for Patients With Parkinson Disease: A Randomized Clinical Trial. JAMA Neurol. 2022 Dec 1;79(12):1232–41.

155. Global Parkinson's Genetics Program. GP2: The Global Parkinson's Genetics Program. Mov Disord. 2021 Apr;36(4):842–51.

156. Parkinson Progression Marker Initiative. The Parkinson Progression Marker Initiative (PPMI). Prog Neurobiol. 2011 Dec;95(4):629–35.

157. Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov Disord. 2008 Nov 15;23(15):2129–70.

158. Eggers C, Pedrosa DJ, Kahraman D, Maier F, Lewis CJ, Fink GR, et al. Parkinson subtypes progress differently in clinical course and imaging pattern. PLoS One. 2012 Oct 8;7(10):e46813.

159. Louis ED, Tang MX, Cote L, Alfaro B, Mejia H, Marder K. Progression of parkinsonian signs in Parkinson disease. Arch Neurol. 1999 Mar;56(3):334–7.

160. Oh JY, Kim YS, Choi BH, Sohn EH, Lee AY. Relationship between clinical phenotypes and cognitive impairment in Parkinson's disease (PD). Arch Gerontol Geriatr. 2009 Nov;49(3):351–4.

161. Rajput AH, Voll A, Rajput ML, Robinson CA, Rajput A. Course in Parkinson disease subtypes: A 39-year clinicopathologic study. Neurology. 2009 Jul 21;73(3):206–12.

162. OPDC home [Internet]. [cited 2022 May 10]. Available from: https://www.dpag.ox.ac.uk/opdc

163. Rosenthal LS, Drake D, Alcalay RN, Babcock D, Bowman FD, Chen-Plotkin A, et al. The NINDS Parkinson's disease biomarkers program. Mov Disord. 2016 Jun;31(6):915–23.

164. Hoehn MM, Yahr MD. Parkinsonism: onset, progression and mortality. Neurology. 1967;17(5):427–42.

165. Rankin J. Cerebral vascular accidents in patients over the age of 60. II. Prognosis. Scott Med J. 1957 May;2(5):200–15.

166. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc. 2005 Apr;53(4):695–9.

167. Elwyn G, Frosch D, Thomson R, Joseph-Williams N, Lloyd A, Kinnersley P, et al. Shared decision making: a model for clinical practice. J Gen Intern Med. 2012 Oct;27(10):1361–7.

168. Marinus J, Visser M, Verwey NA, Verhey FRJ, Middelkoop HAM, Stiggelbout AM, et al. Assessment of cognition in Parkinson's disease. Neurology. 2003 Nov 11;61(9):1222–8.

169. Stiasny-Kolster K, Mayer G, Schäfer S, Möller JC, Heinzel-Gutenbrunner M, Oertel WH. The REM sleep behavior disorder screening questionnaire--a new diagnostic instrument. Mov Disord. 2007 Dec;22(16):2386–93.

170. Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. Sleep. 1991 Dec;14(6):540–5.

171. Yesavage JA, Brink TL, Rose TL, Lum O, Huang V, Adey M, et al. Development and validation of a geriatric depression screening scale: a preliminary report. J Psychiatr Res. 1982;17(1):37–49.

172. Schwab. Projection technique for evaluating surgery in Parkinson's disease. Third symposium on Parkinson's disease [Internet]. Available from: https://cir.nii.ac.jp/crid/1572543024245760768

173. Doty RL, Deems DA, Stellar S. Olfactory dysfunction in parkinsonism: a general deficit unrelated to neurologic signs, disease stage, or disease duration. Neurology. 1988 Aug;38(8):1237–44.

174. Peto V, Jenkinson C, Fitzpatrick R. PDQ-39: a review of the development, validation and application of a Parkinson's disease quality of life questionnaire and its associated measures. J Neurol. 1998 May;245 Suppl 1:S10–4.

175. Chaudhuri KR, Rizos A, Trenkwalder C, Rascol O, Pal S, Martino D, et al. King's Parkinson's disease pain scale, the first scale for pain in PD: An international validation. Mov Disord. 2015 Oct;30(12):1623–31.

176. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res. 1975;12(3):189–98.

177. Fahn, Elton. UPDRS program members. Unified Parkinsons disease rating scale. Recent acquis.

178. Semple SE, Dick F, Cherrie JW, Geoparkinson Study Group. Exposure assessment for a population-based case-control study combining a job-exposure matrix with interview data. Scand J Work Environ Health. 2004 Jun;30(3):241–8.

179. Holden SK, Finseth T, Sillau SH, Berman BD. Progression of MDS-UPDRS Scores Over Five Years in De Novo Parkinson Disease from the Parkinson's Progression Markers Initiative Cohort. Movement Disorders Clinical Practice. 2018;5(1):47–53.

180. Regnault A, Boroojerdi B, Meunier J, Bani M, Morel T, Cano S. Does the MDS-UPDRS provide the precision to assess progression in early Parkinson's disease? Learnings from the Parkinson's progression marker initiative cohort. J Neurol. 2019 Aug 1;266(8):1927–36.

181. Evers LJW, Krijthe JH, Meinders MJ, Bloem BR, Heskes TM. Measuring Parkinson's

disease over time: The real-world within-subject reliability of the MDS-UPDRS. Mov Disord. 2019 Oct 1;34(10):1480–7.

182. Schrag A, Dodel R, Spottke A, Bornschein B, Siebert U, Quinn NP. Rate of clinical progression in Parkinson's disease. A prospective study. Mov Disord. 2007;22(7):938–45.

183. Maetzler W, Liepelt I, Berg D. Progression of Parkinson's disease in the clinical phase: potential markers. Lancet Neurol. 2009 Dec;8(12):1158–71.

184. Faust-Socher A, Duff-Canning S, Grabovsky A, Armstrong MJ, Rothberg B, Eslinger PJ, et al. Responsiveness to Change of the Montreal Cognitive Assessment, Mini-Mental State Examination, and SCOPA-Cog in Non-Demented Patients with Parkinson's Disease. Dement Geriatr Cogn Disord. 2019 Jul 17;47(4-6):187–97.

185. Lessig S, Nie D, Xu R, Corey-Bloom J. Changes on brief cognitive instruments over time in Parkinson's disease. Mov Disord. 2012 Aug;27(9):1125–8.

186. Hu MTM, Szewczyk-Królikowski K, Tomlinson P, Nithi K, Rolinski M, Murray C, et al. Predictors of cognitive impairment in an early stage Parkinson's disease cohort. Mov Disord. 2014 Mar;29(3):351–9.

187. Chen L, Yu C, Zhang N, Liu J, Liu W. Cognitive impairment in patients with Parkinson's disease: A 30-month follow-up study. Clin Neurol Neurosurg. 2016 Dec;151:65–9.

188. Kim HM, Nazor C, Zabetian CP, Quinn JF, Chung KA, Hiller AL, et al. Prediction of cognitive progression in Parkinson's disease using three cognitive screening measures. Clin Park Relat Disord. 2019 Oct 20;1:91–7.

189. Sikorska K, Rivadeneira F, Groenen PJF, Hofman A, Uitterlinden AG, Eilers PHC, et al. Fast linear mixed model computations for genome-wide association studies with longitudinal data. Stat Med. 2013 Jan 15;32(1):165–80.

190. Ning C, Wang D, Zhou L, Wei J, Liu Y, Kang H, et al. Efficient multivariate analysis algorithms for longitudinal genome-wide association studies. Bioinformatics. 2019 Dec 1;35(23):4879–85.

191. Ko S, German CA, Jensen A, Shen J, Wang A, Mehrotra DV, et al. GWAS of longitudinal trajectories at biobank scale. Am J Hum Genet. 2022 Mar 3;109(3):433–45.

192. Wang Z, Wang N, Wang Z, Jiang L, Wang Y, Li J, et al. HiGwas: how to compute longitudinal GWAS data in population designs. Bioinformatics. 2020 Aug 15;36(14):4222–4.

193. Yuan M, Xu XS, Yang Y, Zhou Y, Li Y, Xu J, et al. SCEBE: an efficient and scalable algorithm for genome-wide association studies on longitudinal outcomes with mixed-effects modeling. Brief Bioinform [Internet]. 2021 May 20;22(3). Available from: http://dx.doi.org/10.1093/bib/bbaa130

194. Sikorska K, Lesaffre E, Groenen PJF, Rivadeneira F, Eilers PHC. Genome-wide Analysis of Large-scale Longitudinal Outcomes using Penalization -GALLOP algorithm. Sci Rep. 2018 May 1;8(1):6815.

195. Gorski M, Rasheed H, Teumer A, Thomas LF, Graham SE, Sveinbjornsson G, et al. Genetic loci and prioritization of genes for kidney function decline derived from a meta-

analysis of 62 longitudinal genome-wide association studies. Kidney Int. 2022 Sep;102(3):624–39.

196. Ning C, Kang H, Zhou L, Wang D, Wang H, Wang A, et al. Performance Gains in Genome-Wide Association Studies for Longitudinal Traits via Modeling Time-varied effects. Sci Rep. 2017 Apr 4;7(1):590.

197. Couto Alves A, De Silva NMG, Karhunen V, Sovio U, Das S, Taal HR, et al. GWAS on longitudinal growth traits reveals different genetic factors influencing infant, child, and adult BMI. Sci Adv. 2019 Sep;5(9):eaaw3095.

198. Adkins DE, Clark SL, Copeland WE, Kennedy M, Conway K, Angold A, et al. Genome-Wide Meta-Analysis of Longitudinal Alcohol Consumption Across Youth and Early Adulthood. Twin Res Hum Genet. 2015 Aug;18(4):335–47.

199. Tang W, Kowgier M, Loth DW, Soler Artigas M, Joubert BR, Hodge E, et al. Large-scale genome-wide association studies and meta-analyses of longitudinal change in adult lung function. PLoS One. 2014 Jul 1;9(7):e100776.

200. Allen RJ, Oldham JM, Jenkins DA, Leavy OC, Guillen-Guio B, Melbourne CA, et al. Longitudinal lung function and gas transfer in individuals with idiopathic pulmonary fibrosis: a genome-wide association study. Lancet Respir Med. 2023 Jan;11(1):65–73.

201. Smith EN, Chen W, Kähönen M, Kettunen J, Lehtimäki T, Peltonen L, et al. Longitudinal genome-wide association of cardiovascular disease risk factors in the Bogalusa heart study. PLoS Genet. 2010 Sep 9;6(9):e1001094.

202. He D, Liu H, Wei W, Zhao Y, Cai Q, Shi S, et al. A longitudinal genome-wide association study of bone mineral density mean and variability in the UK Biobank. Osteoporos Int. 2023 Nov;34(11):1907–16.

203. Yang Z, Hu L, Zhen J, Gu Y, Liu Y, Huang S, et al. Genetic basis of pregnancy-associated decreased platelet counts and gestational thrombocytopenia. Blood. 2024 Apr 11;143(15):1528–38.

204. Kockum I, Huang J, Stridh P. Overview of Genotyping Technologies and Methods. Curr Protoc. 2023 Apr;3(4):e727.

205. Heller MJ. DNA MICROARRAY TECHNOLOGY: Devices, Systems, and Applications. Annu Rev Biomed Eng. 2002;4:129–53.

206. Bandres-Ciga S, Faghri F, Majounie E, Koretsky MJ, Kim J, Levine KS, et al. NeuroBooster Array: A Genome-Wide Genotyping Platform to Study Neurological Disorders Across Diverse Populations. medRxiv [Internet]. 2023 Nov 14; Available from: http://dx.doi.org/10.1101/2023.11.06.23298176

207. Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. Am J Hum Genet. 2001 Jul;69(1):1–14.

208. Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008 Jun;9(6):477–85.

209. Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet. 2001 Oct;29(2):217–22.

210. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, et al. Linkage

disequilibrium in the human genome. Nature. 2001 May 10;411(6834):199–204.

211. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007 Oct 18;449(7164):851–61.

212. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. Nat Genet. 2001 Oct;29(2):229–32.

213. Wall JD, Pritchard JK. Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet. 2003 Aug;4(8):587–97.

214. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet. 2004 Jan;74(1):106–20.

215. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015 Mar;47(3):291–5.

216. Neale BM. Introduction to linkage disequilibrium, the HapMap, and imputation. Cold Spring Harb Protoc. 2010 Mar;2010(3):db.top74.

217. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010 Jul;11(7):499–511.

218. Tewhey R, Bansal V, Torkamani A, Topol EJ, Schork NJ. The importance of phase information for human genomics. Nat Rev Genet. 2011 Mar;12(3):215–23.

219. Belmont JW, Boudreau A, Leal SM, Hardenbol P, Pasternak S, Wheeler DA, et al. A haplotype map of the human genome. Nature. 2005 Oct 27;437(7063):1299–320.

220. Project Consortium G, Author C, Committee S, Group P, College of Medicine B, Institute of MIT B, et al. A map of human genome variation from population-scale sequencing The 1000 Genomes Project Consortium*. 2011; Available from: http://www.ncbi.nlm.nih.gov/snp

221. UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015 Oct 1;526(7571):82–90.

222. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016 Oct;48(10):1279–83.

223. TOPMed imputation server [Internet]. [cited 2024 Mar 25]. Available from: https://imputation.biodatacatalyst.nhlbi.nih.gov/

224. Michigan imputation server [Internet]. [cited 2024 Mar 25]. Available from: https://imputationserver.sph.umich.edu/index.html

225. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3 . 2011 Nov;1(6):457–70.

226. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016 Oct;48(10):1284–7.

227. Browning BL, Browning SR. Genotype Imputation with Millions of Reference Samples. Am J Hum Genet. 2016 Jan 7;98(1):116–26.

228. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet. 2018 Sep 6;103(3):338–48.

229. Loh PR, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016 Nov;48(11):1443–8.

230. Loh PR, Palamara PF, Price AL. Fast and accurate long-range phasing in a UK Biobank cohort. Nat Genet. 2016 Jul;48(7):811–6.

231. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods. 2013 Jan;10(1):5–6.

232. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007 Nov;81(5):1084–97.

233. McCarthy Tools [Internet]. [cited 2024 Sep 23]. Available from: https://www.chg.ox.ac.uk/~wrayner/tools/

234. McCarthy Tools [Internet]. [cited 2024 Jan 16]. Available from: https://www.well.ox.ac.uk/~wrayner/tools

235. Minimac4 - genome analysis wiki [Internet]. [cited 2024 Mar 25]. Available from: https://genome.sph.umich.edu/wiki/Minimac4

236. Loh PR. Eagle v2.4.1 User Manual [Internet]. 2018 [cited 2024 Mar 25]. Available from: https://alkesgroup.broadinstitute.org/Eagle/

237. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011 Jan 7;88(1):76–82.

238. International HapMap Consortium. The International HapMap Project. Nature. 2003 Dec 18;426(6968):789–96.

239. Input filtering [Internet]. [cited 2024 Jan 16]. Available from: https://www.cog-genomics.org/plink/1.9/filter

240. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017 Jul 6;101(1):5–22.

241. Ehret GB. Genome-wide association studies: contribution of genomics to understanding blood pressure and essential hypertension. Curr Hypertens Rep. 2010 Feb;12(1):17–25.

242. Munafò MR, Flint J. Meta-analysis of genetic association studies. Trends Genet. 2004 Sep;20(9):439–44.

243. Lin D, Zeng D. Meta-Analysis of Genome-Wide Association Studies: No Efficiency Gain in Using Individual Participant Data. Genet Epidemiol. 2010 Jan;34(1):60–6.

244. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010 Jul 8;26(17):2190.

245. Tseng GC, Ghosh D, Feingold E. Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res. 2012 May;40(9):3785–99.

246. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010 Jul;42(7):565–9.

247. Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet. 2013 Mar;9(3):e1003348.

248. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020 Sep;15(9):2759–72.

249. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. Bioinformatics. 2015 May 1;31(9):1466–8.

250. Gallagher MD, Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. Am J Hum Genet. 2018 May 3;102(5):717–30.

251. Crick F. Central dogma of molecular biology. Nature. 1970 Aug 8;227(5258):561–3.

252. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012 Sep 7;337(6099):1190–5.

253. Farh KKH, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015 Feb 19;518(7539):337–43.

254. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012 Sep 6;489(7414):75–82.

255. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57–74.

256. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015 Feb 19;518(7539):317–30.

257. Nasser J, Bergman DT, Fulco CP, Guckelberger P, Doughty BR, Patwardhan TA, et al. Genome-wide enhancer maps link risk variants to disease genes. Nature. 2021 May;593(7858):238–43.

258. Boix CA, James BT, Park YP, Meuleman W, Kellis M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. Nature. 2021 Feb;590(7845):300–7.

259. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013 Sep 26;501(7468):506–11.

260. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, et al. Genetic effects on gene expression across human tissues. Nature. 2017 Oct 11;550(7675):204–13.

261. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010 Apr 1;6(4):e1000888.

262. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet. 2018 Aug;19(8):491–504.

263. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics. 2016 May 15;32(10):1493–501.

264. Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012 Mar 18;44(4):369–75, S1–3.

265. Galarneau G, Palmer CD, Sankaran VG, Orkin SH, Hirschhorn JN, Lettre G. Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. Nat Genet. 2010 Dec;42(12):1049–51.

266. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014 May;10(5):e1004383.

267. Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine mapping. J R Stat Soc Series B Stat Methodol. 2020 Dec;82(5):1273–300.

268. Wallace C. A more accurate method for colocalisation analysis allowing for multiple causal variants. PLoS Genet. 2021 Sep;17(9):e1009440.

269. Võsa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021 Sep;53(9):1300–10.

270. de Klein N, Tsai EA, Vochteloo M, Baird D, Huang Y, Chen CY, et al. Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases. Nat Genet. 2023 Mar;55(3):377–88.

271. Wallace C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. PLoS Genet. 2020 Apr;16(4):e1008720.

272. [No title] [Internet]. [cited 2024 Jan 16]. Available from: https://chr1swallace.shinyapps.io/coloc-priors/

273. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017 Nov 28;8(1):1826.

274. Wakefield J. Bayes factors for genome-wide association studies: comparison with P-values. Genet Epidemiol. 2009 Jan;33(1):79–86.

275. Zou Y, Carbonetto P, Wang G, Stephens M. Fine-mapping from summary data with the “Sum of Single Effects” model. PLoS Genet. 2022 Jul;18(7):e1010299.

276. Weissbrod O, Hormozdiari F, Benner C, Cui R, Ulirsch J, Gazal S, et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. Nat Genet.

2020 Dec;52(12):1355–63.

277. Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, Price AL, et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. 2014 Oct;10(10):e1004722.

278. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018 Oct;562(7726):203–9.

279. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol. 2010 Oct;28(10):1045–8.

280. Noguchi S, Arakawa T, Fukuda S, Furuno M, Hasegawa A, Hori F, et al. FANTOM5 CAGE profiles of human and mouse samples. Sci Data. 2017 Aug 29;4:170112.

281. Nott A, Holtman IR, Coufal NG, Schlachetzki JCM, Yu M, Hu R, et al. Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. Science. 2019 Nov 29;366(6469):1134–9.

282. MAGMA_Celltyping: Find causal cell-types underlying complex trait genetics [Internet]. Github; [cited 2024 Jun 4]. Available from: https://github.com/neurogenomics/MAGMA_Celltyping

283. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015 Nov;47(11):1228–35.

284. Bulik-Sullivan B. Partitioned Heritability [Internet]. Github; [cited 2024 Jun 5]. Available from: https://github.com/bulik/ldsc/wiki/Partitioned-Heritability

285. Bryois J, Skene NG, Hansen TF, Kogelman LJA, Watson HJ, Liu Z, et al. Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson's disease. Nat Genet. 2020 May;52(5):482–93.

286. Sanderson E, Glymour MM, Holmes MV, Kang H, Morrison J, Munafò MR, et al. Mendelian randomization. Nat Rev Methods Primers [Internet]. 2022 Feb 10;2. Available from: http://dx.doi.org/10.1038/s43586-021-00092-5

287. Henderson EJ, Lord SR, Brodie MA, Gaunt DM, Lawrence AD, Close JCT, et al. Rivastigmine for gait stability in patients with Parkinson's disease (ReSPonD): a randomised, double-blind, placebo-controlled, phase 2 trial. Lancet Neurol. 2016 Mar;15(3):249–58.

288. Zeng Z, Wang L, Shi W, Xu L, Lin Z, Xu X, et al. Effects of Unilateral Stimulation in Parkinson's Disease: A Randomized Double-Blind Crossover Trial. Front Neurol [Internet]. 2022;12. Available from: https://www.frontiersin.org/articles/10.3389/fneur.2021.812455

289. Goetz CG, Luo S, Wang L, Tilley BC, LaPelle NR, Stebbins GT. Handling missing values in the MDS-UPDRS. Mov Disord. 2015 Oct;30(12):1632–8.

290. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.

291. Fahn S, Oakes D, Shoulson I, Kieburtz K, Rudolph A, Lang A, et al. Levodopa and

the progression of Parkinson's disease. N Engl J Med. 2004 Dec 9;351(24):2498–508.

292. Luke SG. Evaluating significance in linear mixed-effects models in R. Behav Res Methods. 2017 Aug;49(4):1494–502.

293. Schilder BM, Humphrey J, Raj T. echolocatoR: an automated end-to-end statistical and functional genomic fine-mapping pipeline. Bioinformatics [Internet]. 2021 Sep 16; Available from: http://dx.doi.org/10.1093/bioinformatics/btab658

294. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014 Mar 27;507(7493):455–61.

295. PsychENCODE Consortium, Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, et al. The PsychENCODE project. Nat Neurosci. 2015 Dec;18(12):1707–12.

296. de Klein N, Tsai EA, Vochteloo M, Baird D, Huang Y, Chen CY, et al. Brain expression quantitative trait locus and network analysis reveals downstream effects and putative drivers for brain-related diseases [Internet]. bioRxiv. 2021 [cited 2022 Oct 7]. p. 2021.03.01.433439. Available from: https://www.biorxiv.org/content/10.1101/2021.03.01.433439v1

297. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010 Sep 15;26(18):2336–7.

298. Malek N, Kanavou S, Lawton MA, Pitz V, Grosset KA, Bajaj N, et al. L-dopa responsiveness in early Parkinson's disease is associated with the rate of motor progression. Parkinsonism Relat Disord. 2019 Aug;65:55–61.

299. Hiroyama M, Takenawa T. Isolation of a cDNA Encoding Human Lysophosphatidic Acid Phosphatase That Is Involved in the Regulation of Mitochondrial Lipid Biosynthesis*. J Biol Chem. 1999 Oct 8;274(41):29172–80.

300. ACP6 protein expression summary - The Human Protein Atlas [Internet]. [cited 2024 Jan 16]. Available from: https://www.proteinatlas.org/ENSG00000162836-ACP6

301. Schapira AHV. Mitochondria in the aetiology and pathogenesis of Parkinson's disease. Lancet Neurol. 2008 Jan;7(1):97–109.

302. Chen RH, Brady DM, Smith D, Murray AW, Hardwick KG. The spindle checkpoint of budding yeast depends on a tight complex between the Mad1 and Mad2 proteins. Mol Biol Cell. 1999 Aug;10(8):2607–18.

303. Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. Genome-wide association study identifies five new schizophrenia loci. Nat Genet. 2011 Sep 18;43(10):969–76.

304. Ruderfer DM, Fanous AH, Ripke S, McQuillin A, Amdur RL, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. Mol Psychiatry. 2014 Sep;19(9):1017–24.

305. NCI, CBIIT, DCEG, Machiela. LDlink [Internet]. [cited 2022 Jul 12]. Available from: https://analysistools.cancer.gov/LDlink/?var=rs11764590&pop=CEU%2BTSI%2BFIN%2

BGBR%2BIBS&genome_build=grch37&r2_d=r2&window=500000&collapseTranscript=true&tab=ldproxy

306. GTEx Portal [Internet]. [cited 2022 Jul 12]. Available from: https://www.gtexportal.org/home/gene/MAD1L1

307. Trost S, Diekhof EK, Mohr H, Vieker H, Krämer B, Wolf C, et al. Investigating the Impact of a Genome-Wide Supported Bipolar Risk Variant of MAD1L1 on the Human Reward System. Neuropsychopharmacology. 2016 Oct;41(11):2679–87.

308. Rowe JB, Hughes L, Ghosh BCP, Eckstein D, Williams-Gray CH, Fallon S, et al. Parkinson's disease and dopaminergic therapy--differential effects on movement, reward and cognition. Brain. 2008 Aug;131(Pt 8):2094–105.

309. Panigrahi A, O'Malley BW. Mechanisms of enhancer action: the known and the unknown. Genome Biol. 2021 Apr 15;22(1):108.

310. Rose AB. Introns as Gene Regulators: A Brick on the Accelerator. Front Genet. 2018;9:672.

311. Stevanovic M, Drakulic D, Lazic A, Ninkovic DS, Schwirtlich M, Mojsin M. SOX Transcription Factors as Important Regulators of Neuronal and Glial Differentiation During Nervous System Development and Adult Neurogenesis. Front Mol Neurosci. 2021 Mar 31;14:654031.

312. Scott CE, Wynn SL, Sesay A, Cruz C, Cheung M, Gomez Gaviro MV, et al. SOX9 induces and maintains neural stem cells. Nat Neurosci. 2010 Oct;13(10):1181–9.

313. Stolt CC, Lommes P, Friedrich RP, Wegner M. Transcription factors Sox8 and Sox10 perform non-equivalent roles during oligodendrocyte development despite functional redundancy. Development. 2004 May;131(10):2349–58.

314. Stolt CC, Wegner M. SoxE function in vertebrate nervous system development. Int J Biochem Cell Biol. 2010 Mar;42(3):437–40.

315. Jankovic J, McDermott M, Carter J, Gauthier S, Goetz C, Golbe L, et al. Variable expression of Parkinson's disease: a base-line analysis of the DATATOP cohort. The Parkinson Study Group. Neurology. 1990;40(10):1529–34.

316. Schiess MC, Zheng H, Soukup VM, Bonnen JG, Nauta HJ. Parkinson's disease subtypes: clinical classification and ventricular cerebrospinal fluid analysis. Parkinsonism Relat Disord. 2000 Apr 1;6(2):69–76.

317. Sauerbier A, Jenner P, Todorova A, Chaudhuri KR. Non motor subtypes and Parkinson's disease. Parkinsonism Relat Disord. 2016 Jan;22 Suppl 1:S41–6.

318. Erro R, Picillo M, Vitale C, Palladino R, Amboni M, Moccia M, et al. Clinical clusters and dopaminergic dysfunction in de-novo Parkinson disease. Parkinsonism Relat Disord. 2016 Jul;28:137–40.

319. van Rooden SM, Colas F, Martínez-Martín P, Visser M, Verbaan D, Marinus J, et al. Clinical subtypes of Parkinson's disease. Mov Disord. 2011 Jan;26(1):51–8.

320. Müller MLTM, Bohnen NI. Cholinergic dysfunction in Parkinson's disease. Curr Neurol Neurosci Rep. 2013 Sep;13(9):377.

321. Lee SH, Lee MJ, Lyoo CH, Cho H, Lee MS. Impaired finger dexterity and nigrostriatal

dopamine loss in Parkinson's disease. J Neural Transm. 2018 Sep;125(9):1333–9.

322. Gibb WR, Lees AJ. The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease. J Neurol Neurosurg Psychiatry. 1988 Jun;51(6):745–52.

323. Verschuur CVM, Suwijn SR, Boel JA, Post B, Bloem BR, van Hilten JJ, et al. Randomized Delayed-Start Trial of Levodopa in Parkinson's Disease. N Engl J Med. 2019 Jan 24;380(4):315–24.

324. Jankovic J, Tan EK. Parkinson's disease: etiopathogenesis and treatment. J Neurol Neurosurg Psychiatry. 2020 Aug;91(8):795–808.

325. Espay AJ, Morgante F, Merola A, Fasano A, Marsili L, Fox SH, et al. Levodopa-induced dyskinesia in Parkinson disease: Current and evolving concepts. Ann Neurol. 2018 Dec;84(6):797–811.

326. Manson A, Stirpe P, Schrag A. Levodopa-induced-dyskinesias clinical features, incidence, risk factors, management and impact on quality of life. J Parkinsons Dis. 2012;2(3):189–98.

327. Tran TN, Vo TNN, Frei K, Truong DD. Levodopa-induced dyskinesia: clinical features, incidence, and risk factors. J Neural Transm. 2018 Aug;125(8):1109–17.

328. Cilia R, Akpalu A, Sarfo FS, Cham M, Amboni M, Cereda E, et al. The modern pre-levodopa era of Parkinson's disease: insights into motor complications from sub-Saharan Africa. Brain. 2014 Oct;137(Pt 10):2731–42.

329. Khan NL, Graham E, Critchley P, Schrag AE, Wood NW, Lees AJ, et al. Parkin disease: a phenotypic study of a large case series. Brain. 2003 Jun;126(Pt 6):1279–92.

330. van Duijn CM, Dekker MC, Bonifati V, Galjaard RJ, Houwing-Duistermaat JJ, Snijders PJ, et al. Park7, a novel locus for autosomal recessive early-onset parkinsonism, on chromosome 1p36. Am J Hum Genet. 2001 Sep;69(3):629–34.

331. Lin MK, Farrer MJ. Genetics and genomics of Parkinson's disease. Genome Med. 2014 Jun 30;6(6):48.

332. Lohmann E, Thobois S, Lesage S, Broussolle E, du Montcel ST, Ribeiro MJ, et al. A multidisciplinary study of patients with early-onset PD with and without parkin mutations. Neurology. 2009 Jan 13;72(2):110–6.

333. Oliveri RL, Annesi G, Zappia M, Civitelli D, Montesanti R, Branca D, et al. Dopamine D2 receptor gene polymorphism and the risk of levodopa-induced dyskinesias in PD. Neurology. 1999 Oct 22;53(7):1425–30.

334. Darmopil S, Martín AB, De Diego IR, Ares S, Moratalla R. Genetic inactivation of dopamine D1 but not D2 receptors inhibits L-DOPA-induced dyskinesia and histone activation. Biol Psychiatry. 2009 Sep 15;66(6):603–13.

335. Falla M, Di Fonzo A, Hicks AA, Pramstaller PP, Fabbrini G. Genetic variants in levodopa-induced dyskinesia (LID): A systematic review and meta-analysis. Parkinsonism Relat Disord. 2021 Mar;84:52–60.

336. de Lau LML, Verbaan D, Marinus J, Heutink P, van Hilten JJ. Catechol-O-methyltransferase Val158Met and the risk of dyskinesias in Parkinson's disease. Mov Disord. 2012 Jan;27(1):132–5.

337. Yin Y, Liu Y, Xu M, Zhang X, Li C. Association of COMT rs4680 and MAO-B rs1799836 polymorphisms with levodopa-induced dyskinesia in Parkinson's disease—a meta-analysis. Neurol Sci. 2021 Oct 1;42(10):4085–94.

338. Solís O, García-Montes JR, Garcia-Sanz P, Herranz AS, Asensio MJ, Kang G, et al. Human COMT over-expression confers a heightened susceptibility to dyskinesia in mice. Neurobiol Dis. 2017 Jun;102:133–9.

339. Kusters CDJ, Paul KC, Guella I, Bronstein JM, Sinsheimer JS, Farrer MJ, et al. Dopamine receptors and BDNF-haplotypes predict dyskinesia in Parkinson's disease. Parkinsonism Relat Disord. 2018 Feb;47:39–44.

340. Foltynie T, Cheeran B, Williams-Gray CH, Edwards MJ, Schneider SA, Weinberger D, et al. BDNF val66met influences time to onset of levodopa induced dyskinesia in Parkinson's disease. J Neurol Neurosurg Psychiatry. 2009 Feb;80(2):141–4.

341. Cheshire P, Bertram K, Ling H, O'Sullivan SS, Halliday G, McLean C, et al. Influence of single nucleotide polymorphisms in COMT, MAO-A and BDNF genes on dyskinesias and levodopa use in Parkinson's disease. Neurodegener Dis. 2014;13(1):24–8.

342. Bialecka M, Kurzawski M, Klodowska-Duda G, Opala G, Tan EK, Drozdzik M. The association of functional catechol-O-methyltransferase haplotypes with risk of Parkinson's disease, levodopa treatment response, and complications. Pharmacogenet Genomics. 2008 Sep;18(9):815–21.

343. König E, Nicoletti A, Pattaro C, Annesi G, Melotti R, Gialluisi A, et al. Exome-wide association study of levodopa-induced dyskinesia in Parkinson's disease. Sci Rep. 2021 Oct 1;11(1):19582.

344. Publications: Frances Mary Ashcroft. OPDC home [Internet]. [cited 2022 Sep 5]. Available from: https://www.dpag.ox.ac.uk/opdc

345. The R Project for Statistical Computing [Internet]. [cited 2024 Sep 24]. Available from: https://www.r-project.org/

346. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996 Sep 13;273(5281):1516–7.

347. Iwaki H, Leonard HL, Makarious MB, Bookman M, Landin B, Vismer D, et al. Accelerating Medicines Partnership: Parkinson's Disease. Genetic Resource. Mov Disord. 2021 Aug;36(8):1795–804.

348. Home [Internet]. [cited 2022 Sep 27]. Available from: http://lidpd.eurac.edu/

349. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience [Internet]. 2019 Jul 1;8(7). Available from: https://doi.org/10.1093/gigascience/giz082

350. Ku S, Glass GA. Age of Parkinson's disease onset as a predictor for the development of dyskinesia. Mov Disord. 2010 Jul 15;25(9):1177–82.

351. Sharma JC, Bachmann CG, Linazasoro G. Classifying risk factors for dyskinesia in Parkinson's disease. Parkinsonism Relat Disord. 2010 Sep;16(8):490–7.

352. Warren Olanow C, Kieburtz K, Rascol O, Poewe W, Schapira AH, Emre M, et al. Factors predictive of the development of Levodopa-induced dyskinesia and wearing-off in Parkinson's disease. Mov Disord. 2013 Jul;28(8):1064–71.

353. Passarella D, Ciampi S, Di Liberto V, Zuccarini M, Ronci M, Medoro A, et al. Low-Density Lipoprotein Receptor-Related Protein 8 at the Crossroad between Cancer and Neurodegeneration. Int J Mol Sci [Internet]. 2022 Aug 10;23(16). Available from: http://dx.doi.org/10.3390/ijms23168921

354. Hiesberger T, Trommsdorff M, Howell BW, Goffinet A, Mumby MC, Cooper JA, et al. Direct binding of Reelin to VLDL receptor and ApoE receptor 2 induces tyrosine phosphorylation of disabled-1 and modulates tau phosphorylation. Neuron. 1999 Oct;24(2):481–9.

355. Deutschlander AB, Konno T, Soto-Beasley AI, Walton RL, van Gerpen JA, Uitti RJ, et al. Association of MAPT subhaplotypes with clinical and demographic features in Parkinson's disease. Ann Clin Transl Neurol. 2020 Sep;7(9):1557–63.

356. Schierding W, Farrow S, Fadason T, Graham OEE, Pitcher TL, Qubisi S, et al. Common Variants Coregulate Expression of GBA and Modifier Genes to Delay Parkinson's Disease Onset. Mov Disord. 2020 Aug;35(8):1346–56.

357. Szwedo AA, Dalen I, Pedersen KF, Camacho M, Bäckström D, Forsgren L, et al. GBA and APOE Impact Cognitive Decline in Parkinson's Disease: A 10-Year Population-Based Study. Mov Disord. 2022 May;37(5):1016–27.

358. Snaith RP, Bridge GW, Hamilton M. The Leeds scales for the self-assessment of anxiety and depression. Br J Psychiatry. 1976 Feb;128:156–65.

359. Spielberger CD, Gorsuch RL, Lushene RE. STAI Manual for the State-trait Anxiety Inventory ("Self-evaluation Questionnaire"). Consulting Psychologists Press; 1970. 24 p.

360. Reddy SS, Connor TE, Weeber EJ, Rebeck W. Similarities and differences in structure, expression, and functions of VLDLR and ApoER2. Mol Neurodegener. 2011 May 9;6:30.

361. Hirota Y, Kubo KI, Fujino T, Yamamoto TT, Nakajima K. ApoER2 Controls Not Only Neuronal Migration in the Intermediate Zone But Also Termination of Migration in the Developing Cerebral Cortex. Cereb Cortex. 2016 Dec 1;28(1):223–35.

362. Zarouchlioti C, Parfitt DA, Li W, Gittings LM, Cheetham ME. DNAJ Proteins in neurodegeneration: essential and protective factors. Philos Trans R Soc Lond B Biol Sci [Internet]. 2018 Jan 19;373(1738). Available from: http://dx.doi.org/10.1098/rstb.2016.0534

363. Sun AG, Wang J, Shan YZ, Yu WJ, Li X, Cong CH, et al. Identifying distinct candidate genes for early Parkinson's disease by analysis of gene expression in whole blood. Neuro Endocrinol Lett. 2014;35(5):398–404.

364. Goetz CG, Nutt JG, Stebbins GT. The Unified Dyskinesia Rating Scale: presentation and clinimetric profile. Mov Disord. 2008 Dec 15;23(16):2398–403.

365. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014 Nov;46(11):1173–86.

366. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015 Feb 12;518(7538):197–206.

367. Abdellaoui A, Yengo L, Verweij KJH, Visscher PM. 15 years of GWAS discovery: Realizing the promise. Am J Hum Genet. 2023 Feb 2;110(2):179–94.

368. Abifadel M, Varret M, Rabès JP, Allard D, Ouguerram K, Devillers M, et al. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. Nat Genet. 2003 Jun;34(2):154–6.

369. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet. 2008 Feb;40(2):161–9.

370. Horton JD, Cohen JC, Hobbs HH. Molecular biology of PCSK9: its role in LDL metabolism. Trends Biochem Sci. 2007 Feb;32(2):71–7.

371. Sabatine MS, Giugliano RP, Keech AC, Honarpour N, Wiviott SD, Murphy SA, et al. Evolocumab and clinical outcomes in patients with cardiovascular disease. N Engl J Med. 2017 May 4;376(18):1713–22.

372. Raedler LA. Praluent (alirocumab): First PCSK9 inhibitor approved by the FDA for hypercholesterolemia. Am Health Drug Benefits. 2016 Mar;9(Spec Feature):123–6.

373. Makarious MB, Leonard HL, Vitale D, Iwaki H, Sargent L, Dadu A, et al. Multi-modality machine learning predicting Parkinson's disease. NPJ Parkinsons Dis. 2022 Apr 1;8(1):35.

374. Martinez-Carrasco A, Real R, Lawton M, Iwaki H, Tan MMX, Wu L, et al. Genetic meta-analysis of levodopa induced dyskinesia in Parkinson's disease. NPJ Parkinsons Dis. 2023 Aug 31;9(1):128.

375. Martínez Carrasco A, Real R, Lawton M, Hertfelder Reynolds R, Tan M, Wu L, et al. Genome-wide Analysis of Motor Progression in Parkinson Disease. Neurol Genet. 2023 Oct;9(5):e200092.

376. Tan MMX, Lawton MA, Pollard MI, Brown E, Real R, Martinez Carrasco A, et al. Genome-wide determinants of mortality and motor progression in Parkinson's disease [Internet]. bioRxiv. 2022. Available from: https://www.medrxiv.org/content/10.1101/2022.07.07.22277297v2

377. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017 Apr 11;35(4):316–9.

378. Overview — nextflow documentation [Internet]. [cited 2024 Apr 23]. Available from: https://www.nextflow.io/docs/latest/overview.html

379. Getting Started — Longitudinal-GWAS-Pipeline 2021 documentation [Internet]. [cited 2024 Jun 28]. Available from: https://longitudinal-gwas-pipeline-osa.readthedocs.io/en/latest/getting_started.html

380. UCSC Genome Browser Downloads [Internet]. [cited 2024 May 1]. Available from: https://hgdownload.soe.ucsc.edu/downloads.html

381. Tan A, Abecasis GR, Kang HM. Unified representation of genetic variants. Bioinformatics. 2015 Jul 1;31(13):2202–4.

382. Cox DR. Regression Models and Life-Tables. J R Stat Soc Series B Stat Methodol. 1972;34(2):187–220.

383. Home [Internet]. [cited 2024 May 22]. Available from: https://amp-pd.org/

384. Commandline parameters — longitudinal-GWAS-pipeline 2021 documentation [Internet]. [cited 2024 Jun 28]. Available from: https://longitudinal-gwas-pipeline.readthedocs.io/en/latest/parameters.html

385. Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, et al. Comprehensive functional genomic resource and integrative model for the human brain. Science [Internet]. 2018 Dec 14;362(6420). Available from: http://dx.doi.org/10.1126/science.aat8464

386. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020 Sep 11;369(6509):1318–30.

387. Imai Y, Soda M, Murakami T, Shoji M, Abe K, Takahashi R. A Product of the Human Gene Adjacent to parkin Is a Component of Lewy Bodies and Suppresses Pael Receptor-induced Cell Death *. J Biol Chem. 2003 Dec 19;278(51):51901–10.

388. Taylor JM, Song YJC, Huang Y, Farrer MJ, Delatycki MB, Halliday GM, et al. Parkin Co-Regulated Gene (PACRG) is regulated by the ubiquitin-proteasomal system and is present in the pathological features of Parkinsonian diseases. Neurobiol Dis. 2007 Aug;27(2):238–47.

389. Siderowf A, Concha-Marambio L, Lafontant DE, Farris CM, Ma Y, Urenia PA, et al. Assessment of heterogeneity among participants in the Parkinson's Progression Markers Initiative cohort using α-synuclein seed amplification: a cross-sectional study. Lancet Neurol. 2023 May;22(5):407–17.

390. Rizig M, Bandres-Ciga S, Makarious MB, Ojo OO, Crea PW, Abiodun OV, et al. Identification of genetic risk loci and causal insights associated with Parkinson's disease in African and African admixed populations: a genome-wide association study. Lancet Neurol. 2023 Nov;22(11):1015–25.

391. Sinnott-Armstrong N, Naqvi S, Rivas M, Pritchard JK. GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. Elife [Internet]. 2021 Feb 15;10. Available from: http://dx.doi.org/10.7554/eLife.58615

392. Gamazon ER, Segrè AV, van de Bunt M, Wen X, Xi HS, Hormozdiari F, et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nat Genet. 2018 Jul;50(7):956–67.

393. Kim-Hellmuth S, Aguet F, Oliva M, Muñoz-Aguirre M, Kasela S, Wucher V, et al. Cell type-specific genetic regulation of gene expression across human tissues. Science [Internet]. 2020 Sep 11;369(6509). Available from: http://dx.doi.org/10.1126/science.aaz8528

394. Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, Graybuck LT, et al. Conserved cell types with divergent features in human versus mouse cortex. Nature. 2019 Sep;573(7772):61–8.

395. Lake BB, Chen S, Sos BC, Fan J, Kaeser GE, Yung YC, et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat Biotechnol. 2018 Jan;36(1):70–80.

396. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by

single-cell RNA-seq. Science. 2015 Mar 6;347(6226):1138–42.

397. Ongen H, Brown AA, Delaneau O, Panousis NI, Nica AC, GTEx Consortium, et al. Estimating the causal tissues for complex traits and diseases. Nat Genet. 2017 Dec;49(12):1676–83.

398. Skene NG, Bryois J, Bakken TE, Breen G, Crowley JJ, Gaspar HA, et al. Genetic identification of brain cell types underlying schizophrenia. Nat Genet. 2018 Jun;50(6):825–33.

399. Finucane HK, Reshef YA, Anttila V, Slowikowski K, Gusev A, Byrnes A, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat Genet. 2018 Apr;50(4):621–9.

400. Gagliano SA, Pouget JG, Hardy J, Knight J, Barnes MR, Ryten M, et al. Genomics implicates adaptive and innate immunity in Alzheimer's and Parkinson's diseases. Ann Clin Transl Neurol. 2016 Dec;3(12):924–33.

401. Reynolds RH, Botía J, Nalls MA, International Parkinson's Disease Genomics Consortium (IPDGC), System Genomics of Parkinson's Disease (SGPD), Hardy J, et al. Moving beyond neurons: the role of cell type-specific gene regulation in Parkinson's disease heritability. NPJ Parkinsons Dis. 2019 Apr 17;5:6.

402. Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, van der Zwan J, et al. Molecular Architecture of the Mouse Nervous System. Cell. 2018 Aug 9;174(4):999–1014.e22.

403. Tan MMX, Lawton MA, Pollard MI, Brown E, Real R, Carrasco AM, et al. Genome-wide determinants of mortality and motor progression in Parkinson's disease. NPJ Parkinsons Dis. 2024 Jun 7;10(1):113.

404. van Rheenen W, van der Spek RAA, Bakker MK, van Vugt JJFA, Hop PJ, Zwamborn RAJ, et al. Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. Nat Genet. 2021 Dec;53(12):1636–48.

405. Pardiñas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nat Genet. 2018 Mar;50(3):381–9.

406. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet. 2015 Oct;47(10):1121–30.

407. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. Hum Mol Genet. 2018 Oct 15;27(20):3641–9.

408. Murphy AE, Schilder BM, Skene NG. MungeSumstats: a Bioconductor package for the standardization and quality control of many GWAS summary statistics. Bioinformatics. 2021 Dec 7;37(23):4593–6.

409. Porcu E, Rüeger S, Lepik K, eQTLGen Consortium, BIOS Consortium, Santoni FA, et al. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. Nat Commun. 2019 Jul 24;10(1):3300.

410. Raitamaa L, Kautto J, Tuunanen J, Helakari H, Huotari N, Järvelä M, et al. Association of body-mass index with physiological brain pulsations across adulthood - a fast fMRI study. Int J Obes [Internet]. 2024 Mar 29; Available from: http://dx.doi.org/10.1038/s41366-024-01515-5

411. Kwon DK, Kwatra M, Wang J, Ko HS. Levodopa-Induced Dyskinesia in Parkinson's Disease: Pathogenesis and Emerging Treatment Strategies. Cells [Internet]. 2022 Nov 23;11(23). Available from: http://dx.doi.org/10.3390/cells11233736

412. Zhou FM. Chapter 25 - The striatal medium spiny neurons: what they are and how they link with Parkinson's disease. In: Martin CR, Preedy VR, editors. Genetics, Neurology, Behavior, and Diet in Parkinson's Disease. Academic Press; 2020. p. 395–412.

413. ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020 Jul;583(7818):699–710.

414. Breschi A, Muñoz-Aguirre M, Wucher V, Davis CA, Garrido-Martín D, Djebali S, et al. A limited set of transcriptional programs define major cell types. Genome Res. 2020 Jul;30(7):1047–59.

415. Mahmoud O, Dudbridge F, Davey Smith G, Munafo M, Tilling K. A robust method for collider bias correction in conditional genome-wide association studies. Nat Commun. 2022 Feb 2;13(1):619.

416. Cai M, Liu Z, Li W, Wang Y, Xie A. Association between rs823128 polymorphism and the risk of Parkinson's disease: A meta-analysis. Neurosci Lett. 2018 Feb 5;665:110–6.

417. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 2009;10:451–81.

418. Bustos BI, Billingsley K, Blauwendraat C, Gibbs JR, Gan-Or Z, Krainc D, et al. Genome-wide contribution of common Short-Tandem Repeats to Parkinson's Disease genetic risk [Internet]. bioRxiv. medRxiv; 2021. Available from: http://medrxiv.org/lookup/doi/10.1101/2021.07.01.21259645

419. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016 Oct 12;538(7624):161.

420. Bandres-Ciga S, Saez-Atienzar S, Kim JJ, Makarious MB, Faghri F, Diez-Fairen M, et al. Large-scale pathway specific polygenic risk and transcriptomic community network analysis identifies novel functional pathways in Parkinson disease. Acta Neuropathol. 2020 Sep;140(3):341–58.

421. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021 Oct;18(10):1196–203.

422. Nguyen E, Poli M, Faizi M, Thomas A, Birch-Sykes C, Wornow M, et al. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. ArXiv [Internet]. 2023 Nov 14; Available from: https://www.ncbi.nlm.nih.gov/pubmed/37426456

423. Zhou Z, Ji Y, Li W, Dutta P, Davuluri R, Liu H. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome [Internet]. arXiv [q-bio.GN]. 2023. Available from: http://arxiv.org/abs/2306.15006

424. Dalla-Torre H, Gonzalez L, Revilla JM, Carranza NL, Grzywaczewski AH, Oteri F, et al. The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics [Internet]. bioRxiv. 2023 [cited 2024 Jun 6]. p. 2023.01.11.523679. Available from: https://www.biorxiv.org/content/10.1101/2023.01.11.523679v1

425. Schiff Y, Kao CH, Gokaslan A, Dao T, Gu A, Kuleshov V. Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling [Internet]. arXiv [q-bio.GN]. 2024. Available from: http://arxiv.org/abs/2403.03234

426. Bressan E, Reed X, Bansal V, Hutchins E, Cobb MM, Webb MG, et al. The Foundational Data Initiative for Parkinson Disease: Enabling efficient translation from genetic maps to mechanism. Cell Genom. 2023 Mar 8;3(3):100261.