Human Molecular Genetics, 2015, Vol. 24, No. 7 2023–2034

doi: 10.1093/hmg/ddu616
Advance Access Publication Date: 11 December 2014
Original Article

ORIGINAL ARTICLE

Mendelian genes for Parkinson's disease contribute to the sporadic forms of the disease†

Nino Spataro[1], Francesc Calafell[1], Laura Cervera-Carles[2,3], Ferran Casals[4], Javier Pagonabarraga[2,3], Berta Pascual-Sedano[2,3], Antònia Campolongo[2,3], Jaime Kulisevsky[2,3,5], Alberto Lleó[2,3], Arcadi Navarro[1,6,7,8], Jordi Clarimón[2,3] and Elena Bosch[1,*]

[1]Institute of Evolutionary Biology (CSIC-UPF), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, 08003 Barcelona, Spain, [2]Department of Neurology, Institut d'Investigacions Biomèdiques Sant Pau-Hospital de Sant Pau, Universitat Autònoma de Barcelona, 08025 Barcelona, Spain, [3]Center for Networking Biomedical Research in Neurodegenerative Diseases (CIBERNED), Madrid, Spain, [4]Genomics Core Facility, Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), 08003 Barcelona, Spain, [5]Health Sciences Department, Universitat Oberta de Catalunya, Catalonia, Spain, [6]National Institute for Bioinformatics (INB), Barcelona Biomedical Research Park (PRBB), 08003 Barcelona, Spain, [7]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona Biomedical Research Park (PRBB), 08003 Barcelona, Spain and [8]Center for Genomic Regulation (CRG), Barcelona Biomedical Research Park (PRBB), 08003 Barcelona, Spain

*To whom correspondence should be addressed at: Elena Bosch, Institute of Evolutionary Biology (CSIC-UPF), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, C/Doctor Aiguader 88, 08003 Barcelona, Spain. Tel: +34 933160841; Fax: +34 933960901; Email: elena.bosch@upf.edu

Abstract

Parkinson's disease (PD) can be divided into familial (Mendelian) and sporadic forms. A number of causal genes have been discovered for the Mendelian form, which constitutes 10–20% of the total cases. Genome-wide association studies have successfully uncovered a number of susceptibility loci for sporadic cases but those only explain a small fraction (6–7%) of PD heritability. It has been observed that some genes that confer susceptibility to PD through common risk variants also contain rare causing mutations for the Mendelian forms of the disease. These results suggest a possible functional link between Mendelian and sporadic PD and led us to investigate the role that rare and low-frequency variants could have on the sporadic form. Through a targeting approach, we have resequenced at 49× coverage the exons and regulatory regions of 38 genes (including Mendelian and susceptibility PD genes) in 249 sporadic PD patients and 145 unrelated controls of European origin. Unlike susceptibility genes, Mendelian genes show a clear general enrichment of rare functional variants in PD cases, observed directly as well as with Tajima's *D* statistic and several collapsing methods. Our findings suggest that rare variation on PD Mendelian genes may have a role in the sporadic forms of the disease.

†Sequence data have been deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/), under accession number EGAS00001000973.

Received: September 10, 2014. **Revised:** November 7, 2014. **Accepted:** December 8, 2014

Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disorder, affecting up to 1–2% of the population over the age of 65 years (1). Over the past decade, mutations in several genes have been identified as cause of familial PD, either autosomal dominant (*SNCA*, *LRRK2*, *VPS35*) or recessive (*PARK2*, *PINK1*, *DJ1*, *ATP13A2*, *FBXO7* and *PLA2G6*) PD (2). Familial Parkinsonism has also been associated with different copy number variants in the gene encoding α-synuclein (*SNCA*) (3), one of the major constituents of the Lewy bodies, the pathological hallmark of PD. However, these Mendelian monogenic forms represent <10% of the PD cases (4). The most common form is late-onset sporadic PD, which is thought to result from complex interactions among different genetic and environmental factors. Common variants in *SNCA*, the microtubule-associated protein tau (*MAPT*) region, *LRRK2* and rare mutations of *GBA* have been repeatedly validated as genetic susceptibility factors in candidate gene association studies (2,4). Genome-wide association studies (GWAS) in PD have provided broader association evidence at several loci, but not always at the genome-wide level and with population-specific differences. GWAS on individuals of European ancestry (5–10) have confirmed the known association of PD with the *SNCA* and *MAPT* genes and identified suggestive associations with other genome regions. Recently, the use of large datasets such as those of the 23andMe database (11), the International PD Genomics Consortium and the Wellcome Trust Case Control Consortium (12) as well as the imputation of sequence variants from the 1000 Genomes Project in meta-analyses of previous GWAS (13–15), has substantially increased the number of loci achieving genome-wide significance. An exhaustive and up-to-date compilation and meta-analysis of PD association studies is freely available on the PDGene database (14). However, these susceptibility factors have been estimated to explain only a small fraction (6–7%) of the genetic variation in PD liability (11).

Many hypotheses have been proposed to explain the so-called 'missing heritability' in complex phenotypes (16). Thanks to recent advances in sequencing technologies, part of the debate has focused on the role of rare and low-frequency variants, which are not captured by the usual SNP arrays used in GWAS and do not cause sufficiently large effects to be detected in family studies (17). Indeed, recent large-scale studies of human variation report that most human genetic variation occurs at very low frequency in populations and that such a feature of the site frequency spectrum (SFS) of our species is probably the result of our recent explosive population growth (18–20). Moreover, rare coding variation is particularly enriched for deleterious alleles (21). Therefore, it seems likely that this non-common variation could have a major role in disease susceptibility (22). Interestingly, some of the most significant susceptibility genes for PD are mutated in the Mendelian forms of the disease, suggesting that the Mendelian and sporadic forms of the disease are etiologically related (8). Furthermore, the finding of rare causing mutations on susceptibility loci could suggest a continuum in the frequencies of variants influencing individual PD risk and thus that additional rare variation may also have a role in the etiology of sporadic PD.

Association tests for common variants are clearly underpowered for detecting low-frequency variants given their small number of observations (even in very large samples) (23,24). This has motivated the development of several methodologies to detect genetic association which basically collapse in different ways all the variant information in a given gene or region to compare it between different cohorts (25,26). These collapsing methods have successfully demonstrated the contribution of rare susceptibility variants in candidate genes for many diverse complex phenotypes such as colorectal cancer (27), plasma high-density lipoprotein cholesterol level (28), hypertryglyceridemia (29), type 1 diabetis (30) and blood pressure (31), among others. However, when exploring the power that these gene burden tests results would have had at the level of exome or whole-genome analysis, it seems clear that most of these candidate genes would have not reached genome-wide significance after applying multiple-test correction for the number of genes in our genome (32). Detecting the effects of rare and low-frequency variants from deep sequencing of human exomes will probably require very large sample sizes. However, it has also been suggested that accounting for the SFS on gene-by-gene basis should provide more powerful association tests (19).

In this study, we have adopted the latter approach to investigate the contribution of rare variants to the etiology of idiopathic PD. Our working hypothesis is that an excess of rare variants may indicate the involvement of a gene in a complex disease such as idiopathic PD and that such excess of rare variants can be measured, even in moderate sample sizes by using statistics coming from the field of molecular evolution (33). In particular, we used Tajima's *D* statistic (34) to test for deviations in the allele frequency spectrum between cases and controls in both individual genes and different gene groups. To this aim, we have re-sequenced, at high coverage, the coding and regulatory sequences of 38 candidate genes in a cohort of 249 idiopathic PD cases and 145 unrelated controls of European ancestry. The selected genes include genes associated with the sporadic forms of PD and genes causing Mendelian PD. Additionally, since sequencing data give access to the full genetic variability on the resequenced regions, we also explored whether common functional variants in PD susceptibility genes could explain each corresponding GWAS hit.

Results

Sequencing summary statistics and cohorts characteristics

We sequenced the protein-coding and regulatory regions of 38 genes, including 9 genes previously demonstrated to cause familial forms of PD and 33 genes that had significant association in GWAS of sporadic PD (see full list and overlap in Table 1). All sequencing data were generated with an Illumina Hiseq2000 instrument after enrichment with a custom NimbleGen array (Supplementary Material, Table S1) to a mean depth of 49× (see Materials and Methods; Supplementary Material, Figure S1 and Table S2 for more detail on depth of coverage). After base-calling and quality control analysis, we identified a total of 3649 biallelic SNPs and 377 biallelic indels (see information by gene and sample set in Supplementary Material, Table S3). From those, 3486 SNPs and 334 indels presented valid genotypes in all samples and comprise the dataset that has been used throughout the analyses (Table 2). A total of 327 SNPs were non-synonymous, and when considering at least two different prediction tools 174 of them were predicted to have functional impact (see further details in Supplementary Material, Table S4). Among the non-synonymous substitutions, we identified up to eight different instances of four previously known Mendelian mutations for Parkinson in *LRRK2* (p.G2019S) and *PARK2* (p.M192L, p.R234Q, p.T415N); all of them in idiopathic PD cases (Table 3). On the contrary, previously described mutations in two unconfirmed Mendelian loci such as *GIGYF2* (35) (p.N457T, p.T112A) and *HTRA2*

Table 1. Resequenced genes and variant counts

Gene	Chr	Size[a] (bp)	Group	GWAS	Total[b]	Splicing[b]	CAV[c]	Frameshift	Non-frameshift	Nonsense
RAB25	1	6633	Complex	y	36 (5)	4	3 (2)			
NUCKS1	1	13 060	Complex	y	99 (9)	2	3 (1)			
RAB7L1	1	8745	Complex	y	79 (10)	8	2 (1)			
GBA	1	10 706	Complex	y	62 (5)	11	12 (7)	1		
SYT11	1	10 285	Complex	y	52 (5)	6	2			
ACMSD	2	8246	Complex	y	51 (10)	6	4 (3)			1
STK39	2	12 027	Complex	y	77 (7)	5	3 (3)			
MCCC1	3	11 255	Complex	y	78 (5)	14	9 (6)			
STBD1	4	10 005	Complex	y	75 (7)	4	6 (4)			
GAK	4	24 140	Complex	y	283 (23)	24	14 (12)			
DGKQ	4	14 155	Complex	y	100 (5)	1	8 (3)			
BST1	4	8977	Complex	y	96 (5)	11	9 (8)			
SCARB2	4	12 517	Complex	y	109 (7)	5	6 (3)			
HLA-DRB5	6	5471	Complex	y	NA[d]	NA[d]	NA[d]			
GPNMB	7	12 944	Complex	y	94 (5)	18	12 (4)		1	1
FGF20	8	5914	Complex	y	62 (4)	5	2 (2)			
ITGA8	10	14 903	Complex	y	124 (11)	15 (1)	14 (9)	1		
HIP1R	12	18 470	Complex	y	168 (21)	26	17 (8)		1	
STX1B	16	12 267	Complex	y	71 (12)	10 (10)	1 (1)			
SETD1A	16	14 233	Complex	y	72 (5)	22	11 (7)			
SREBF1	17	17 253	Complex	y	112 (10)	6 (6)	18 (9)			
MED13	17	18 303	Complex	y	108 (8)	18	14 (9)			
RAI1	17	18 021	Complex	y	132 (8)	38	23 (8)		1	
MAPT	17	23 017	Complex	y	242 (28)	16	15 (6)	1		
RIT2	18	6208	Complex	y	58 (4)	3	5 (2)		1	
SNCA	4	9954	Mendelian D	y	92 (8)	1	1			
LRRK2	12	23 168	Mendelian D	y	158 (12)	36	27 (12)	1		1
VPS35	16	15 221	Mendelian D	–	85 (10)	2	1 (1)			
PINK1	1	11 171	Mendelian R	y	96 (13)	12	10 (1)			1
DJ1	1	8616	Mendelian R	y	53 (7)	4	3 (1)			
ATP13A2	1	15 133	Mendelian R	y	109 (6)	29	22 (11)			
PARK2	6	11 123	Mendelian R	y	88 (7)	9	9 (7)			
FBX07	22	13 080	Mendelian R	–	135 (13)	11	9 (3)			
PLA2G6	22	18 632	Mendelian R	–	176 (9)	24	13 (8)			
GIGYF2	2	23 952	U Mendelian	y	171 (13)	23	14 (5)		1	
HTRA2	2	6416	U Mendelian	–	30 (2)	6	4 (2)			
EIF4G1	3	18 503	U Mendelian	–	137 (12)	23	14 (5)		1	
UCHL1	4	6946	U Mendelian	y	50 (3)	2	1			

Chr, chromosome; Mendelian D, Mendelian genes in the Dominant group; Mendelian R, Mendelian genes in the Recessive group; U Mendelian, unconfirmed Mendelian genes; GWAS, genome-wide association study; y, evidence of association in GWAS; Total, total number of variants; CAV, code-altering variants (non-synonymous SNPs, nonsense mutations and coding indels); Splicing, putative splice-altering variants.

[a]Total length sequenced.

[b]In brackets, number of indels included in each category.

[c]In brackets, non-synonymous SNPs with predicted functional effects (see Materials and Methods).

[d]*HLA-DRB5* was excluded from the analysis due to low coverage. All figures refer to the total number of variants with valid genotypes for all samples.

Table 2. Gene groups and variant counts

Group	Total[a]	CAV[a]	Damaging[a]	Splicing[a]	Genes
Complex	2440 (219)	213 (7)	127 (7)	262 (1)	*RAB25, STBD1, GPNMB, FGF20, ACMSD, RIT2, STX1B, MCCC1, NUCKS1, SETD1A, SCARB2, ITGA8, SYT11, GAK, STK39, DGKQ, BST1, RAB7L1, GBA, SREBF1, MAPT, MED13, RAI1, HIP1R*
Mendelian	992 (85)	95 (1)	47 (1)	128 (0)	*PARK2, PINK1, DJ1, ATP13A2, FBX07, PLA2G6, SNCA, LRRK2, VPS35*
Dominant	335 (30)	29 (1)	15 (1)	39 (0)	*SNCA,LRRK2,VPS35*
Recessive	657 (55)	66 (0)	32 (0)	89 (0)	*PARK2, PINK1, DJ1, ATP13A2, FBX07, PLA2G6*
All	3820 (334)	341 (10)	188 (10)	444 (1)	Complex + Mendelian + *UCHL1* + *GIGYF2* + *HTRA2* + *EIF4G1*

Total, total number of variants; CAV, code-altering variants (non-synonymous SNPs, nonsense mutations and coding indels); Damaging, putative code-damaging variants (non-synonymous SNPs with predicted functional effects, nonsense mutations and coding indels); Splicing, putative splice-altering variants.

[a]In brackets, number of indels included in each category. All figures refer to the total number of variants with valid genotypes for all samples.

Table 3. Individuals carrying known Mendelian mutations

Known Mendelian mutations	Sample ID
LRRK2 (p.G2019S)	Cas213, Cas226, Cas113
PARK2 (p.R234Q)	Cas74, Cas172, Cas214
PARK2 (p.T415N)	Cas211[a]
PARK2 (p.M192L)	Cas76

[a]The same individual carried a frameshift indel in *PARK2*.

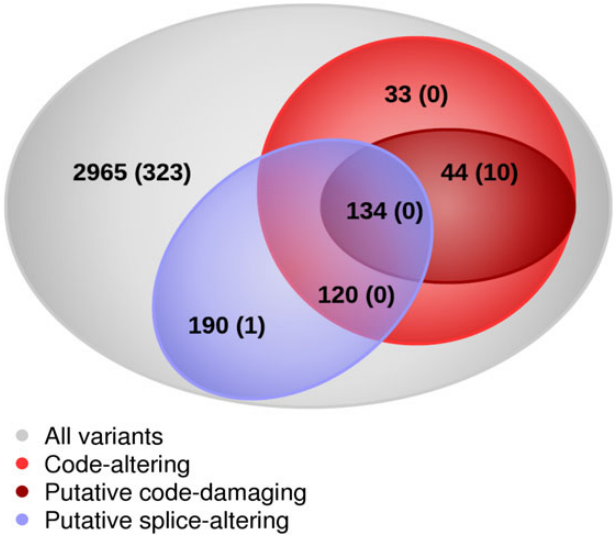

Figure 1. Summary of all biallelic variants. Numbers of different variant types and their overlap in the whole dataset of 394 individuals. In brackets, number of indels included in each category.

(36) (p.G399S) were found in three control individuals. As for indels, most were found in non-coding regions (324) but we also found four different frameshift, six non-frameshift as well as four nonsense mutations with valid genotypes in all samples (Table 2; see complete list in Supplementary Material, Tables S3 and S4). To take into account the possible differential functional impact of the SNPs and indels detected, we distinguished three groups of functional variants: code-altering (i.e. non-synonymous SNPs, nonsense mutations and exonic indels), putative code-damaging (non-synonymous SNPs with predicted functional effects, nonsense mutations and exonic indels) and putative splice-altering variants (variants predicted to alter the splicing; see details in Materials and Methods). A summary of the total number of biallelic SNPs and indels fitting each type of variant used throughout the analyses as well as their overlap is shown in Figure 1.

Principal component analysis (PCA) using all biallelic variants (SNPs and indels) displayed three clear clusters (Supplementary Material, Fig. S2A), which can be explained by the individual genotypes for the human inverted (H2) and non-inverted (H1) haplotypes in the *MAPT* (microtubule-associated protein tau) locus (37). ANOVA statistics for case–control differentiation on the first principal component (PC1) revealed significant differences ($P = 0.011$) in agreement with previous studies, in which an association was described between the H1 haplotype and PD (38). Effectively, the H1 haplotype (as inferred by considering genotypes at rs1800547 as in 38) was significantly overrepresented in PD patients compared with controls (383 out of a total of 498 PD chromosomes carried the H1 haplotype while in controls only 199 out of 290; $P = 0.0107$). After excluding all variants from the *MAPT* region (242 in total), all samples from the two cohorts fully overlapped in the first three principal components (Supplementary Material, Fig. S2B), as expected given their shared self-identified ancestry. Since different populations have different levels of polymorphism, it may be the case that heterogeneous ancestries in a sample result in one or more individuals having an excess of rare variants. Again, no outliers were identified when the number of singletons per individual was explored (Supplementary Material, Fig. S3).

Common variants

Given our limited sample size, none of the observed biallelic variants with a minor allele frequency (MAF) $\geq 5\%$ exhibited evidence of association after multiple testing correction ($P < 0.05$) when comparing cases versus controls. However, we can use our data to find possible candidate variants that would explain the association signal seen in previous GWAS. To this aim, we searched for functional variants (i.e. code-altering, putative code-damaging and splice-altering variants) in the same linkage disequilibrium blocks that contain tagSNPs previously associated with PD as described in the GWAS catalog (http://www.genome.gov/gwastudies, accessed on 10 October 2014) (39) and the PDGene database (14).

We found potential code-damaging variants for only three susceptibility genes (Supplementary Material, Table S5): a Pro to Thr change (rs11649804) in *RAI1*, which is in the same LD block and has similar frequencies than the intronic tagSNP rs11868035; a Glu to Lys change (rs3733250) in *STBD1*, which is in the same LD block and has similar frequencies than the intronic tagSNP SNP rs6812193 plus a Pro to Leu change (rs63750417) and an Arg to Trp change (rs17651549) in *MAPT* with allele frequencies matching that of human inverted (H2) haplotype. See Supplementary Material, Table S5 for a detailed catalog of all additional functional variants in each corresponding tagSNP block.

Rare variants

Of note, 68.56% of the detected biallelic variants show a MAF $\leq 1\%$ in our whole dataset of 394 individuals. However, variants categorized as putatively functional showed higher proportions below the 1% threshold. In particular, 78.60% of the putative splice-altering variants, 82.99% of the code-altering variation and 86.17% of the putative code-damaging variants displayed frequencies below 1% in our dataset.

As a first approximation to investigate whether the numbers of rare variants differed between PD cases and controls, we compared between the two cohorts the proportion of biallelic variants with MAF $\leq 1\%$ in different gene groups (Supplementary Material, Fig. S5 and Table S6). When considering all types of variants with MAF $\leq 1\%$, we detected no significant differences between cases and controls in any set containing only strictly GWAS susceptibility genes (Complex group), Mendelian genes, or both (see Fig. 2A and Supplementary Material, Fig. S5). However, when exploring for variants with a more likely functional role such as code-altering variants, PD cases displayed significantly higher proportions of variants with MAF $\leq 1\%$ but only in the Mendelian gene set ($P = 0.036$). When restricting the analysis to putative code-damaging variants or to putative splicing variants, only the Mendelian group showed a trend towards higher number of variants with MAF $\leq 1\%$ in PD cases although none of the trends reached

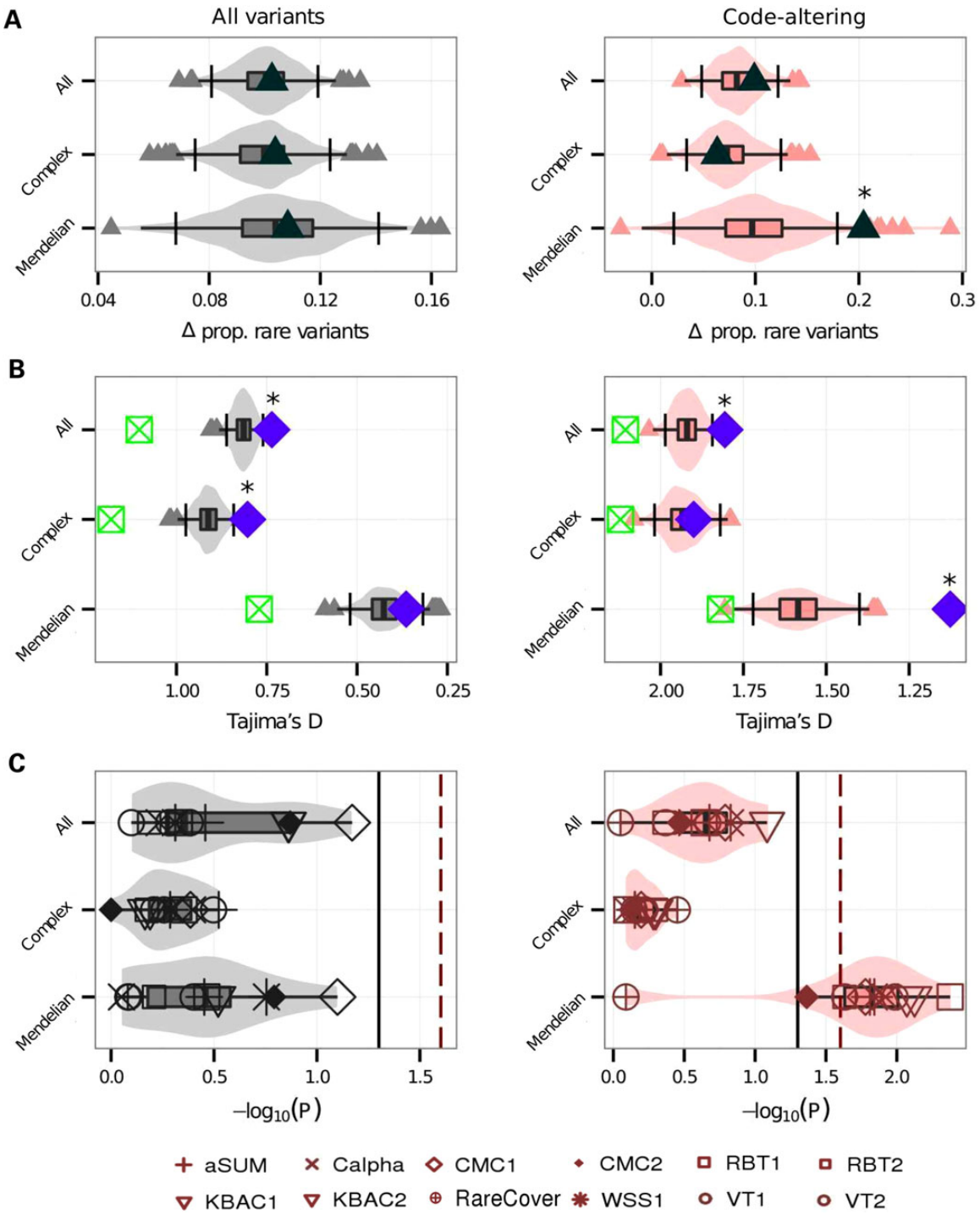

Figure 2. Cases for idiopathic Parkinson show excess of rare code-altering variation in Mendelian genes. (**A**) Difference in the proportions of rare variants (defined as those with a MAF ≤ 1%) between cases and controls when considering all types of variants (left plot), and code-altering variants (right plot). Violin plots represent the distribution of difference in proportions of rare variants when permuting 1000 times the individuals in the two cohorts. Dark green triangles represent the actual case–control difference in the proportion of rare variants. (*) Significant values ($P < 0.05$ after Bonferroni correction). (**B**) Difference of Tajima's *D* values between cases and controls when considering all types of variants (left plot), and code-altering variants (right plot). Violin plots represent the distribution of Tajima's *D* values when resampling 1000 times a subset of 145 cases; blue diamonds represent the Tajima's *D* value in controls, and green squares are the Tajima's *D* value over all the 249 PD cases. (*) Significant values ($P < 0.05$ after Bonferroni correction). (**C**) Distribution of statistical significances in several collapsing tests, when considering all types of variants (left plot) and code-altering variants (right plot). Violin plot representing the $-\log_{10}(p)$. Black line, 0.05 significance level; red dashed line, after Bonferroni correction.

significance (Supplementary Material, Fig. S5 and Table S6). Next, we repeated the same analysis but with the proportion of singletons (Supplementary Material, Table S7 and Fig. S6). As with the proportion of variants with MAF ≤ 1%, we only detected significant differences when analyzing the proportion of code-altering singletons in the Mendelian group ($P = 0.036$).

To capture further differences in the allele frequency spectrum between the two sample sets, we also computed the Tajima's *D* statistic (34), a widely used metric in evolutionary biology for testing neutrality using DNA sequence information. Given the formulation of the statistic, an excess of rare variants in a particular group (e.g. either risk variants in PD cases or protective variants in controls) for a given gene or set of genes influencing PD susceptibility will result in a smaller Tajima's *D* value in that group (see Supplementary Material, Supplementary Note). In agreement with the observed higher proportions of singletons and variants with MAF ≤ 1%, Mendelian genes displayed significantly lower Tajima's *D* values in PD cases when analyzing code-altering variants ($P < 0.001$; Fig. 2B and Supplementary Material, Table S8) but also when restricting the analysis to putative code-damaging variants ($P = 0.008$; Supplementary Material, Table S8 and Fig. S7). On the contrary, no significant differences were found on Tajima's *D* values between PD cases and controls when analyzing all variant types together on the Mendelian group (Fig. 2B and Supplementary Material, Fig. S7). Moreover, code-altering variants (and putative code-damaging variants) also displayed significantly lower Tajima's *D* values in PD cases when grouping all genes in the All group ($P = 0.012$ and $P < 0.001$, respectively) but not in the Complex set (Supplementary Material, Fig. S7 and Table S8). These results suggest that the excess of those variant types when grouping all genes probably results from the inclusion of PD Mendelian genes and that such enrichment is not a general feature of the genes associated only with the sporadic forms of PD. As for putative splicing variants, only the Mendelian group showed significantly lower Tajima's *D* values in PD cases ($P < 0.001$). Intriguingly, when using all variants we detected significantly lower Tajima's *D* values in PD cases in the All ($P < 0.001$) and the Complex groups ($P = 0.004$).

Next, we applied a set of different collapsing strategies for gene-based association analysis as available in Variant Tools (40), setting a MAF threshold of 0.5%. As described above, we performed the analysis in different gene groups, collapsing the information for all variants and for particular functional types (Fig. 2C and Supplementary Material, Table S9). Again, no association with rare variants was detected when considering all variants together, yet the analysis of code-altering variants, putative code-damaging and putative splicing variants in the Mendelian group displayed consistent associations in the CMC Fisher (41) (one-tailed), KBAC(42) (one- and two-tailed) or the RBT(43) (one-tailed test) tests. Note, however, that code-altering and putative splice-altering variants in the Mendelian group showed significant results in many other additional tests (Supplementary Material, Table S9).

The Mendelian group comprises genes of dominant and recessive inheritance. Although with lower statistical support, the previous observed trend of enrichment for rare putative functional variants in PD cases was also consistently detected in the two groups when analyzed separately (Supplementary Material, Fig. S8). In that case, only the Tajima's *D* statistic and some collapsing tests reached statistical significance (Supplementary Material, Tables S6–S9). In particular, Mendelian genes of dominant inheritance displayed significantly higher Tajima's *D* values in controls ($P = 0.012$) as well as associations in three collapsing tests only when analyzing code-altering variants. In contrast, recessive Mendelian genes displayed significant Tajima's *D* differences between cases and controls for code-altering variants ($P < 0.001$), putative code-damaging variants ($P = 0.036$) and for putative splice-altering variants (< 0.001), although only the putative splice-altering variants reached significant associations in the collapsing tests.

Genes and pathways

Given the limited number of individuals analyzed, our study is underpowered to perform gene-by-gene analyses. With the exception of the Tajima's *D* statistic, no differentiation between cases and controls remained significant when applying multiple testing corrections for the number of genes analyzed (Supplementary Material, Tables S10–S13). However, it is notable that for at least three genes (*MED13*, *SREBF1* and *LRRK2*) we observed consistent patterns of enrichment for rare variants through all the analyses (Fig. 3 and Supplementary Material, Figs S9–S12). For *MED13*, we detected trends towards higher proportions of code-altering singletons in controls and suggestive associations in three collapsing association tests (Supplementary Material, Tables S11 and S13). Similar trends were found for *SREBF1* in six collapsing association tests and when analyzing code-altering variants with MAF ≤ 1% (Supplementary Material, Tables S10 and S13). The Tajima's *D* statistic, however, displayed significantly lower values in controls for both genes ($P < 0.001$ in both genes; Supplementary Material, Fig. S10 and Table S12). Among all the Mendelian genes, *LRRK2* displayed the most clear and consistent signal for an enrichment of rare code-altering variants and putative splice-altering variants in PD cases (Fig. 3 and Supplementary Material, Figs S10 and S12). Again, only the Tajima's *D* statistic displayed significantly lower values in PD cases but only when analyzing the putative splice-altering variants ($P < 0.001$).

Finally, we also tested for rare variant enrichment in groups of genes classified by functional pathway or cellular compartment (ensuring that each group contained at least four genes): mitochondria (*DJ1*, *UCHL1*, *PINK1*, *MAPT*, *PLA2G6*, *MCC1*, *SNCA*, *HTRA2*, *LRRK2* and *PARK2*), lysosome (*ATP13A2*, *SCARB2*, *GBA*, *VPS35*) and ubiquitin (*UCHL1*, *PINK1*, *FBXO7*, *LRRK2* and *PARK2*). When analyzing all variants together no general differences were observed between cases and controls in any of the analyses for any of the three pathways with only one exception: the Tajima's *D* statistic displayed significant differences when analyzing all variants in the mitochondria pathway ($P < 0.042$; Fig. 3 and Supplementary Material, Tables S6–S9 and Fig. S13). However, the mitochondria and the ubiquitin datasets showed consistent excesses of code-altering and putative splice-altering rare variants in PD cases along the analyses. In particular, while we detected significantly lower Tajima's *D* values in PD cases in the two pathways for both variant types, only the ubiquitin set presented a significant excess of code-altering and putative splice-altering variants with MAF ≤ 1% (Supplementary Material, Table S8). In addition, in both pathways, at least five collapsing methods for rare variants reached significance when considering the code-altering variants or the putative splice-altering variants (Fig. 3C; Supplementary Material, Table S9).

Discussion

We found that most consistent differences between sporadic PD cases and controls involved potentially functional variants in genes associated with the Mendelian forms of the disease: broadly, code-altering variants with MAF ≤ 1% (and code-altering singletons) were relatively more abundant in cases; if using only code-altering variants, Tajima's *D* in Mendelian genes was significantly lower in PD cases than in controls, and tests devised specifically to account for the burden of rare variants were again significant for code-altering variants in the set of Mendelian genes. In contrast, PD susceptibility genes as group displayed no general excess of rare functional variation. And with three

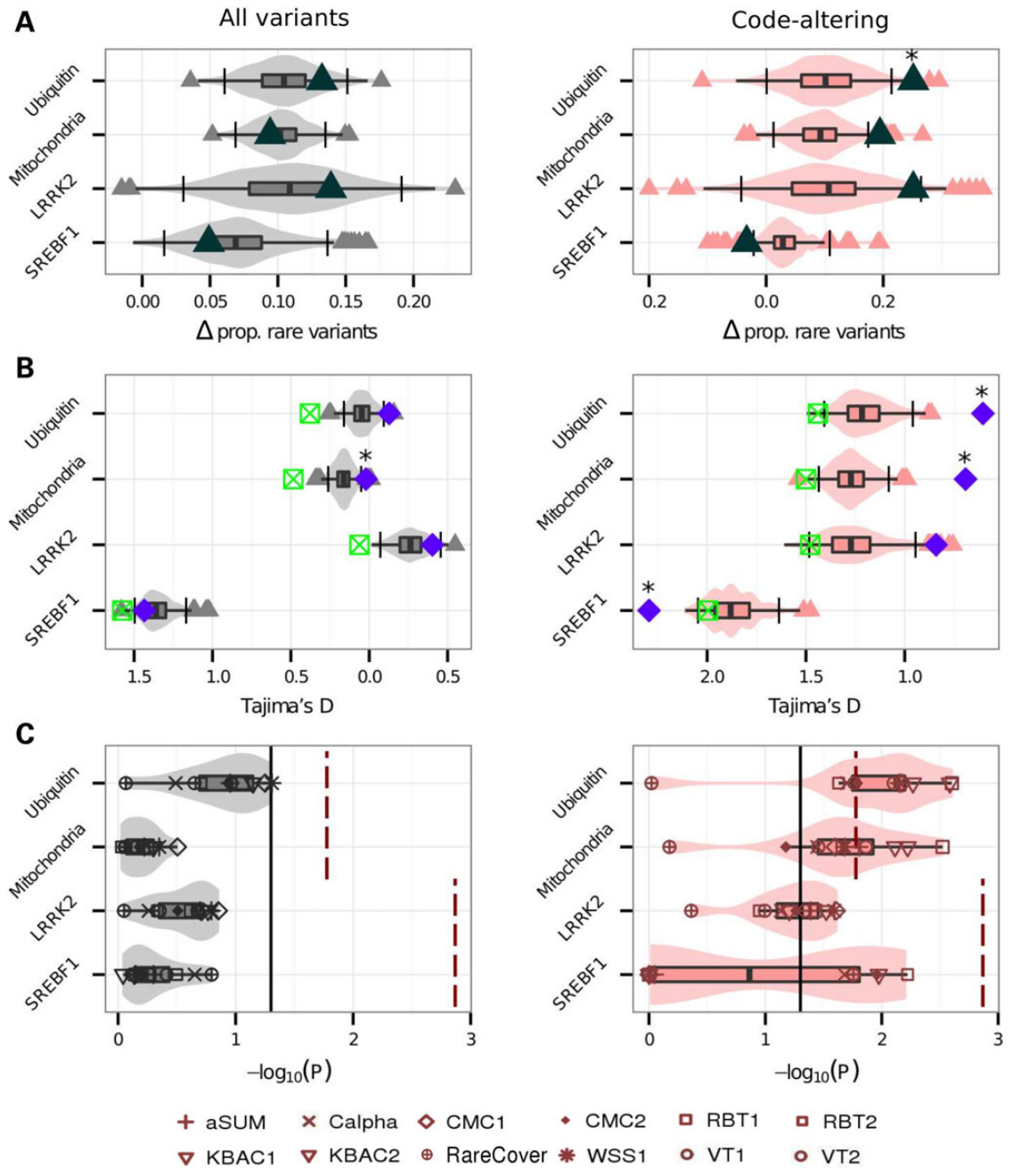

Figure 3. Excess of rare variation in particular pathways and genes. (**A**) Difference in the proportions of rare variants (defined as those with a MAF ≤ 1%) between cases and controls when considering all variants (left plot), and code-altering variants (right plot). Violin plots represent the distribution of differences in proportions of rare variants when permuting 1000 times the individuals in the two cohorts. Dark green triangles the actual case–control difference in the proportion of rare variants. (*): Significant values ($P < 0.05$ after Bonferroni correction). (**B**) Difference of Tajima's *D* values between cases and controls when considering all types of variants (left plot), and code-altering variants (right plot). Violin plots represent the distribution of Tajima's *D* values when resampling 1000 times a subset of 145 cases. Blue diamonds represent the Tajima's *D* in controls and green squares the Tajima's *D* calculated over all the 249 PD cases. (*) Significant values ($P < 0.05$ after Bonferroni correction). (**C**) Distribution of statistical significances in several collapsing tests, when considering all types of variants (left plot) and code-altering variants (right plot). Violin plot representing the $-\log_{10}(p)$. Black line, 0.05 significance level; red dashed lines, after Bonferroni correction.

exceptions, our sequence data could not contribute to identify plausible causal common variants in these genes. Thus, the causal mechanisms behind the original association signals may either lay in other regions not targeted here (i.e. in other regulatory regions not covered in our design or even in other genes within the genomic vicinity of these GWAS hits) or involve other types of regulatory variants not considered in our functional analysis.

Even with a limited number of PD cases, our study revealed that up to 3.61% of patients with sporadic PD are carriers of known Mendelian mutations (Table 3). However, when removing these mutations from the analysis, the excess of rare code-altering variation observed in sporadic PD cases on Mendelian genes remained (see Supplementary Material, Fig. S14 and Tables S6–S9). Thus, previously known Mendelian mutations cannot explain our observations. Moreover, when removing the

known Mendelian mutations as well as all nonsense mutations and coding indels, the excess of remaining non-synonymous variation in PD cases was still significant (Supplementary Material, Fig. S14). Similarly, rare code-altering variation was also detected when analyzing the mitochondria and the ubiquitin pathways, which involve different subsets of these PD Mendelian genes, respectively (Supplementary Material, Fig. S13). Although with lower statistical power, Mendelian genes also displayed a trend towards an enrichment of rare variants in PD cases when restricting the analysis to putative splice-altering or to putative code-damaging variants. Such enrichment of rare functional variants is detected not only in the whole Mendelian group but also when splitting Mendelian genes according to their mode of inheritance (Supplementary Material, Fig. S8). Finally, while all individual Mendelian genes showed a clear trend towards higher proportions of rare functional variation in PD cases, genes on the Complex group showed more diverse patterns (Supplementary Material, Figs. S10–12).

Intriguingly, when analyzing all variant types together, we detected significantly increased Tajima's *D* values in controls in the Complex group (as well as in the All group and in the mitochondria pathway). To evaluate what may cause this pattern, we decomposed Tajima's *D* into its two component summary statistics π and θ_W (Watterson's theta). Both are measures of nucleotide diversity, while π measures the per-nucleotide average number of differences between individuals, θ_W is a measure of the population mutation rate based on *S* (the number of segregating sites) (see Supplementary Material, Supplementary Note). We observed that Complex genes have decreased pairwise diversity (π) in cases relative to controls, while no differences in the number of segregating sites are generally observed (Supplementary Material, Supplementary Note, Fig. SN2). In contrast, the Tajima's *D* differences found at code-altering variants for Mendelian genes are due to an extreme excess of segregating sites in PD cases since no differences in the level of π are observed between cases and controls. Thus, only in the Mendelian group the Tajima's *D* differences can be unequivocally attributed to a differential enrichment of rare variants. Because the pattern detected in the Complex group does not directly affect functional SNPs, it does not contribute to our understanding of the underlying architecture of genetic liability for PD. Moreover, this may be due to the presence of the *MAPT* locus among the Complex genes (note that this gene is also present in the All group and in the mitochondria pathway). Indeed, when *MAPT* was removed from the analysis, differences in Tajima's *D* between cases and controls in the Complex group ceased to be significant (Supplementary Note, Fig. SN2). As explained above, *MAPT* lies in an inverted region with long range LD and two main haplotypes (37) that raise nucleotide diversity to notably high levels.

Given our sample size and the fact that each individual in our sample has a ~1% probability of carrying a code-altering singleton in a particular gene, we estimated (44) that our study design had 80% power to detect an association for any given gene if the odd ratio (OR) was 5.0. For variants with MAF ≤ 1%, this probability was 1.7% and the minimum detectable OR was 3.7. Thus, it is expected that for individual genes, we just find suggestive enrichments for code-altering singletons or code-altering variants with MAF ≤ 1%. Interestingly, these have been observed in sporadic PD cases for the *LRRK2* gene but in controls for *MED13* and *SREBF1* (and thus, probably suggesting the presence of protective variants). A similar analysis of non-synonymous variation on PD-related pathogenic and susceptibility genes in East Asians recently detected significant excess of rare variants in the *LRRK2* gene (45). Our results, in a sense, replicate the *LRRK2* results in the Foo *et al.* (2014) (45) study, and provide further evidence for the involvement of this gene in late-onset sporadic PD. However, we suggest that such an involvement is a consequence of *LRRK2* being a Mendelian gene for PD, since we find that rare code-altering variation in the Mendelian genes as a set contribute to genetic risk in late-onset PD cases. When removing *LRRK2* from the analysis, the excess of rare code-altering variation detected in the Mendelian group remained significant in the Tajima's *D* and several collapsing tests but not when analyzing the proportion of SNPs with MAF ≤ 1% (Supplementary Material, Fig. S14 and Tables S6–S9). Moreover, the same pattern is observed when removing *ATP13A2* or *FBX07* from the Mendelian group but not for any other single Mendelian gene. Therefore, we can conclude that the enrichment of rare code-altering variants is a common feature of all the Mendelian genes, even if particular genes seem to contribute more strongly.

Rare variants are seldom captured by the classical association designs. Here we have explored their role on sporadic PD through several alternative approaches, based on the comparison of the proportion of SNPs with MAF ≤ 1% and on the Tajima's *D* differences observed between cases and controls, besides applying several collapsing strategies. Tajima's *D* test explores the complete site frequency spectrum and thus may have additional power than a simple inspection of the fraction of rare variation. Given the formulation of this statistic, we expect that an excess of rare variants in a particular sample set will result in a smaller Tajima's *D* value in that sample (see Supplementary Material, Supplementary Note). Thus, we have focused on the Tajima's *D* differences between cases and controls for a given gene or set of genes but not on the actual Tajima's *D* value, which would be directly influenced by demography, and, in a few cases, by natural selection (46). Given that in our study design PD cases and controls share the same ancestry and evolutionary history, significant differences at particular loci between them cannot be attributed to past adaptive events. Although low depth next-generation sequencing data have been suggested to influence the calculation of the statistic (47), we have ensured good coverage and processed both sample sets in parallel to ensure the same potential bias (if any) in the two sample sets. All rare variant burden association tests applied gain power when compared with traditional association analysis by collectively analyzing the considered variants, but at the price of not pinpointing the actual variants involved in the disease. Thus, whereas our analysis may effectively detect groups of genes or individual genes that are enriched for rare variants, we cannot identify which variants actually contribute to susceptibility (or protection) to the disease. Still, this is valuable knowledge about the etiology of the disease, and that can be used to guide future research.

Overall, our results provide additional support for the notion that the Mendelian and sporadic forms of PD should not be regarded as two different entities, but rather as ends of a continuous spectrum of genetic architectures of disease: whereas single high penetrance mutations in particular genes cause the Mendelian forms of the disease, many different mutations in the same genes seem to contribute to sporadic PD, in addition to the low-risk common polymorphisms identified by GWAS.

Materials and Methods

Subjects

A total of 249 non-Mendelian PD patients and 145 unrelated controls, all of European origin, participated in this study, which was approved by the local institutional review board (Comitè Ètic

d'Investigació Clínica—Institut Municipal d'Assistència Sanitària, CEIC—IMAS). Written informed consent was obtained from all participants. The 249 PD patients (139 males and 110 females) were collected among outpatients regularly attending the Movement Disorders Unit, at the Hospital Sant Pau, Barcelona, Spain. The mean age of disease onset was 53.40 ± 17.97 years, while the average age at blood sampling was 66.37 ± 15.79. All PD patients fulfilled the diagnostic criteria described by Hughes *et al.* (1992) for idiopathic PD (48) and up to 39.85% of them had reported non-Mendelian family history of PD. For the 145 unrelated controls (56 males and 89 females), average age at examination was 66.23 ± 8.19 years. All control participants underwent thorough neurological examination and complete neuropsychological assessment to rule out any possible neurological illness. Genomic DNA was isolated using a standard phenol–chloroform extraction protocol for whole blood. Around 3–6 μg of genomic DNA per subject were then used to construct Illumina TrueSeq DNA libraries as described below.

Capture, sequencing and base calling

We selected 38 genes known to be associated with PD by GWASs or involved in the Mendelian forms of the disease (Table 1). These genes were extracted from the PDGene database (http://www.pdgene.org, accessed on 1 June 2012) (14). For each gene, coordinates for each possible exon were obtained from BioMart [Ensembl (49) genes 67, GRCh37.p7] and then extended in order to include putative splicing sites (with 50 intronic base pairs at both ends of each exon) and any possible regulatory sequence (retrieved from Ensembl; http://www.ensembl.org, accessed on 4 June 2012) overlapping them. We also targeted 2500 bp upstream from the transcription start site of each gene to capture the promoter region. In addition, we included up to 1000 bp of intronic sequence per gene free of any additional coding and regulatory feature as well as 100 bp centered around the SNP showing the strongest association according to the PD gene database(14) (only in the susceptibility genes).

From a total of 512 460 target bases submitted to design, 58 041 (11.25%) could not be included in the NimbleGen Sequence Capture Array (Supplementary Material, Table S1); the median fractions of bases not covered in the array were 6 and 37% in the exonic and intronic regions, respectively. All target capturing and sequencing procedures were performed at BGI Hong Kong (now BGI Tech Solutions). Briefly, genomic DNA was randomly fragmented with a Covaris instrument to yield fragments between 200 and 300 bp. Following NimbleGen's recommendations for capturing Illumina DNA libraries, adapters were ligated to both ends of the resulting fragments, which were then subsequently amplified by ligation-mediated polymerase chain reaction (LM-PCR). Upon purification and after checking the quality of the amplified DNA with a Bioanalyzer instrument, libraries were hybridized to the customized NimbleGen array for target enrichment. After washing, each captured library was then eluted, amplified and purified according to the standard manufacturer's instructions. Optimal sample quality was assessed again with a Bioanalyzer instrument and successful capturing verified by quantitative PCR. Qualified enriched libraries were then finally loaded on a Hiseq2000 platform to perform high-throughput sequencing with paired-end reads of 90 bp.

Raw reads were first mapped to the human reference genome (hg19) using the BWA aligner(50) and then subsequently processed using the GATK pipeline (51) in order to realign reads around indels, remove PCR duplicates and perform base quality score recalibration. The mean coverage per sample and target was 49.39× and 91% of the bases on target were covered at ≥15× depth (see Supplementary Material, Fig. S1 and Table S2 for details regarding coverage by gene, region and sample). Only 10 of the initial 641 fragments to target yielded a mean coverage <5× over different samples and were removed from our analysis (Supplementary Material, Table S2). We also discarded the *HLA-DRB5* gene region due to low overall mean coverage (below 10×) in comparison with the remaining targeted loci (Supplementary Material, Fig. S1 and Table S2). Variant discovery was performed using the Unified Genotyper tool of GATK with the parameters for SNPs and indel filtering described on GATK documentation for target sequencing projects. We identified a total of 4026 biallelic polymorphic variants which included 377 indels (334 of them with valid genotypes over all samples) and 3649 SNPs (3486 with valid genotypes over all samples, see Table 1). The Transition/Transversion ratio in our final dataset was 2.28. For variant annotation, we used ANNOVAR, a tool suited for functional annotations of variants detected from high-throughput sequencing data (52).

Principal component analysis (PCA) and H1/H2 MAPT haplotypes

Population substructure within cases and controls was investigated by means of principal component analysis (PCA). PCA was performed with the SmartPCA software package (53) and outputting the first four principal components (-k4). All PD cases and controls in our dataset had been previously characterized for the polymorphic inversion on 17q21 (54 or in additional unpublished genotyping). We found 100% concordance between the inferred *MAPT* haplotypes from our sequence data and their corresponding H1/H2 genotypes.

Search of functional variants in GWAs hits

SNPs associated to PD were compiled from the PDGene database (14) and the GWAS catalog (http://www.genome.gov/gwastudies, accessed on 10 October 2014) (39). Genotypes around a 1 Mb region centered on each of them were extracted from the CEU population in the 1000 genome's project (55). Blocks of LD were then inferred with PLINK version 1.07 (http://pngu.mgh.harvard.edu/purcell/plink/) (56) using default parameters. We subsequently identified which variants in our dataset were found in each block, annotated those putatively functional (code-altering variants, putative code-damaging and putative splice-altering) and compared their corresponding allele frequencies with that of the original associated tagSNP.

Concatenated groups and variant types

Genes were categorized into five different groups (see Table 2). Genes unequivocally related to Mendelian forms of Parkinson have been labeled as "Mendelian" genes and according to their mode of inheritance were further split into Mendelian "Dominant" and "Recessive" genes. Under the "Complex" category, we grouped all genes associated with idiopathic PD but without mutations known or suggested to cause Mendelian forms of Parkinson. Finally, the "All" set corresponds to the pool of all the genes sequenced in this study, including four genes (i.e. *UCHL1*, *HTRA2*, *EIF4G1*, *GIGYF2*) that have been suggested to cause Mendelian forms of PD but remain unconfirmed (35,36,57,58).

Most of the statistical analyses have been performed using all detected variants and different groups of putatively functional variants. To take into account different potential functional

impact levels, we distinguished between the following (partly overlapping) categories (Fig. 1): code-altering, putative code-damaging and putative splice-altering variants. We considered as "code-altering variants" all the non-synonymous SNPs, the nonsense mutations and exonic indels (frameshift and non-frameshift) found. We included as "putative code-damaging" all the nonsense mutations and exonic indels together with all those non-synonymous SNPs predicted to have pathological consequences by at least two different prediction algorithms after using SIFT, Polyphen, MutationTaster and MutationAssessor as implemented in ANNOVAR(52). "Splicing" variants included variants (indels and SNPs) on exonic splicing enhancers (ESE), on exonic splicing silencers (ESS) and variants predicted to affect splicing by ANNOVAR (52). All variant types were directly annotated with ANNOVAR (52), except for ESE and ESS, whose co-ordinates were retrieved from the Human Splicing Finder Version 2.4.1 (http://www.umd.be/HSF/, accessed on 7 February 2014), and variants falling in the detected motif fragments were then collected separately as ESE and ESS variants.

Rare variant association analysis

For the different gene groups and for each single gene, we performed multiple statistical analyses to assess the contribution of rare variants to PD. In these analyses, we considered only biallelic SNPs and indels having valid genotypes in all samples. Moreover, given the different sizes of the two sample sets (249 PD cases and 145 controls), two different correction strategies based on resampling and permutation have been applied when appropriate.

Proportion and number of rare variants

Singletons and variants with MAF $\leq$ 1% in the whole dataset (394 individuals) were annotated per individual (Supplementary Material, Figs S3 and S4) and in each specific case/control sample set. Differences between cases and controls in the numbers of singletons and non-singletons (as well as in the number of variants with MAF $\leq$ 1% and MAF > 1%) were then evaluated. As the number of singletons and rare variants increases with sample size (20) we proceeded as follows. First, we permuted the case/control status of the samples in the whole dataset and recalculated 1000 times the numbers of singletons and non-singleton variants (as well as of variants with MAF $\leq$ 1% and MAF > 1%) in two permuted sample sets of 249 and 145 individuals. The distribution of obtained differences on the proportion of singletons (and variants with MAF $\leq$ 1%) between the 1000 permuted sets was then used to assess the significance of the real case–control proportion difference observed by considering the 2.5 and 97.5% percentiles as thresholds. Finally, multiple testing in this and other tests was addressed by means of the Bonferroni correction, dividing α (the type I error rate) by the number of independent gene categories. All the corresponding P values in the text have been corrected by multiple testing with the Bonferroni method.

Tajima's *D* test

Tajima's D values (34) for the two original case and control PD sample sets were computed using custom Java scripts (59). To correct for the sample size differences between the two cohorts, we re-sampled 1000 times 145 individuals among the 249 cases and for each new re-sampled set a new Tajima's D value was computed as previously described. We then inferred in which percentile the observed value of Tajima's D in controls falls within these 1000 re-sampled Tajima's D values for cases and considered as significantly different only those control values that fell in the corresponding 2.5% lower and upper tails of the re-sampled distribution.

Collapsing methods

A number of association tests for rare variants were performed using the Variant Association Tools options within the Variant Tools software version 2.0 (http://varianttools.sourceforge.net/) (40): Combined and Multivariate Collapsing (CMC) (41), $c(\alpha)$ (60), Kernel Based Adaptive Clustering (KBAC) (42), Replication Based Test (RBT) (43), RareCover (61), Variable Thresholds method (VT) (62), Weighted Sum Statistic (WSS) (63) and data-adaptive Sum test (aSUM)(64). For this analysis, we considered as rare variants those with MAF $\leq$ 0.5% in the whole dataset of 394 individuals.

Supplementary Material

Supplementary Material is available at *HMG* online.

Acknowledgements

We thank all participants in the study as well as two anonymous reviewers and David Hughes for useful suggestions and comments to improve the manuscript.

Conflict of Interest statement. None declared.

Funding

This work was supported by Ministerio de Ciencia e Innovación, Spain (SAF2011-29239 to E.B. and BFU2012-38236 to A.N.), by Direcció General de Recerca, Generalitat de Catalunya (2009SGR-1101 and 2014SGR-866), by the Spanish National Institute of Bioinfomatics of the Instituto de Salud Carlos III (PT13/0001/0026), CIBERNED and by FEDER (Fondo Europeo de Desarrollo Regional)/FSE (Fondo Social Europeo).

References

1. Van Den Eeden, S.K. (2003) Incidence of Parkinson's disease: variation by age, gender, and race/ethnicity. *Am. J. Epidemiol.*, **157**, 1015–1022.
2. Clarimón, J. and Kulisevsky, J. (2013) Parkinson's disease : from genetics to clinical practice. *Curr. Genomics*, **14**, 560–567.
3. Farrer, M., Kachergus, J., Forno, L., Lincoln, S., Wang, D.-S., Hulihan, M., Maraganore, D., Gwinn-Hardy, K., Wszolek, Z., Dickson, D. *et al.* (2004) Comparison of kindreds with parkinsonism and alpha-synuclein genomic multiplications. *Ann. Neurol.*, **55**, 174–179.
4. Lesage, S. and Brice, A. (2009) Parkinson's disease: from monogenic forms to genetic susceptibility factors. *Hum. Mol. Genet.*, **18**, R48–R59.
5. Maraganore, D.M., de Andrade, M., Lesnick, T.G., Strain, K.J., Farrer, M.J., Rocca, W.A., Pant, P.V.K., Frazer, K.A., Cox, D.R. and Ballinger, D.G. (2005) High-resolution whole-genome association study of Parkinson disease. *Am. J. Hum. Genet.*, **77**, 685–693.
6. Fung, H.-C., Scholz, S., Matarin, M., Simón-Sánchez, J., Hernandez, D., Britton, A., Gibbs, J.R., Langefeld, C., Stiegert, M. L., Schymick, J. *et al.* (2006) Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first

stage analysis and public release of data. *Lancet Neurol.*, **5**, 911–916.
7. Pankratz, N., Wilk, J.B., Latourelle, J.C., DeStefano, A.L., Halter, C., Pugh, E.W., Doheny, K.F., Gusella, J.F., Nichols, W.C., Foroud, T. *et al.* (2009) Genomewide association study for susceptibility genes contributing to familial Parkinson disease. *Hum. Genet.*, **124**, 593–605.
8. Simón-Sánchez, J., Schulte, C., Bras, J.M., Sharma, M., Gibbs, J.R., Berg, D., Paisan-Ruiz, C., Lichtner, P., Scholz, S.W., Hernandez, D.G. *et al.* (2009) Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat. Genet.*, **41**, 1308–1312.
9. Edwards, T.L., Scott, W.K., Almonte, C., Burt, A., Powell, E.H., Beecham, G.W., Wang, L., Züchner, S., Konidari, I., Wang, G. *et al.* (2010) Genome-wide association study confirms SNPs in SNCA and the MAPT region as common risk factors for Parkinson disease. *Ann. Hum. Genet.*, **74**, 97–109.
10. Hamza, T.H., Zabetian, C.P., Tenesa, A., Laederach, A., Montimurro, J., Yearout, D., Kay, D.M., Doheny, K.F., Paschall, J., Pugh, E. *et al.* (2010) Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. *Nat. Genet.*, **42**, 781–785.
11. Do, C.B., Tung, J.Y., Dorfman, E., Kiefer, A.K., Drabant, E.M., Francke, U., Mountain, J.L., Goldman, S.M., Tanner, C.M., Langston, J.W. *et al.* (2011) Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS Genet*, **7**, e1002141.
12. Parkinson, I., Consortium, G., Trust, W. and Control, C. (2011) A two-stage meta-analysis identifies several new loci for Parkinson's disease. *PLoS Genet.*, **7**, e1002142.
13. Nalls, M.A., Plagnol, V., Hernandez, D.G., Sharma, M., Sheerin, U.-M., Saad, M., Simón-Sánchez, J., Schulte, C., Lesage, S., Sveinbjörnsdóttir, S. *et al.* (2011) Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet*, **377**, 641–649.
14. Lill, C.M., Roehr, J.T., McQueen, M.B., Kavvoura, F.K., Bagade, S., Schjeide, B.-M.M., Schjeide, L.M., Meissner, E., Zauft, U., Allen, N.C. *et al.* (2012) Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: the PDGene database. *PLoS Genet.*, **8**, e1002548.
15. Nalls, M.A., Pankratz, N., Lill, C.M., Do, C.B., Hernandez, D.G., Saad, M., DeStefano, A.L., Kara, E., Bras, J., Sharma, M. *et al.* (2014) Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.*, **46**, 989–993.
16. Maher, B. (2008) Personal genomes: The case of the missing heritability. *Nature*, **456**, 18–21.
17. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
18. Keinan, A. and Clark, A.G. (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, **336**, 740–743.
19. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G. *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64–69.
20. Coventry, A., Bull-Otterson, L.M., Liu, X., Clark, A.G., Maxwell, T.J., Crosby, J., Hixson, J.E., Rea, T.J., Muzny, D.M., Lewis, L.R. *et al.* (2010) Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat. Commun.*, **1**, 131.
21. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J. *et al.* (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, **493**, 216–220.
22. Lohmueller, K.E. (2014) The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genet.*, **10**, e1004379.
23. Casals, F., Idaghdour, Y., Hussin, J. and Awadalla, P. (2012) Next-generation sequencing approaches for genetic mapping of complex diseases. *J. Neuroimmunol.*, **248**, 10–22.
24. Lee, S., Abecasis, G.R., Boehnke, M. and Lin, X. (2014) Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.*, **95**, 5–23.
25. Bansal, V., Libiger, O., Torkamani, A. and Schork, N.J. (2010) Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.*, **11**, 773–785.
26. Asimit, J. and Zeggini, E. (2010) Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.*, **44**, 293–308.
27. Fearnhead, N.S., Wilding, J.L., Winney, B., Tonks, S., Bartlett, S., Bicknell, D.C., Tomlinson, I.P., Mortensen, N.J. and Bodmer, W.F. (2004) Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 15992–15997.
28. Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R. and Hobbs, H.H. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, **305**, 869–872.
29. Johansen, C.T., Wang, J., Lanktree, M.B., Cao, H., McIntyre, A. D., Ban, M.R., Martins, R.A., Kennedy, B.A., Hassell, R.G., Visser, M.E. *et al.* (2010) Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat. Genet.*, **42**, 684–687.
30. Nejentsev, S., Walker, N., Riches, D., Egholm, M. and Todd, J.A. (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, **324**, 387–389.
31. Ji, W., Foo, J.N., O'Roak, B.J., Zhao, H., Larson, M.G., Simon, D. B., Newton-Cheh, C., State, M.W., Levy, D. and Lifton, R.P. (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet*, **40**, 592–599.
32. Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L. *et al.* (2012) Exome sequencing and the genetic basis of complex traits. *Nat Genet*, **44**, 623–630.
33. Sazzini, M., Zuntini, R., Farjadian, S., Quinti, I., Ricci, G., Romeo, G., Ferrari, S., Calafell, F. and Luiselli, D. (2009) An evolutionary approach to the medical implications of the tumor necrosis factor receptor superfamily member 13B (TNFRSF13B) gene. *Genes Immun.*, **10**, 566–578.
34. Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
35. Meeus, B., Nuytemans, K., Crosiers, D., Engelborghs, S., Pals, P., Pickut, B., Peeters, K., Mattheijssens, M., Corsmit, E., Cras, P. *et al.* (2011) GIGYF2 has no major role in Parkinson genetic etiology in a Belgian population. *Neurobiol. Aging*, **32**, 308–312.
36. Simón-Sánchez, J. and Singleton, A.B. (2008) Sequencing analysis of OMI/HTRA2 shows previously reported pathogenic mutations in neurologically normal controls. *Hum. Mol. Genet.*, **17**, 1988–1993.
37. Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V.G. *et al.* (2005) A common inversion under selection in Europeans. *Nat. Genet.*, **37**, 129–137.

38. Zabetian, C.P., Hutter, C.M., Factor, S.A., Nutt, J.G., Higgins, D.S., Griffith, A., Roberts, J.W., Leis, B.C., Kay, D.M., Yearout, D. *et al.* (2007) Association analysis of MAPT H1 haplotype and subhaplotypes in Parkinson's disease. *Ann. Neurol.*, **62**, 137–144.
39. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.
40. San Lucas, F.A., Wang, G., Scheet, P. and Peng, B. (2012) Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. *Bioinformatics*, **28**, 421–422.
41. Li, B. and Leal, S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.
42. Liu, D.J. and Leal, S.M. (2010) A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet.*, **6**, e1001156.
43. Ionita-Laza, I., Buxbaum, J.D., Laird, N.M. and Lange, C. (2011) A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet.*, **7**, e1001289.
44. Purcell, S., Cherny, S.S. and Sham, P.C. (2003) Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*, **19**, 149–150.
45. Foo, J.N., Tan, L.C., Liany, H., Koh, T.H., Irwan, I.D., Ng, Y.Y., Ahmad-Annuar, A., Au, W.-L., Aung, T., Chan, A.Y.Y. *et al.* (2014) Analysis of non-synonymous-coding variants of Parkinson's disease-related pathogenic and susceptibility genes in East Asian populations. *Hum. Mol. Genet.*, **23**, 3891–3897.
46. Jobling, M., Hollox, E., Hurles, M., Kivisild, T. and Tyler-Smith, C. (2013) *Human Evolutionary Genetics*, 2nd ed. Garland Science, New York.
47. Korneliussen, T.S., Moltke, I., Albrechtsen, A. and Nielsen, R. (2013) Calculation of Tajima's *D* and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinform.*, **14**, 289.
48. Hughes, A.J., Daniel, S.E., Kilford, L. and Lees, A.J. (1992) Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. *J. Neurol. Neurosurg. Psychiatry*, **55**, 181–184.
49. Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S. *et al.* (2014) Ensembl 2014. *Nucleic Acids Res.*, **42**, D749–D755.
50. Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, **26**, 589–595.
51. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J. R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, MA., Hanna, M. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.
52. Wang, K., Li, M. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
53. Patterson, N., Price, A.L. and Reich, D. (2006) Population structure and eigenanalysis. *PLoS Genet.*, **2**, e190.
54. Setó-Salvia, N., Clarimón, J., Pagonabarraga, J., Pascual-Sedano, B., Campolongo, A., Combarros, O., Mateo, J.I., Regaña, D., Martínez-Corral, M., Marquié, M. *et al.* (2011) Dementia risk in Parkinson disease: disentangling the role of MAPT haplotypes. *Arch. Neurol.*, **68**, 359–364.
55. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R. M., Gibbs, R.A., Hurles, M.E. and McVean, G.A. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
56. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
57. Levecque, C., Destée, A., Mouroux, V., Becquet, E. and Defebvre, L. (2001) No genetic association of the Ubiquitin Carboxy-terminal Hydrolase-L1 gene S18Y polymorphism with familial Parkinson's disease. *J Neural Transm.*, **108**, 979–984.
58. Lesage, S., Condroyer, C., Klebe, S., Lohmann, E., Durif, F., Damier, P., Tison, F., Anheim, M., Honoré, A., Viallet, F. *et al.* (2012) EIF4G1 in familial Parkinson's disease: pathogenic mutations or rare benign variants? *Neurobiol. Aging*, **33**, 2233. e1–2233.e5.
59. Ramírez-Soriano, A., Ramos-Onsins, S.E., Rozas, J., Calafell, F. and Navarro, A. (2008) Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics*, **179**, 555–567.
60. Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K. and Daly, M.J. (2011) Testing for an unusual distribution of rare variants. *PLoS Genet.*, **7**, e1001322.
61. Bhatia, G., Bansal, V., Harismendy, O., Schork, N.J., Topol, E.J., Frazer, K. and Bafna, V. (2010) A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS Comput. Biol.*, **6**, e1000954.
62. Price, A.L., Kryukov, G.V., de Bakker, P.I.W., Purcell, S.M., Staples, J., Wei, L.-J. and Sunyaev, S.R. (2010) Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.*, **86**, 832–838.
63. Morris, A.P. and Zeggini, E. (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.*, **34**, 188–193.
64. Han, F. and Pan, W. (2010) A data-adaptive sum test for disease association with multiple common or rare variants. *Hum. Hered.*, **70**, 42–54.