Parkinson's Families Project: a UK-wide study of early onset and familial Parkinson's disease

Theresa M. Schmaderer (1)*, Clodagh Towns (1)*, Simona Jasaityte (1), Manuela M. X. Tan (3), Miriam Pollard (1), Megan Hodgson (1,2), Lesley Wu (1), Russel Tilney (1), Robyn Labrum (4), Jason Hehir (4), James Polke (4), Kailash P. Bhatia (1,2), Henry Houlden (5), Nicholas W. Wood (1), Paul R. Jarman (6), Huw R. Morris (1,2)^, Raquel Real (1,2)^, on behalf of the PFP Study Group

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

2 UCL Movement Disorders Centre, University College London, London, UK

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

4 Neurogenetics Laboratory, National Hospital for Neurology & Neurosurgery, Queen Square, London, UK

5 Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London, UK

6 National Hospital for Neurology & Neurosurgery, Queen Square, London, UK

*These authors contributed equally to this work.

^ Joint senior authors.

Correspondence:

Dr Raquel Real, Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, Queen Square House, London WC1N 3BG

r.real@ucl.ac.uk

Professor Huw R Morris, Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, Royal Free Hospital, Rowland Hill Street, London NW3 2PF,

h.morris@ucl.ac.uk

Keywords: Parkinson's disease; genetics; familial PD; early onset PD; monogenic; clinical features; phenotype

Abstract

The Parkinson's Families Project is a UK-wide study aimed at identifying genetic variation associated with familial and early-onset Parkinson's disease (PD). We recruited individuals with a clinical diagnosis of PD and age at motor symptom onset ≤ 45 years and/or a family history of PD. Where possible, we also recruited affected and unaffected relatives. We analysed DNA samples with a combination of single nucleotide polymorphism (SNP) array genotyping, multiplex ligation-dependent probe amplification (MLPA), and whole genome sequencing (WGS). We investigated the association between identified pathogenic mutations and demographic and clinical factors such as age at motor symptom onset, family history, motor symptoms (MDS-UPDRS) and cognitive performance (MoCA). We completed baseline genetic analysis in 714 families, of which 196 had sporadic early-onset PD (sEOPD), 112 had familial early-onset PD (fEOPD) and 406 had late-onset familial PD (fLOPD). 53 (7.4%) of these families carried known pathogenic variants causing PD. We identified pathogenic mutations in *LRRK2* in 4.1% of families, and bi-allelic pathogenic mutations in *PRKN* in 2.4% of families. We also identified pathogenic mutations in two families with *SNCA* duplications, and single families with pathogenic mutations in *VCP*, *PINK1*, *PNPLA6*, *PLA2G6* and *SPG7*. Most early-onset and familial PD cases do not have a known genetic cause, indicating that there are likely to be further monogenic causes for PD.

Introduction

Parkinson's disease (PD) is the second most common neurodegenerative condition after Alzheimer's Disease (AD) and its prevalence is rapidly increasing[1]. PD becomes more common with advancing age, and both common and rare genetic variants can increase the risk of PD. Additionally, rare variants in genes at 20 loci have been reported to cause monogenic PD, although some of these genes have not been widely replicated, and some cause syndromes that are clinically and/or pathologically distinct from sporadic late-onset PD[2]. First-degree relatives of PD patients have been estimated to have an approximately 2-fold increased risk of developing the condition compared to unrelated individuals[3–5]. A family history of PD and an early age at onset (AAO) are associated with an increased likelihood of carrying a pathogenic mutation[6,7]. In unselected PD populations, known rare causal variants account for around 1-2% of cases, whereas they are found in around 5% of patients with familial PD and 20-40% of patients with an age of onset ≤30[8]. Pathogenic mutations in *LRRK2*, *SNCA* and *VPS35* have been consistently identified in autosomal dominant PD, and bi-allelic mutations in *PRKN* (*PARK2*), *PINK1*, *DJ-1* (*PARK7*), and *ATP13A2* (*PARK9*) in autosomal recessive PD. Rare single variants in the Gaucher disease-causing *GBA* gene are an important genetic risk factor for PD, with approximately 5-10 % of Northern European PD patients carrying single *GBA* variants[9]. For the vast majority of early-onset and familial PD cases, a known genetic cause has not been identified, suggesting either that there are additional monogenic forms to discover and/or that some PD families have more complex inheritance[10,11].

Globally, efforts are underway to collect clinical and genetic data of diagnosed PD cases to elucidate the multifactorial pathogenesis of this complex disease[12–16]. However, a major obstacle when it comes to identifying and validating candidate monogenic variants is the availability of DNA samples from affected and unaffected family members. Classic linkage analysis and whole exome/genome sequencing strategies have been used to show a causal association between genetic variation and monogenic PD, both of which require access to DNA samples from multiple family members across several generations[17].

The Parkinson's Families Project (PFP) is an ongoing UK-wide study aiming to identify new monogenic forms of PD by recruiting PD patients who are more likely to have a strong genetic contribution to the development of the condition, as well as their affected and unaffected relatives. UK-based studies of Parkinson's have previously shown that early-onset PD (EOPD) with age at symptom onset <45 years, as well as PD families with three or more affected members are particularly likely to carry a pathogenic mutation [7]. Here, we have built on this approach by recruiting early-onset and/or familial PD cases together with their genetically related family

members, to enable further genetic investigation of PD. The aims of the PFP study are: i) to build a cohort of families in which new monogenic variants may be discovered, and candidate variants may be replicated, through segregation studies; ii) to define the frequency and clinical features of known mutations in a large scale multicentre study; iii) to define a cohort of patients eligible for drug trials. PFP started recruitment in 2015 and will continue to do so until January 2030, with a target recruitment of over 1,200 families, comprising over 2,400 participants. Here, we describe the study protocol and the preliminary findings from our genetic screening of the first 714 families.

RESULTS

Cohort description

From January 1st, 2015 to February 24th, 2020, we recruited 1,035 participants from 840 families to the PFP study. Of these, we evaluated 914 individuals from 754 families using at least one of the genetic methods described below. We then excluded 40 index cases from further analysis due to either a diagnosis of secondary parkinsonism (*n* = 2), atypical parkinsonism (*n* = 5), non-parkinsonism disorder (*n* = 2), sLOPD (*n* = 13), missing clinical data (*n* = 4), consent withdrawal (*n* = 1), or failed genetic quality control (*n* = 13) (Supplementary Figure 1). Relatives of excluded index cases were also excluded (*n* = 16 relatives). In total, data were available from 858 participants from 714 families.

Baseline demographics and PD family history for the 714 index cases included are shown in Table 1. 27.5% (196/714) of index cases have sEOPD, 15.7% (112/714) have fEOPD, and 56.9% (406/714) have fLOPD. Using PCA analysis to define ancestry, 92.9% of all index cases were of European ancestry. The rate of concordance between self-reported (when available) and genetically determined ancestry was 97.3% (677/696). We recruited 144 relatives (34 affected, 110 unaffected), in family units ranging from two to six family members. 15% (*n* = 107) of the families recruited consisted of the index case plus at least one additional family member. Of these multiplex families, 72.9% consisted of the index case and one single relative, while 19.6% had two relatives recruited, and 5.6% three relatives. The remaining families had five and six individuals recruited (one each). Kinship analysis identified four families with occult relatedness. In all of these cases, individuals from the same extended family were independently recruited at different study sites.

Identification of PD-causing mutations

Following the analysis of Illumina NeuroChip genotyping array (*n* = 709 of index cases), MLPA (*n* = 540), and sequencing (*n* = 136) data, we identified known PD-causing mutations in 53 families (7.4%, 53/714; Supplementary Table 1). Mutations in autosomal dominant genes explained PD occurrence

in 32 families (4.5%; Table 2). Mutations in *LRRK2* were the most commonly identified genetic cause, accounting for PD in 29 families (4.1%). The *LRRK2* p.G2019S variant was identified in all but one of these families, and the p.R1441C variant was carried by the remaining family. One *LRRK2* p.G2019S carrier presented with an additional heterozygous frameshift mutation in *GBA* (p.L29Afs*18). The majority (*n* = 22; 75.9%) of the *LRRK2* mutation-positive families had fLOPD. Interestingly, five *LRRK2* p.G2019S carriers had sEOPD, reflecting incomplete penetrance and the likely presence of disease modifiers. One unaffected and one affected relative of a PD-LRRK2 index case was found to carry the *LRRK2* p.G2019S mutation. Penetrance of *LRRK2* p.G2019S increases with age and is influenced by genetic and environmental factors[18]. Other causal dominant mutations identified include two cases of heterozygous *SNCA* gene duplication. *SNCA* CNVs are typically associated with fEOPD[19], but both these cases presented as sEOPD. We have also identified one case with the heterozygous p.R159C *VCP* missense variant, who presented with fLOPD.

Pathogenic bi-allelic autosomal recessive mutations were identified in 21 families (2.9%; Table 3). Compound heterozygous or homozygous mutations in *PRKN* were the second most common cause of monogenic PD, accounting for PD in 17 families (2.4%). All individuals with bi-allelic *PRKN* mutations presented with early-onset PD, and nine (52.9%) cases did not have a family history of PD. Consanguinity was reported in 7.1% of bi-allelic *PRKN* mutation carriers compared to 1.6% of early-onset PD cases without mutations (P = 0.240, Fisher's Exact test). The remaining bi-allelic recessive cases carried homozygous mutations in *PINK1* (p.Y258*) and *PNPLA6* (p.P1297S), and compound heterozygous mutations in *PLA2G6* (p.T319M, p.L354P) and *SPG7* (p.N288*, p.A510V). Mutations in *PNPLA6* cause Hereditary Spastic Paraplegia 39 (OMIM 612020), but levodopa-responsive parkinsonism has been reported in association with bi-allelic mutations, generally with additional clinical features[20,21]. This case presented initially with early-onset sporadic levodopa-responsive akinetic-rigid parkinsonism, later developing lower limb spasticity and other pyramidal signs, as well as distal axonal motor neuropathy. Bi-allelic pathogenic mutations in *PLA2G6* are responsible for a broad spectrum with clinical syndromes, including autosomal recessive Parkinson's disease (OMIM 612953), frequently in association with additional clinical features such as prominent dystonia[22,23]. This case, previously reported by Magrinelli and colleagues, presented with early-onset dystonia of the right arm, followed by asymmetric akinetic-rigid parkinsonism, with good initial response to dopaminergic treatment, and pyramidal signs[22]. Finally, mutations in *SPG7* cause Hereditary Spastic Paraplegia 7 (OMIM 607259), which typically presents as pure spastic paraplegia but is often associated with complex phenotypes, including movement disorders[24]. This case presented with typical early-onset asymmetric akinetic-rigid syndrome responsive to levodopa, with later

development of subtle cerebellar signs and cerebellar atrophy on the MRI. Cases presenting with levodopa-responsive parkinsonism in association with bi-allelic *SPG7* mutations have been previously reported[25,26].

Supplementary Table 1 lists demographic data and genetic findings of all participants with a known PD-causing mutation and their relatives. We further identified 29 index cases with a single heterozygous mutation in either *PRKN* or *PINK1* (Supplementary Table 2). A list of all the unique mutations identified (*n* = 56, including *GBA* risk variants) is provided in Supplementary Table 3.

Demographic characteristics of pathogenic mutation carriers

As expected, known causal mutations were more common in participants with an early AAO, defined by symptom onset before age 46 (Table 4). We identified a monogenic cause in 9.7% (30/308) of patients with EOPD (≤ 45 years) compared to 5.7% (23/406) of patients with LOPD (χ^2 = 4.2, df = 1, P = 0.0397, Chi-squared test). However, when looking into juvenile and young onset PD (≤ 35 years), a monogenic cause was present in 20.4% (19/93) of patients with symptom onset ≤ 35, compared to 5.5% (34/621) of patients with AAO > 35 (χ2 = 26.3, df = 1, P = 2.88-06, Chi-squared test). In particular, 19.3% (18/93) of patients with symptom onset ≤ 35 carried homozygous or compound heterozygous mutations in recessive genes, compared to only 0.48% in patients with onset > 35 (3/621; P = 2.395e-14, Fisher's exact test). Among patients with a family history of PD, dominant mutations were more frequent than bi-allelic recessive mutations (4.8% vs 1.7%; χ2 = 7.8, df = 1, P = 0.005, Chi-squared test). Furthermore, each additional affected family member increased the odds of having a dominant mutation by a factor of 1.4, after adjusting the logistic regression for sex and age at symptom onset (95% confidence interval [CI] = 1.08-1.86, P = 0.0044). The majority of pathogenic mutation carriers were of European ancestry, except for one participant of South East Asian ancestry with homozygous pathogenic mutations in *PINK1* (p.Y258*), and four participants of Ashkenazi Jewish ancestry (three heterozygous *LRRK2* p.G2019S carriers and one homozygous *PNPLA6* p.P1297S carrier).

Clinical features of *LRRK2* mutation carriers

Among *LRRK2* mutation carriers, 82.8% (24/29) had a positive family history of PD and the majority experienced symptom onset > 45 years (75.9%, 22/29). Demographic characteristics of *LRRK2* mutation carriers are described in Supplementary Table 4. Clinical features of PD-*LRRK2* mutation carriers compared to mutation-negative index cases (i.e., no identified dominant or bi-allelic/single recessive mutations in PD genes) are presented in Table 2. Age at onset was similar in PD-*LRRK2* and

PD cases without mutations in PD genes (56.9 ± 12.9 vs 53.0 ± 14.7 years; P = 0.177, Mann-Whitney U test). Disease duration at study assessment was also similar between groups (6.9 ± 4.7 vs 8.3 ± 8.1 years in PD-*LRRK2* and mutation-negative PD, respectively; P = 0.897, Mann-Whitney U test). While the majority of *LRRK2* mutation carriers were European, 10.7% were of Ashkenazi Jewish ancestry compared to 0.65% of mutation-negative PD (P = 0.002, Fisher's exact test). We compared the PD motor subtype in PD-*LRRK2* and mutation-negative PD using multinomial logistic regression, adjusted for sex, age, and disease duration. PD-*LRRK2* cases have an increased odds ratio (OR) of having a postural instability and gait difficulty (PIGD)-dominant compared to a tremor-dominant motor subtype (OR = 3.2, 95%CI = 1.05 - 9.93, P = 0.041). There was no difference in motor severity, as measured by MDS-UPDRS part III, between PD-*LRRK2* and mutation-negative PD (25.7 ± 14.9 vs 26.3 vs 16.9, respectively; P = 0.922, Mann-Whitney U test). Regarding motor complications, while the rate of dystonia was similar in PD-*LRRK2* and mutation-negative PD cases, dyskinesia and motor fluctuations were more common in PD-*LRRK2* (Chi-squared test: χ^2 = 4.8, df = 1, P = 0.029, and χ^2 = 5.7, df = 1, P = 0.017, respectively). We then adjusted for sex, age, and disease duration in a logistic regression model, which confirmed the association between *LRRK2* mutations and dyskinesia and motor fluctuations (dyskinesia: OR = 3.3, 95%CI = 1.31 - 8.07, P = 0.009; motor fluctuations: OR = 3.9, 95%CI = 1.47 - 11.72, P = 0.009). No other comparisons of clinical features between PD-*LRRK2* and mutation-negative PD cases approached significance.

Clinical features of bi-allelic *PRKN* mutation carriers

The demographic and clinical features of bi-allelic *PRKN* mutation carriers are summarised in Supplementary Table 4 and Table 3, respectively. 47% (8/17) of bi-allelic *PRKN* mutation carriers had a positive family history of PD. The majority had symptom onset ≤ 35 years (82.3%, 14/17), while 35.3% (6/17) had juvenile PD (i.e., symptom onset ≤ 21). Accordingly, bi-allelic *PRKN* mutation carriers had a significantly earlier age of symptom onset compared to non-carriers (27.2 ± 9.9 vs 53.0 ± 14.7 years; P = 1.23e-09, Mann-Whitney U test). Disease duration was also significantly longer at study assessment compared to mutation-negative PD cases (22.6 ± 14.9 vs 8.34 ± 8.12; P = 1.44e-05, Mann-Whitney U test). All bi-allelic *PRKN* mutation carriers were of European ancestry. There were no differences in motor scores or motor subtype between groups. However, given that bi-allelic *PRKN* mutation carriers had significantly longer disease duration, we adjusted motor severity to disease duration by dividing MDS-UPDRS part III scores at assessment by disease duration. Bi-allelic *PRKN* mutation carriers had significantly lower adjusted motor severity scores compared to PD without a monogenic cause (1.91 ± 1.80 vs 6.58 ± 7.15; P = 1.29e-04, Mann-Whitney U test), indicating a slower rate of motor symptom progression. Concordantly, individuals with bi-allelic

PRKN mutations performed better in motor aspects of activities of daily living, as measured by MDS-UPDRS part II, after adjusting for confounding variables including disease duration (beta = -9.0, standard error = 2.12, P = 2.78e-05). The frequency of motor fluctuations was similar between bi-allelic *PRKN* mutation carriers and mutation-negative PD (50% vs 42.7%; χ^2 = 0.30, df = 1, P = 0.586, Chi-squared test). However, bi-allelic *PRKN* mutations were associated with a reduced likelihood of experiencing motor fluctuations after adjusting for disease duration (OR = 0.12, 95%CI = 0.02 - 0.57, P = 0.0097). In addition, bi-allelic *PRKN* mutation carriers were associated with reduced odds of urinary dysfunction (OR = 0.30, 95%CI = 0.09 - 0.94, P = 0.042). No other clinical features diverged significantly between bi-allelic *PRKN* mutation carriers and mutation-negative PD cases. However, it should be emphasised that, although we performed this analysis without single *PRKN* mutation carriers, results may be biased by the presence of undiagnosed bi-allelic *PRKN* cases, given that 7.5% of EOPD index cases did not undergo MLPA assay and the rate of NGS or WGS in this group was only 24.0% (Supplementary Table 5).

Clinical features of *GBA* mutation carriers

We identified 20 carriers of *GBA* variants (Table 5 and Supplementary Table 6), the majority of which (75%, 15/20) presented with a family history of PD suggestive of dominant inheritance (i.e., affected individuals in at least two generations). One *GBA* mutation carrier also had a *LRRK2* p.G2019S mutation; a second *GBA* mutation carrier had a concomitant heterozygous *PINK1* exon 5 deletion. Fifty percent (10/20) of *GBA* mutation carriers had motor symptom onset ≤ 45 years. This is in accordance with previous studies suggesting earlier symptom onset in *GBA* mutation carriers[27]. In addition, 15% (3/20) of *GBA* carriers were of Ashkenazi Jewish ancestry. *GBA* mutation carriers had significantly higher frequency of off-dystonia and dyskinesias (P = 0.036 and P = 0.049, respectively; Fisher's exact test), but the odds of developing these motor complications was not significantly higher after correcting for disease duration. Compared to mutation-negative PD, *GBA* mutation carriers had an increased risk of impulse control disorder (OR = 6.6, 95%CI = 1.83 – 24.4, P = 0.004), as previously described[28]. Other motor and other non-motor features at baseline were similar between *GBA* mutation carriers and mutation-negative PD.

Polygenic risk score analysis

A monogenic cause for PD was not identified in the vast majority of families, despite the significant enrichment in cases with early onset and/or family history of PD, which carry an increased *a priori* probability of a positive genetic finding. A further 2.8% of cases carry a *GBA* variant that increases the risk of PD. We therefore wondered if other seemingly familial cases could be the result of

increased risk of PD due to the cumulative effect of several risk variants, each contributing only a small fraction to the overall PD risk[29]. To answer this question, we calculated the PD polygenic risk score for each individual (Supplementary Figure 2), but found that unit changes in the z-transformed PRS were not associated with PD mutation status (OR = 1.01, 95%CI = -0.75 - 1.36, P = 0.947).

DISCUSSION

The UK-based PFP study consists of early-onset and familial PD cases and their relatives, with a collection of detailed demographic, clinical, lifestyle, and environmental data, as well as biological samples for genetic testing. It aims to provide support for monogenic PD gene discovery while contributing to the characterisation of genotype-phenotype relationships of known monogenic forms of PD. The first phase of genetic screening for mutations in known causal PD genes has been successfully completed for 714 families. Pathogenic causal mutations have been identified in 53 families, providing an overall diagnostic yield of 7.4% (9.7% in EOPD and 5.7% in fLOPD). This is in line with previous studies that found pathogenic mutations in known PD genes to account for 5-10% of monogenic PD cases[30].

Unsurprisingly, mutations in *LRRK2* were the most common cause of monogenic PD and were more frequent in the fLOPD group, although 17.2% of cases did not report a family history of PD and 24.1% had age of motor symptom onset ≤45 years. Age of symptom onset for *LRRK2* is reported to average 58–61 years, yet it frequently varies even within the same family[31], with the range of age at symptom onset probably reflecting the presence of disease-modifying genetic factors[32,33]. Our findings suggest that the implementation of routine diagnostic genetic testing for *LRRK2* variants in EOPD patients, even without a family history of PD, is justified. In addition, the seemingly sporadic nature of *LRRK2*-associated PD in many individuals is likely due to its incomplete penetrance, which has been extensively described[31,34,35]. While clinical characteristics are largely indistinguishable from idiopathic PD[31], it has been suggested that *LRRK2*-associated PD has a milder phenotype and slower disease progression[37]. In this study, *LRRK2* carrier status was associated with a PIGD-dominant motor subtype, which corroborates the findings by Alcalay and colleagues[38]. Other studies did not find an association between *LRRK2* carrier status and motor subtypes[7], or report an association with a tremor-dominant phenotype[39]. Comparisons across studies may however be difficult to interpret, given the different methodologies used to classify motor subtype[38–40]. We found that *LRRK2* mutations were associated with an increased risk of dyskinesia and motor fluctuations compared to mutation-negative PD cases. This is in line with a large meta-analysis by Shu and colleagues, who report an increased likelihood of developing motor complications in *LRRK2* p.G2019S carriers[41].

Other studies comparing *LRRK2*-PD with idiopathic PD did not find an association between *LRRK2* status and incidence of dyskinesias[39,42].

Bi-allelic mutations in *PRKN* were the second most frequently found and explained PD in 2.4% of families, all with EOPD. These individuals had an earlier age at symptom onset compared to mutation-negative PD cases, consistent with findings reported elsewhere[6,7]. We also observed lower MDS-UPDRS motor severity scores after adjusting for disease duration, indicating slower progression of motor symptoms compared to mutation-negative PD cases. Additionally, there was significant association between bi-allelic *PRKN* carrier status and a decrease in the MDS-UPDRS part II scores, which suggests reduced impact of motor symptoms on experiences of daily living. These findings are consistent with other studies, which have shown slower progression in bi-allelic *PRKN* carriers[6]. Previous studies have reported that postural symptoms[7], dystonia, and psychiatric symptoms may be more common in *PRKN* carriers[6,43], but we did not find evidence of this in our cohort. However, we found that bi-allelic *PRKN* mutations have a reduced likelihood of urinary dysfunction, consistent with previous reports of reduced incidence of autonomic dysfunction in *PRKN* mutation carriers[44].

In 92.6% of cases, no pathogenic mutations could be identified, which suggests that additional causative or contributing genetic factors are yet to be discovered. It is possible that not all cases with familial PD have a monogenic form of the disease. We have found a heterozygous *GBA* risk variant in 2.8% of our cohort, which increases the risk of PD in families that share *GBA* risk variants. The incidence of *GBA* mutations is significantly higher among PD patients, but the degree of pathogenicity and penetrance of different mutations is still debated[45]. Likewise, we have found a single heterozygous mutation in a recessive PD gene in another 4.1% of all index cases, including in 10.7% of familial EOPD. These could represent truly monogenic PD, where the second mutation has yet to be identified due to technical constraints. All but one single heterozygous *PRKN* and *PINK1* mutation carriers were investigated with a combination of NeuroChip and MLPA, but sequencing was performed in only 20.7%. Alternatively, some studies have indicated that heterozygous *PRKN* carriers may have an increased risk of developing PD symptoms, but with highly reduced penetrance[46–48]. Therefore, single heterozygous mutations in PD genes could increase the risk of PD in family members sharing the same risk variant. Another possibility is that familial PD can be polygenic in nature, with relatives sharing multiple risk variants, each with a small risk effect, that increase the overall risk of PD among family members that share the same genetic background[49]. However, our findings did not support this hypothesis, since there was no association between the PD polygenic risk score and mutation status (i.e., PD mutation-negative cases did not have an

increased polygenic PD risk compounded by the cumulative effect of many common risk variants, compared to mutation-positive PD). Finally, only 19% of all index cases were investigated with either gene panel next-generation or whole-genome sequencing, limiting the capacity to identify pathogenic mutations, particularly novel. NeuroChip is a reliable, high-throughput, and cost-effective screening tool for molecular diagnostics in neurodegenerative diseases[50]. However, it is limited to the identification of previously characterized and annotated variants, and thus only offers limited capabilities in exploratory genetic research. Of the 53 index cases with identified pathogenic mutations, 11.3% would not have been recognised without sequencing. We identified two novel *PRKN* frameshift mutations in compound heterozygosity that would have been missed if screening for mutations exclusively with the genotyping array. The novel frameshift *PRKN* variants p.C166Hfs*18 and p.P132Tfs*9 are predicted to cause premature termination of the parkin protein, and the resultant mRNA to be targeted for nonsense-mediated mRNA decay. An additional variant of unknown significance in *PRKN* was detected in compound heterozygosity exclusively by next-generation sequencing. The missense p.C166Y variant is not present in control population databases but has been previously reported in a PD case[51]. In addition, this variant is predicted to be deleterious by *in silico* tools, and other mutations at the same codon are classified as pathogenic. This variant is therefore also very likely to be pathogenic. In addition to novel mutations, we identified bi-allelic mutations in three genes (*PNPLA6*, *SPG7* and *PLA2G6*) that can present as early-onset levodopa-responsive parkinsonism. All of these rare cases were identified exclusively with whole-genome sequencing. We expect that additional pathogenic mutations will be identified in this cohort if sequencing methods are employed on a larger scale.

Limitations

To date, over 90% of all recruited participants are of European ancestry, meaning that mutation rates cannot be generalised across populations. Further efforts are needed to recruit individuals from other ethnic groups. Despite our efforts to recruit family members, the number of recruited relatives is still relatively small. Several reasons account for this, namely, the fact that in adult-onset disorders such as PD, family members from older generations might no longer be available for study participation. In addition, the fact that this is a cross-sectional study without longitudinal follow-up might hamper recruitment of newly affected relatives at a future date. We cannot rule out a recruitment bias inherent to the study design, given the inability to recruit all eligible PD cases in a clinic-based study as compared to a community-based study.

Conclusion

Following genetic screening for pathogenic mutation in known causal PD genes, we have identified a monogenic form of PD in 7.4% of recruited families. We have succeeded in building a cohort enriched for known causal mutations, which will aid further characterization of genotype-phenotype associations, important for accurate diagnosis and prognosis prediction. The large number of families with a seemingly strong genetic component that remain without a molecular diagnosis presents an opportunity to uncover novel causative or high-risk conferring genetic variants and will be the focus of the next phase of the analysis. Currently, efforts are being made to recruit additional relatives from these unexplained families, in particular targeting families with a very early age at symptom onset or with multiple affected family members. As more samples are whole-genome sequenced from both affected and unaffected family members, segregation studies will be possible for demonstrating gene-disease associations, thereby facilitating new genetic discoveries. In addition, unaffected mutation carriers will allow for the examination of penetrance modifiers, thus providing insights into disease mechanisms and potential drug targets. PFP will continue to recruit from currently participating and new families until 2030.

METHODS

Subjects and clinical data collection: The PFP study has been reviewed and approved by the London Camden and King's Cross Research Ethics Committee (REC – 15/LO/0097; IRAS ID – 162268) and is sponsored by the University College London Joint Research Office. The study is conducted in compliance with UK General Data Protection Regulation (GDPR) and the principles expressed in the Helsinki Declaration. PFP is registered with www. clinicaltrials.gov (NCT02760108). All participants provided written informed consent to study participation and data sharing. Participants could also opt to consent to confirmatory diagnostic genetic testing in case of a positive genetic finding, and to being re-contacted for further research studies, including therapeutic drug trials.

For this analysis, we included families recruited to PFP between 01/01/2015 and 24/02/2020, at 43 study sites across the UK (Figure 1). Eligible index cases had a clinical diagnosis of PD and met at least one of the following criteria: i) Motor symptom onset at or before the age of 45 (early onset PD); ii) At least one relative up to 3rd degree affected by PD or parkinsonism (familial PD). Whenever possible we also recruited affected and unaffected relatives of index cases. Participating individuals were at least 16 years old and had capacity to consent to participation. Participants were assessed only once during the study. For all participants, we collected demographic, environmental, medical, and family history data through questionnaires and a peripheral blood or saliva sample for DNA

extraction. We also facilitated remote participation of participants who did not live near a study site. These participants completed shortened and simplified assessment booklets from home and donated samples through their local doctor. Patient questionnaires included: Parkinson's Disease Quality of Life Questionnaire (PDQ-8), EQ-5D, Epworth Sleepiness Scale (ESS), REM Sleep Behavior Disorder Screening Questionnaire (RBDSQ), Hospital Anxiety and Depression Scale (HADS), Questionnaire for Impulsive-Compulsive Disorders in Parkinson's Disease (QUIP), Fecal Incontinence and Constipation Questionnaire, Scales for Outcomes in Parkinson's Disease - Autonomic (SCOPA-AUT), Parkinson's Disease Sleep Scale (PDSS). Affected participants recruited on-site were also subject to a standardised structured interview and completed validated scales and questionnaires by experience raters to assess motor and non-motor symptoms, including: Montreal Cognitive Assessment (MoCA), Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS), and the Modified Hoehn and Yahr Stages. Figure 1 shows an overview of the study protocol.

Participants with partially completed MDS-UPDRS ratings that fell below the threshold defined by Goetz and colleagues were excluded from downstream analyses[52]. Subjects were classified into motor subtypes (tremor dominant [TD], postural instability and gait difficulty [PIGD] or intermediate) based on the methodology defined by Stebbins and colleagues[53]. If items required for classification were missing, individuals were labelled as "unclassifiable". To account for differences in disease duration at assessment, we computed a motor severity score that consists of the ratio between the total MDS-UPDRS part III score and disease duration from reported symptom onset. Based on the MDS-UPDRS part IV, we also computed composite scores for dyskinesia (sum of items 4.1 and 4.2) and motor fluctuations (sum of items 4.3-4.5). Items of the MDS-UPDRS were categorised as present if the composite score was ≥ 1, except depression (item 1.3) and apathy (item 1.4), which were considered present only if sustained over more than one day at a time (score ≥ 2). REM sleep behaviour disorder was considered present if the RBDSQ was >5.

Clinical data storage and management: Data collected is held on REDCap® (Research Electronic Data Capture), a secure web-based Hypertext Preprocessor (PHP) software with a MySQL database back-end (https://www.project-redcap.org). It is tried and tested for use in managing clinical studies and trials, longitudinal studies and surveys[54]. The web host, network connection and storage is Information Governance Toolkit (IGT)-compliant and ISO27001-certified, according to data security best practices. Personally identifiable information is held in a database that is separated from the main study database. Members of the study team at each site only have access to records for participants recruited at their site. The databases will be maintained until 2034 for

genetic/epidemiological research, under the custodianship of Prof. Huw Morris to enable the long-term follow-up of patients recruited in this study. All clinical data were processed, stored, and disposed in accordance with all applicable legal and regulatory requirements, including the Data Protection Act 1998 and any amendments thereto.

Sample collection and storage: DNA was extracted from EDTA blood or saliva samples (saliva collection kit: Oragene® OG-500, DNA Genotek Inc.) by LGC Biosearch Technologies™. DNA is stored in secure freezers at University College London. Affected participants additionally donated ACD blood that was sent to the European Collection of Authenticated Cell Cultures (ECCAC, https://www.culturecollections.org.uk/collections/ecacc.aspx), in Wiltshire, UK, for peripheral blood lymphocytes (PBLs) extraction and transformation into lymphoblastoid cell lines. These cell lines provide an ongoing source of DNA for future studies, and may be used for disease models or the generation of induced pluripotent cell lines. Cell lines are stored at the ECACC encoded by the unique PFP study identifier.

Genetic analysis: *SNP Array Genotyping:* Quantity and purity of DNA were determined with a Qubit fluorometric assay (Invitrogen) and a NanoDrop spectrophotometer (Thermo Fisher Scientific, UK), respectively. Samples were diluted to a standard concentration in molecular grade nuclease-free water (Thermo Fisher Scientific, UK). We genotyped 909 DNA samples from 749 families using the Illumina NeuroChip array, which consists of a 306,670 SNP backbone (Infinium HumanCore-24 v1.0) with added custom content covering 179,467 neurodegenerative disease-related variants[50]. We manually clustered the genotypes using Illumina GenomeStudio v2.0 (Illumina Inc., San Diego, CA, USA), based on the protocol by Guo and colleagues[55]. We curated a list of GBA PD risk factors and PD-causing mutations, as well as all pathogenic and likely pathogenic SNVs and indels from 10 PD causing genes (*PRKN, PARK7, PINK1, ATP13A2, FBXO7, SCNA, LRRK2, VCP, VPS35, DCTN1*), from ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/, accessed on the 18/01/2023)[56]. We added any additional variants from PD-associated genes classified as definitely pathogenic in the MDSGene database (https://www.mdsgene.org/, accessed on the 21/02/2023)[57]. 131 of these variants were represented in the Neurochip array and were systematically screened for in all index cases using a custom R script (Supplementary Table 7). Mutations identified by the NeuroChip array were subsequently confirmed by diagnostic targeted mutation analysis, if patients provided consent for this.

For additional downstream analyses, we performed standard quality control in PLINK v1.9[58]. Briefly, we excluded samples with genotype missingness >5% (which can indicate poor quality of DNA sample), mismatch between clinical and genetically determined sex (which could be due to a sample mix-up), and excess heterozygosity defined as individuals who deviate > 3SD from the mean heterozygosity rate (which can indicate sample contamination)[59]. We excluded variants if the call rate was <95%. Pairwise identity-by-descent (IBD) analysis was performed to infer relatedness across all samples and identify cryptic familial relationships using the KING tool (https://www.kingrelatedness.com/)[60]. Ancestry was genetically determined using GenoTools (https://github.com/dvitale199/GenoTools). To perform polygenic risk score analysis, genotypes were imputed against the TOPMed reference panel (version R2; https://www.nhlbiwgs.org/) using the TOPMed Imputation Server (https://imputation.biodatacatalyst.nhlbi.nih.gov) using Minimac4 (version 1.7.3)[61]. Imputed variants were excluded if the imputation info R^2 score was ≤ 0.3. Following imputation, variants with missingness > 5% and minor allele frequencies < 1% were also excluded. Polygenic risk scores were calculated using PRSice-2 (https://choishingwan.github.io/PRSice/)[62], using summary statistics from the largest Parkinson's disease genome-wide association study (GWAS) to date[63].

Multiplex Ligation-dependent Probe Amplification (MLPA): Samples from 562 index cases were screened for copy number variants (CNVs) using the SALSA MLPA EK5-FAM reagent kit according to the manufacturer's instructions (MRC-Holland, Amsterdam, The Netherlands). We prioritised index cases with an AAO ≤ 35 and/or with more than two additional affected family members for CNV screening. PCR fragments were analysed by capillary electrophoresis using an ABI 3730XL genetic analyzer (Applied Biosystems). Data was analysed using the Coffalyser.Net™ (MRC-Holland) or GeneMarker® (SoftGenetics®, PA, USA) software packages, according to the supplied protocols.

Next-Generation Sequencing (NGS): DNA samples of a subset of 43 index individuals underwent diagnostic genetic screening using next-generation sequencing (Illumina MiSeq or HiSeq) of a panel of seven genes (*FBXO7*, *LRRK2*, *PRKN*, *PARK7*, *PINK1*, *SNCA*, *VPS35*) and MLPA gene dosage analysis of three genes (*PRKN*, *PINK1*, *SNCA*), as previously described. Pathogenic or likely pathogenic variants were confirmed by bi-directional Sanger sequencing.

Whole Genome Sequencing (WGS): DNA samples from 117 index cases were analysed with whole genome sequencing as part of the 100,000 Genomes Project[64]. For novel variants and variants of unknown significance (VUS) we searched population databases for allele frequencies (gnomAD,

https://gnomad.broadinstitute.org) and the published literature. We also annotated variants using Ensembl Variant Effect Predictor (https://www.ensembl.org/Tools/VEP) to obtain *in silico* pathogenicity estimates.

Statistical analyses: For statistical analysis, we classified PD cases into the following categories: i) Sporadic early-onset PD (sEOPD): motor symptom onset ≤ 45 years, no family history of PD; ii) Familial early-onset PD (fEOPD): motor symptom onset ≤ 45 years, positive family history of PD; iii) Familial late-onset PD (fLOPD): motor symptom onset > 45 years, positive family history of PD. We compared demographic and clinical features using Mann-Whitney U-test for continuous variables and Fisher's exact tests or Chi-squared tests for proportions. We investigated the effect of the *LRRK2*, *PRKN* and *GBA* genetic status on clinical features using linear regression for continuous scores or logistic regression for categorical scores, adjusting for sex, age, and disease duration at assessment, where appropriate. We used multinomial logistic regression to analyse motor subtype, using the tremor dominant group as the reference. For analysis of the modified Hoehn & Yahr stages, we used the 0-1.5 group as the reference. For the polygenic risk score analysis, scores were z-transformed and added as a covariate in a logistic regression model to predict the dependent variable (mutation status), together with age at onset, sex, and the first five genetic principal components. All p-values are two-tailed. We used R version 4.0.5 to perform statistical analyses[65].

Data availability

Anonymized datasets are available from the corresponding authors upon reasonable request.

References

1. Dorsey, E. R., Sherer, T., Okun, M. S. & Bloem, B. R. The Emerging Evidence of the Parkinson Pandemic. J. Parkinsons. Dis. 8, S3–S8 (2018).
2. Blauwendraat, C., Nalls, M. A. & Singleton, A. B. The genetic architecture of Parkinson's disease. Lancet Neurol. 19, 170–178 (2020).
3. Marder, K. et al. Risk of Parkinson's disease among first-degree relatives: A community-based study. Neurology 47, 155–160 (1996).
4. Torti, M. et al. Effect of family history, occupation and diet on the risk of Parkinson disease: A case-control study. PLoS One 15, e0243612 (2020).
5. Liu, F.-C. et al. Familial aggregation of Parkinson's disease and coaggregation with neuropsychiatric diseases: a population-based cohort study. Clin. Epidemiol. 10, 631–641 (2018).
6. Kasten, M. et al. Genotype-Phenotype Relations for the Parkinson's Disease Genes Parkin, PINK1, DJ1: MDSGene Systematic Review. Mov. Disord. 33, 730–741 (2018).
7. Tan, M. M. X. et al. Genetic analysis of Mendelian mutations in a large UK population-based Parkinson's disease study. Brain 142, 2828–2844 (2019).
8. Alcalay, R. N. et al. Frequency of known mutations in early-onset Parkinson disease: implication for genetic counseling: the consortium on risk for early onset Parkinson disease study. Arch. Neurol. 67, 1116–1122 (2010).
9. Smith, L. & Schapira, A. H. V. GBA Variants and Parkinson Disease: Mechanisms and Treatments. Cells 11, (2022).
10. Bandres-Ciga, S., Diez-Fairen, M., Kim, J. J. & Singleton, A. B. Genetics of Parkinson's disease: An introspection of its journey towards precision medicine. Neurobiol. Dis. 137, 104782 (2020).
11. Skrahina, V. et al. The Rostock International Parkinson's Disease (ROPAD) Study: Protocol and Initial Findings. Mov. Disord. 36, 1005–1010 (2021).
12. Malek, N. et al. Tracking Parkinson's: Study Design and Baseline Patient Data. J. Parkinsons. Dis. 5, 947–959 (2015).
13. Zhao, Y. et al. The role of genetics in Parkinson's disease: a large cohort study in Chinese mainland population. Brain 143, 2220–2234 (2020).
14. Kovanda, A. et al. A multicenter study of genetic testing for Parkinson's disease in the clinical setting. NPJ Parkinsons Dis 8, 149 (2022).
15. Cristina, T.-P. et al. A genetic analysis of a Spanish population with early onset Parkinson's disease. PLoS One 15, e0238098 (2020).
16. Towns, C. et al. Defining the causes of sporadic Parkinson's disease in the global Parkinson's genetics program (GP2). NPJ Parkinsons Dis 9, 131 (2023).
17. Klein, C. & Westenberger, A. Genetics of Parkinson's disease. Cold Spring Harb. Perspect. Med. 2, a008888 (2012).
18. Sierra, M. et al. High frequency and reduced penetrance of LRRK2 G2019S mutation among Parkinson's disease patients in Cantabria (Spain). Mov. Disord. 26, 2343–2346 (2011).
19. Siddiqui, I. J., Pervaiz, N. & Abbasi, A. A. The Parkinson Disease gene SNCA: Evolutionary and structural insights with pathological implication. Sci. Rep. 6, 24475 (2016).
20. Sen, K., Finau, M. & Ghosh, P. Bi-allelic variants in PNPLA6 possibly associated with Parkinsonian features in addition to spastic paraplegia phenotype. J. Neurol. 267, 2749–2753 (2020).
21. Kazanci, S. et al. PNPLA6-Related Disorder with Levodopa-Responsive Parkinsonism. Mov Disord Clin Pract 10, 338–340 (2023).
22. Magrinelli, F. et al. Dissecting the Phenotype and Genotype of PLA2G6-Related Parkinsonism. Mov. Disord. 37, 148–161 (2022).
23. Shen, T. et al. Early-Onset Parkinson's Disease Caused by PLA2G6 Compound Heterozygous Mutation, a Case Report and Literature Review. Front. Neurol. 10, 915 (2019).
24. Sáenz-Farret, M. et al. Spastic Paraplegia Type 7 and Movement Disorders: Beyond the Spastic Paraplegia. Mov Disord Clin Pract 9, 522–529 (2022).
25. Pedroso, J. L. et al. SPG7 with parkinsonism responsive to levodopa and dopaminergic deficit. Parkinsonism Relat. Disord. 47, 88–90 (2018).
26. Phillips, O., Amato, A. M. & Fernandez, H. H. Early-onset parkinsonism and hereditary spastic paraplegia type 7: pearls and pitfalls. Parkinsonism Relat. Disord. 110, (2023).

27. Sidransky, E. & Lopez, G. The link between the GBA gene and parkinsonism. Lancet Neurol. 11, 986–998 (2012).
28. Amami, P., De Santis, T., Invernizzi, F., Garavaglia, B. & Albanese, A. Impulse control behavior in GBA-mutated parkinsonian patients. J. Neurol. Sci. 421, 117291 (2021).
29. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. 50, 1219–1224 (2018).
30. Deng, H., Wang, P. & Jankovic, J. The genetics of Parkinson disease. Ageing Res. Rev. 42, 72–85 (2018).
31. Marder, K. et al. Age-specific penetrance of LRRK2 G2019S in the Michael J. Fox Ashkenazi Jewish LRRK2 Consortium. Neurology 85, 89–95 (2015).
32. Hamza, T. H. & Payami, H. The heritability of risk and age at onset of Parkinson's disease after accounting for known genetic risk factors. J. Hum. Genet. 55, 241–243 (2010).
33. Trinh, J. et al. DNM3 and genetic modifiers of age of onset in LRRK2 Gly2019Ser parkinsonism: a genome-wide linkage and association study. Lancet Neurol. 15, 1248–1256 (2016).
34. Lee, A. J. et al. Penetrance estimate of LRRK2 p.G2019S mutation in individuals of non-Ashkenazi Jewish ancestry. Mov. Disord. 32, 1432–1438 (2017).
35. Goldwurm, S. et al. Evaluation of LRRK2 G2019S penetrance: relevance for genetic counseling in Parkinson disease. Neurology 68, 1141–1143 (2007).
36. Trinh, J. et al. Genotype-phenotype relations for the Parkinson's disease genes SNCA, LRRK2, VPS35: MDSGene systematic review. Mov. Disord. 33, 1857–1870 (2018).
37. Saunders-Pullman, R. et al. Progression in the LRRK2-Asssociated Parkinson Disease Population. JAMA Neurol. 75, 312–319 (2018).
38. Alcalay, R. N. et al. Motor phenotype of LRRK2 G2019S carriers in early-onset Parkinson disease. Arch. Neurol. 66, 1517–1522 (2009).
39. Healy, D. G. et al. Phenotype, genotype, and worldwide genetic penetrance of LRRK2-associated Parkinson's disease: a case-control study. Lancet Neurol. 7, 583–590 (2008).
40. Lim, S.-Y. et al. Parkinson's disease in the Western Pacific Region. Lancet Neurol. 18, 865–879 (2019).
41. Shu, L. et al. Clinical Heterogeneity Among LRRK2 Variants in Parkinson's Disease: A Meta-Analysis. Front. Aging Neurosci. 10, 283 (2018).
42. Yahalom, G. et al. Dyskinesias in patients with Parkinson's disease: effect of the leucine-rich repeat kinase 2 (LRRK2) G2019S mutation. Parkinsonism Relat. Disord. 18, 1039–1041 (2012).
43. Koros, C., Simitsi, A. & Stefanis, L. Chapter Eight - Genetics of Parkinson's Disease: Genotype–Phenotype Correlations. in International Review of Neurobiology (eds. Bhatia, K. P., Chaudhuri, K. R. & Stamelou, M.) vol. 132 197–231 (Academic Press, 2017).
44. Tijero, B. et al. Autonomic involvement in Parkinsonian carriers of PARK2 gene mutations. Parkinsonism Relat. Disord. 21, 717–722 (2015).
45. Riboldi, G. M. & Di Fonzo, A. B. GBA, Gaucher Disease, and Parkinson's Disease: From Genetic to Clinic to New Therapeutic Approaches. Cells 8, (2019).
46. Castelo Rueda, M. P. et al. Frequency of Heterozygous Parkin (PRKN) Variants and Penetrance of Parkinson's Disease Risk Markers in the Population-Based CHRIS Cohort. Front. Neurol. 12, 706145 (2021).
47. Weissbach, A. et al. Influence of L-dopa on subtle motor signs in heterozygous Parkin- and PINK1 mutation carriers. Parkinsonism Relat. Disord. 42, 95–99 (2017).
48. Avenali, M., Blandini, F. & Cerri, S. Glucocerebrosidase Defects as a Major Risk Factor for Parkinson's Disease. Front. Aging Neurosci. 12, 97 (2020).
49. Lubbe, S. J. et al. Assessing the relationship between monoallelic PRKN mutations and Parkinson's risk. Hum. Mol. Genet. 30, 78–86 (2021).
50. Blauwendraat, C. et al. NeuroChip, an updated version of the NeuroX genotyping platform to rapidly screen for variants associated with neurological diseases. Neurobiol. Aging 57, 247.e9–247.e13 (2017).
51. Brooks, J. et al. Parkin and PINK1 mutations in early-onset Parkinson's disease: comprehensive screening in publicly available cases and control. J. Med. Genet. 46, 375–381 (2009).
52. Goetz, C. G. et al. Handling missing values in the MDS-UPDRS. Mov. Disord. 30, 1632–1638 (2015).
53. Stebbins, G. T. et al. How to identify tremor dominant and postural instability/gait difficulty groups with the movement disorder society unified Parkinson's disease rating scale: comparison with the unified Parkinson's disease rating scale. Mov. Disord. 28, 668–670 (2013).
54. Harris, P. A. et al. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J. Biomed. Inform. 42, 377–381 (2009).

55. Guo, Y. et al. Illumina human exome genotyping array clustering and quality control. Nat. Protoc. 9, 2643–2662 (2014).
56. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 46, D1062–D1067 (2018).
57. Klein, C., Hattori, N. & Marras, C. MDSGene: Closing Data Gaps in Genotype-Phenotype Correlations of Monogenic Parkinson's Disease. J. Parkinsons. Dis. 8, S25–S30 (2018).
58. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7 (2015).
59. Marees, A. T. et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int. J. Methods Psychiatr. Res. 27, e1608 (2018).
60. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. Bioinformatics 26, 2867–2873 (2010).
61. Das, S. et al. Next-generation genotype imputation service and methods. Nat. Genet. 48, 1284–1287 (2016).
62. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience 8, (2019).
63. Nalls, M. A. et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet Neurol. 18, 1091–1102 (2019).
64. Caulfield, M. et al. National Genomic Research Library. (2020) doi:10.6084/m9.figshare.4530893.v7.
65. Ripley, B. D. The R project in statistical computing. MSOR Connect. 1, 23–25 (2001).

Author Contributions

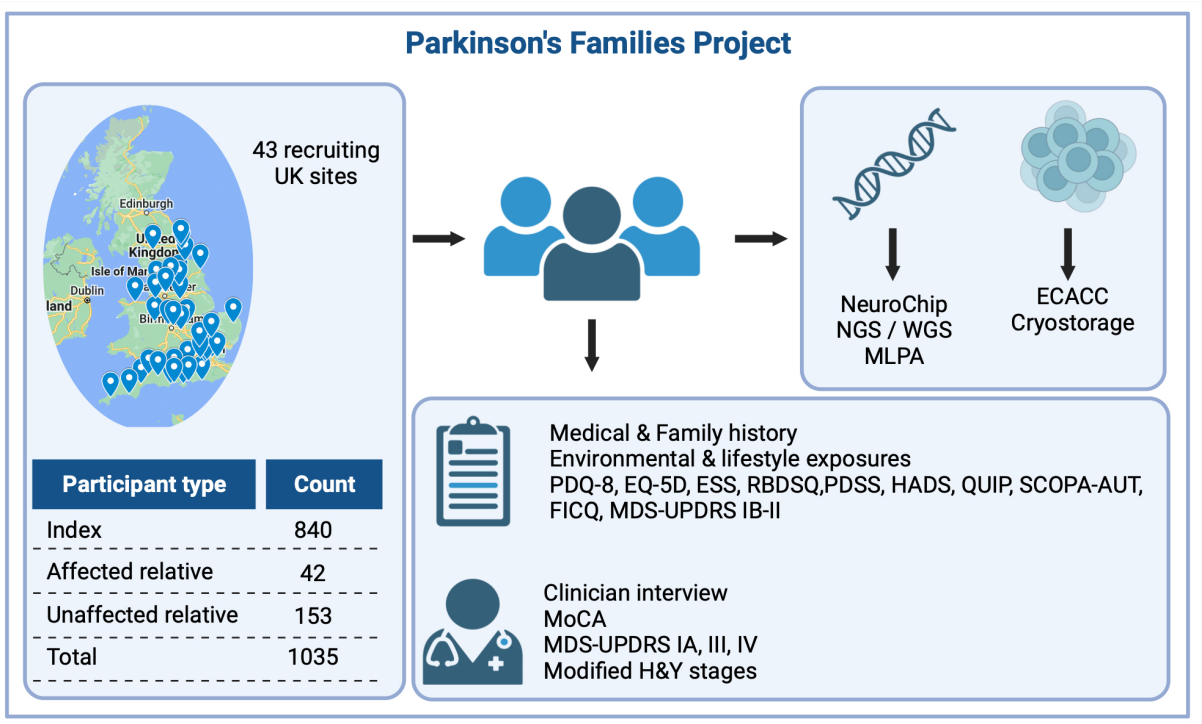
CT, RR, MMXT and HRM designed the study. MMXT, LW and RR prepared samples for SNP array genotyping. RR performed analysis of SNP array data. MMXT, CT and RR performed and analysed MLPA experiments. JH, RL and JP performed and interpreted diagnostic NGS data. CT, SJ and RR performed statistical analysis and interpreted the data. MH, MMXT, MP, RR, RT and SC collected and processed clinical data. TMS, CT and RR wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

PFP has received support from the Janet Owen's bequest fund, the Walker-Peltz charitable fund, the Medical Research Council (MRC-G0700943), Cure Parkinson's Trust, Parkinson's UK (K-1501) and the National Institute for Health Research (NIHR) Clinical Research Network (CRN) North Thames. The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript. The authors would like to thank study participants and referring clinicians, without whom this study would not be possible. A full list of PFP Study Group members is available in Supplementary Materials. Figures created with BioRender.com.

Competing interests

H.R.M. reports paid consultancy from Roche and is a co-applicant on a patent application related to C9ORF72 - Method for diagnosing a neurodegenerative disease (PCT/GB2012/052140).
All other authors declare no financial or non-financial competing interests.



Participant type	Count
Index	840
Affected relative	42
Unaffected relative	153
Total	1035

Figure 1. Parkinson's Families Project overview. Participants are recruited across 43 sites in the UK. Index cases must be ≥18 years, have capacity to consent, have a diagnosis of PD with symptom onset <45 and/or family history of PD. All participants donate a blood sample for DNA extraction. Affected participants additionally donate blood for peripheral blood lymphocyte extraction, which are sent to the European Collection of Authenticated Cell Cultures (ECACC) for transformation into lymphoblastoid cell lines and storage. All affected participants fill out a questionnaire with detailed medical and family history, environmental, drug and lifestyle exposures, as well as the following questionnaires: Parkinson's Disease Questionnaire (PDQ-8), EQ-5D, Epworth Sleepiness Scale (ESS), REM Sleep Behavior Disorder Screening Questionnaire (RBDSQ), Panic Disorder Severity Scale (PDSS), Hospital Anxiety and Depression Scale (HADS), Questionnaire for Impulsive-Compulsive Disorders in Parkinson's Disease (QUIP), Scales for Outcomes in Parkinson's disease - Autonomic Dysfunction (SCOPA-AUT), Fecal Incontinence and Constipation Questionnaire (FICQ), MDS-UPDRS parts IB and II. Affected participants recruited on-site are also assessed by an experienced investigator, who rates the MDS-UPDRS parts IA, III and IV, MoCA and Hohen & Yahr scales.

Table 1 Demographic characteristics of index cases by group

	sEOPD N = 196	fEOPD N = 112	fLOPD N = 406	Total N = 714
Sex (% Female)	40.3	42.0	43.6	42.4
Age at motor onset (Years, mean ± sd)	37.9 (6.5)	36.4 (7.7)	63.0 (8.9)	52.0 (15.1)
Age at Diagnosis (Years, mean ± sd)	42.2 (6.0)	43.8 (8.4)	65.5 (8.9)	55.8 (13.9)
Age at Assessment (Years, mean ± sd)	48.8 (8.9)	52.1 (10.2)	68.9 (8.5)	60.7 (12.9)
Disease duration at assessment (Years, mean ± sd)	10.8 (9.2)	15.6 (12.4)	5.8 (4.8)	8.7 (8.6)
Family history (%)				
No family history	100	0	0	27.5
One affected relative	0	67.9	63.3	46.6
Two affected relatives	0	21.4	25.9	18.1
Three or more affected relatives	0	10.7	10.8	7.8
Genetically Determined Ancestry (%)				
African	1	0	0.2	0.4
American	1	0.9	0	0.4
Ashkenazi Jewish	1	1.8	1.5	1.4
Central Asian	1	0	0.2	0.4
East Asian	0	0	0.2	0.1
European	89.3	94.6	94.3	92.3
Middle East	0	0	0.5	0.3
South-Asian	6.1	1.8	2.5	3.3
Unknown	0.5	0.9	0.5	1.1
Self-reported parental consanguinity (%)	2.4	0.9	0.8	1.2

sEOPD=sporadic early-onset PD. fEOPD=familial early-onset PD. fLOPD=familial late-onset PD.

Table 2 Clinical features of AD mutation carriers vs. mutation-negative index cases

	SNCA N = 2	*VCP* N = 1	*LRRK2* N = 29	Mutation-negative N = 614	Beta (95% CI)	P-value
Age at Onset (mean ± sd)	40.0 (5.7)	55-60*	56.9 (12.9)	53.0 (14.7)	2.50 (-2.24, 7.25)	0.301
Motor Features (mean ± sd)						
MDS-UPDRS Part III	49	38	25.7 (14.9)	26.3 (16.9)	0.06 (-6.69, 6.82)	0.985
Motor Severity Score	4.1	4.7	6.9 (10.1)	6.6 (7.1)	0.16 (-2.99, 3.31)	0.921
Motor Subtype (%)						
Tremor-dominant	0	0	19	38		
PIGD-dominant	100	100	80.9	51.1	1.17 (0.05, 2.30)	**0.041**
Intermediate	0	0	0	10.9	NA	NA
Hoehn and Yahr stage (%)						
0 -1.5	0	0	40.9	36.7		
2 or 2.5	0	100	31.8	39.6	-0.36 (-1.41, 0.70)	0.507
3+	100	0	27.3	23.6	0.09 (-1.02, 1.21)	0.868
Motor Complications (%)						
Dyskinesias	100	0	45.4	24.5	1.19 (0.27, 2.09)	**0.009**
Motor fluctuations	100	100	70	42.7	1.37 (0.39, 2.46)	**0.009**
Off dystonia	100	0	28.6	22.6	0.59 (-0.49, 1.56)	0.253
Motor Aspects of Daily Living (mean ± sd)	28	17	13.6 (7.6)	13.2 (9.0)	0.70 (-2.38, 3.79)	0.653
Autonomic Dysfunction (%)						
Orthostatic Hypotension	100	100	44.4	47.6	-0.12 (-0.92, 0.66)	0.764
Constipation	100	0	51.8	53.2	-0.11 (-0.90, 0.68)	0.776
Urinary Dysfunction	100	0	66.7	66.4	-0.01 (-0.83, 0.87)	0.975
REM Sleep Behaviour Disorder (%)	100	0	45.8	41	0.34 (-0.52, 1.19)	0.427
Neuropsychiatric Symptoms (%)						
Apathy	100	0	22.7	30.5	-0.33 (-1.46, 0.63)	0.531
Depression	100	0	4.5	13.2	-1.04 (-3.94, 0.57)	0.318
Anxiety	100	0	13.6	19.9	-0.37 (-1.84, 0.75)	0.563
Dopamine Dysregulation Syndrome	100	0	25	13	1.01 (-0.18, 2.06)	0.070
Psychosis	100	100	13.6	16.7	-0.13 (-1.60, 0.99)	0.836
MoCA score (mean ± sd)	27	30	26.8 (2.9)	26.1 (3.4)	0.87 (-0.55, 2.29)	0.230

*When N = 1, Age at Onset is presented as a 5-year age bracket. *LRRK2* mutation carriers vs. mutation-negative PD were compared with linear, logistic or multinomial regression as appropriate, after adjustment for sex, age and disease duration (except age at onset, which was adjusted only for sex and disease duration, and motor severity, which was adjusted only for sex and age). Significance level set at <0.05. MDS-UPDRS items were used to define the following clinical features: dyskinesias (items 4.1+4.2 >0); motor fluctuations (items 4.3+4.4+4.5 >0); off-dystonia (item 4.6 >0); orthostatic hypotension (item 1.12 >0); constipation (item 1.11 >0); urinary dysfunction (item 1.10 >0); apathy (item 1.5 >0); depression (item 1.3 >1); anxiety (item 1.4 >1), impulse control disorder (item 1.6 >0); psychosis (item 1.2 >0). Motor severity scores are MDS-UPDRD part III scores divided by disease duration in years. Motor subtypes were defined according to Stebbins *et al.*, 2013. Motor aspects of daily living scores are the sum of MDS-UPDRS part II items. REM sleep behaviour disorder was defined as RBDSQ score >5. NA=not applicable.

Table 3 Clinical features of AR mutation carriers vs. mutation-negative index cases

	PNPLA6 **N = 1**	***PINK1*** **N = 1**	***PLA2G6*** **N = 1**	***SPG7*** **N = 1**	***PRKN*** **N = 17**	**Mutation-negative N = 614**	**Beta (95% CI)**	**P-value**
Age at Onset (mean ± sd)	26-30*	26-30*	31-35*	21-25*	27.2 (9.9)	53.0 (14.7)	-13.6 (-20.0, -7.22)	**3.29E-05**
Motor Features (mean ± sd)								
MDS-UPDRS Part III	23	16	NA	NA	30.1 (17.5)	26.3 (16.9)	-3.17 (-11.9, 5.55)	0.475
Motor Severity Score	1	0.6	NA	NA	1.9 (1.8)	6.6 (7.1)	-4.01 (-7.69, -0.33)	**0.033**
Motor Subtype (%)								
Tremor-dominant	0	0	NA	NA	50	38		
PIGD-dominant	100	0	NA	NA	50	51.1	-1.03 (-2.34, 0.28)	0.122
Intermediate	0	0	NA	NA	0	10.9	NA	NA
Hoehn and Yahr stage (%)								
0 - 1.5	0	0	NA	NA	20	36.7		
2 or 2.5	100	100	NA	NA	40	39.6	-0.27 (-1.96, 1.42)	0.753
3+	0	0	NA	NA	40	23.6	0.21 (-1.69, 2.11)	0.831
Motor Complications (%)								
Dyskinesias	100	100	NA	NA	33.3	24.5	-1.05 (-2.54, 0.28)	0.139
Motor fluctuations	0	0	NA	NA	50	42.7	-2.09 (-3.75, -0.56)	**0.010**
Off dystonia	0	0	NA	NA	28.6	22.6	-1.29 (-2.82, 0.06)	0.076
Motor Aspects of Daily Living (mean ± sd)	20	NA	16	NA	10.1 (7.8)	13.2 (9.0)	-9.0 (-13.1, -4.8)	**2.78E-05**
Autonomic Dysfunction (%)								
Orthostatic Hypotension	100	NA	100	100	35.3	47.6	-1.00 (-2.16, 0.06)	0.073
Constipation	0	NA	100	100	41.2	53.2	-0.55 (-1.68, 0.52)	0.324
Urinary Dysfunction	0	NA	100	100	43.7	66.4	-1.20 (-2.40, -0.06)	**0.042**
REM Sleep Behaviour Disorder (%)	100	NA	100	NA	46.1	41	-0.44 (-1.73, 0.81)	0.493
Neuropsychiatric Symptoms (%)								
Apathy	0	0	NA	0	20	30.5	-1.30 (-2.86, -0.03)	0.065
Depression	0	0	NA	0	6.7	13.2	-1.62 (-4.57, 0.12)	0.135
Anxiety	0	0	NA	100	26.7	19.9	-0.25 (-1.64, 0.95)	0.696
Dopamine Dysregulation Syndrome	0	0	NA	0	13.3	13	-1.24 (-3.19, 0.20)	0.133
Psychosis	0	0	NA	0	6.7	16.7	-1.95 (-4.92, -0.17)	0.076
Total MoCA score (mean ± sd)	26	23	28	25	26.5 (2.6)	26.1 (3.4)	-0.07 (-1.80, 1.65)	0.933

*When N = 1, Age at Onset is presented as a 5-year age bracket. *PRKN* mutation carriers vs. mutation-negative PD were compared with linear, logistic or multinomial regression as appropriate, after adjustment for sex, age and disease duration (except age at onset, which was adjusted only for sex and disease duration, and motor severity, which was adjusted only for sex and age). Significance level set at <0.05. MDS-UPDRS items were used to define the following clinical features: dyskinesias (items 4.1+4.2 >0); motor fluctuations (items 4.3+4.4+4.5 >0); off-dystonia (item 4.6 >0); orthostatic hypotension (item 1.12 >0); constipation (item 1.11 >0); urinary dysfunction (item 1.10 >0); apathy (item 1.5 >0); depression (item 1.3 >1); anxiety (item 1.4 >1), impulse control disorder (item 1.6 >0); psychosis (item 1.2 >0). Motor severity scores are MDS-UPDRD part III scores divided by disease duration in years. Motor subtypes were defined according to Stebbins et al., 2013. Motor aspects of daily living scores are the sum of MDS-UPDRS part II items. REM sleep behaviour disorder was defined as RBDSQ score >5. NA=not applicable.

Table 4 Frequency of pathogenic and risk factor variants identified in index cases

	sEOPD N = 196	fEOPD N = 112	fLOPD N = 406	Total N = 714
PD-causing genes (%)	**9.7**	**9.8**	**5.7**	**7.4**
*LRRK2**	2.6	1.8	5.4	4.1
SNCA	1	0	0	0.3
VCP	0	0	0.2	0.1
PRKN (Bi-allelic)	4.6	7.1	0	2.4
PINK1 (Bi-allelic)	0.5	0	0	0.1
PNPLA6 (Bi-allelic)	0.5	0	0	0.1
PLA2G6 (Bi-allelic)	0.5	0	0	0.1
SPG7 (Bi-allelic)	0	0.9	0	0.1
PD risk factors (%)	**4.6**	**16.1**	**5.4**	**6.8**
*GBA**	2	5.4	2.5	2.8
PRKN (mono-allelic)	2.6	8.9	3.0	3.8
PINK1 (mono-allelic)*	0	1.8	0	0.3

*Two *GBA* mutations carriers have concomitant mono-allelic mutations in *LRRK2* and *PINK1*.
sEOPD=sporadic early-onset PD. fEOPD=familial early-onset PD. fLOPD=familial late-onset PD.

Table 5 Clinical features of GBA mutation carriers vs. mutation-negative index cases

	GBA* N = 19	Mutation-negative N = 614	Beta (95% CI)	P-value
Age at Onset (mean ± sd)	45.6 (10.3)	53.0 (14.7)	-8..23 (-14.1, -2.5)	**0.005**
Motor Features (mean ± sd)				
MDS-UPDRS Part III	31.5 (10.8)	26.3 (16.9)	5.21 (-4.23, 14.7)	0.278
Motor Severity Score	5.0 (2.7)	6.6 (7.1)	-1.13 (-5.39, 3.13)	0.603
Motor Subtype (%)				
Tremor-dominant	27.3	38		
PIGD-dominant	63.6	51.1	0.69 (-0.74, 2.12)	0.345
Intermediate	9.1	10.9	0.37 (-1.95, 2.70)	0.754
Hoehn and Yahr stage (%)				
0 - 1.5	30	36.7		
2 or 2.5	40	39.6	0.07 (-1.51, 1.65)	0.930
3+	30	23.6	0.59 (-1.18, 2.37)	0.511
Motor Complications (%)				
Dyskinesias	55.6	24.5	1.29 (-0.15, 2.78)	0.075
Motor fluctuations	55.6	42.7	0.19 (-1.34, 1.71)	0.806
Off dystonia	55.6	22.6	1.21 (-0.21, 2.69)	0.091
Motor Aspects of Daily Living (mean ± sd)	13.9 (8.5)	13.2 (9.0)	1.26 (-2.68, 5.20)	0.530
Autonomic Dysfunction (%)				
Orthostatic Hypotension	50	47.6	0.12 (-0.85, 1.09)	0.809
Constipation	55.6	53.2	0.32 (-0.64, 1.32)	0.510
Urinary Dysfunction	72.2	66.4	0.54 (-0.48, 1.71)	0.325
REM Sleep Behaviour Disorder (%)	52.9	41	0.57 (-0.46, 1.60)	0.272
Neuropsychiatric Symptoms (%)				
Apathy	36.4	30.5	0.07 (-1.31, 1.31)	0.918
Depression	18.2	13.2	0.13 (-1.79, 1.55)	0.871
Anxiety	36.4	19.9	0.71 (-0.66, 1.95)	0.272
Dopamine Dysregulation Syndrome	54.5	13	1.88 (0.60, 3.19)	**0.004**
Psychosis	36.4	16.7	0.97 (-0.45, 2.24)	0.149
MoCA score (mean ± sd)	25.3 (2.0)	26.1 (3.4)	-1.25 (-2.88, 0.38)	0.133

*Excludes one case with concomitant *LRRK2* p.G2019S mutation. NA=not applicable. *GBA* mutation carriers vs. mutation-negative PD were compared with linear, logistic or multinomial regression as appropriate, after adjustment for sex, age and disease duration (except age at onset, which was adjusted only for sex and disease duration, and motor severity, which was adjusted only for sex and age). Significance level set at <0.05. MDS-UPDRS items were used to define the following clinical features: dyskinesias (items 4.1+4.2 >0); motor fluctuations (items 4.3+4.4+4.5 >0); off-dystonia (item 4.6 >0); orthostatic hypotension (item 1.12 >0); constipation (item 1.11 >0); urinary dysfunction (item 1.10 >0); apathy (item 1.5 >0); depression (item 1.3 >1); anxiety (item 1.4 >1), impulse control disorder (item 1.6 >0); psychosis (item 1.2 >0). Motor severity scores are MDS-UPDRD part III scores divided by disease duration in years. Motor subtypes were defined according to Stebbins et al., 2013. Motor aspects of daily living scores are the sum of MDS-UPDRS part II items. REM sleep behaviour disorder was defined as RBDSQ score >5.