Published in final edited form as:

Mov Disord. 2019 December ; 34(12): 1851–1863. doi:10.1002/mds.27864.

The Genetic Architecture of Parkinson Disease in Spain: Characterizing Population-Specific Risk, Differential Haplotype Structures, and Providing Etiologic Insight

A full list of authors and affiliations appears at the end of the article.

Abstract

Background: The Iberian Peninsula stands out as having variable levels of population admixture and isolation, making Spain an interesting setting for studying the genetic architecture of neurodegenerative diseases.

Objectives: To perform the largest PD genome-wide association study restricted to a single country.

Methods: We performed a GWAS for both risk of PD and age at onset in 7,849 Spanish individuals. Further analyses included population-specific risk haplotype assessments, polygenic risk scoring through machine learning, Mendelian randomization of expression, and methylation data to gain insight into disease-associated loci, heritability estimates, genetic correlations, and burden analyses.

Results: We identified a novel population-specific genome-wide association study signal at *PARK2* associated with age at onset, which was likely dependent on the c.155delA mutation. We replicated four genome-wide independent signals associated with PD risk, including *SNCA*, *LRRK2*, *KANSL1/MAPT*, and *HLA-DQB1*. A significant trend for smaller risk haplotypes at known loci was found compared to similar studies of non-Spanish origin. Seventeen PD-related genes showed functional consequence by two-sample Mendelian randomization in expression and methylation data sets. Long runs of homozygosity at 28 known genes/loci were found to be enriched in cases versus controls.

Conclusions: Our data demonstrate the utility of the Spanish risk haplotype substructure for future fine-mapping efforts, showing how leveraging unique and diverse population histories can benefit genetic studies of complex diseases. The present study points to *PARK2* as a major hallmark of PD etiology in Spain.

Keywords

age at onset; Parkinson’s disease; polygenic risk score; risk haplotype; Spanish population

*Correspondence to: Dr. Andrew Singleton and Dr. Sara Bandres-Ciga, Porter Neuroscience Center, 35 Convent Drive, Bethesda, MD 20892, USA; singleta@mail.nih.gov and sara.bandresciga@nih.gov.

Supporting Data

Additional Supporting Information may be found in the online version of this article at the publisher’s web-site.

Full financial disclosures and author roles may be found in the online version of this article.

Parkinson's disease (PD) is a complex disorder arising from the interplay of polygenic risk, environment, and stochastic factors occurring in an unpredictable manner.[1] Over the last 20 years, extensive work in molecular genetics has dissected the underlying genetic cause of several familial and early-onset patients in which disease was inherited in a Mendelian fashion. However, whereas just a small percentage of PD cases are monogenic, often exhibiting variable penetrance, the vast majority are considered to be sporadic with complex genetic influence.

An important step forward in favor of a genetic contribution to the etiology of idiopathic PD has been taken from genome-wide association studies (GWAS). The implementation of a large-scale, unbiased approach aimed at identifying genetic susceptibility factors has substantially improved our understanding of the pathogenic pathways relevant to disease. To date, 90 loci have been associated with idiopathic PD.[2] Yet, despite great advances in the field, only a small proportion of the heritable component of PD has been mapped. Single-nucleotide polymorphism (SNP)-based heritability attributed to common variants in PD is estimated to be roughly 22%.[2]

GWAS meta-analyses have been crucial to identifying and expanding the biological knowledge of novel disease risk factors. However, one of the main challenges is the heterogeneity across cohorts that might mask genetic associations specific to certain populations. Separated from the rest of Europe by the Pyrenees range of mountains and just 8.9 miles from the north coast of Africa, the Iberian Peninsula represents a cross-link between two continents and stands out as having remarkably variable levels of admixture, which has been reinforced by linguistic and geopolitical boundaries within the territory.[3] The Spanish population has a more diverse haplotypic structure in comparison with other European populations and is somehow isolated within itself in terms of global genetic structure.[4] The long-lasting migratory influences since early centuries and ulterior admixture from a number of civilizations over recent history have left their genetic imprint, thereby creating a particular genome population structure and diversity.[5] Taken together, these observations make Spain an interesting setting to comprehensively study the genetic architecture of PD and other complex diseases.

Here, we have performed the largest PD genome-wide assessment from a single country to date, by characterizing 7,849 Spanish individuals. We compared our new Spanish data set with extant data from various European ancestry populations to relate our findings in the context of larger studies. We also analyzed risk profiles, heritability, and autozygosity in this population as it relates to PD etiology. Of particular interest is our analyses leveraging the unique population genetic structure of Spain to identify smaller risk haplotype blocks than in less admixed Europeans. We envisage that the data generated from this large study, together with other concomitant efforts underway in other European populations, will be key to shed light on the molecular mechanisms involved in the disease process and might pave the way for future therapeutic interventions.

Participants and Methods

Cohort Characteristics

A total of 7,849 individuals (4,783 cases and 3,066 neurologically healthy controls) were recruited from 13 centers across Spain. Idiopathic PD patients were diagnosed by expert movement disorders neurologists following the standard criteria of the United Kingdom PD Society Brain Bank.[6] The respective ethical committees for medical research approved involvement in genetic studies, and all participants gave written informed consent. Controls were generally assessed to detect overall signs of neurological condition. Detailed demographic characteristics of each Spanish subcohort are summarized in Supporting Information Table S1, with an average age at onset (AAO) for cases of 61.23 ± 11.47, age at recruitment for controls of 61.79 ± 11.05, 42.27% of female cases, and 56.06% female controls.

Genotyping and Quality-Control Analyses

Samples were genotyped using the customized NeuroChip Array (v.1.0 or v.1.1; Illumina, San Diego, CA).[7] NeuroChip Array (v1.0; old version) contains a backbone of 306,670 variants whereas NeuroChip Array (v1.1; new version) contains 307,907 variants. These tagging variants, based on Infinium HumanCore-24 v1.0, densely cover ancestry informative markers, markers for determination of identity by descent, and X-chromosome SNPs for sex determination. Additionally, both versions of NeuroChip contain a custom content consisting of 179,467 neurodegenerative disease-related variants.

Genotypes were clustered using Illumina GenomeStudio (v.2.0). Quality-control (QC) analysis was performed as follows: Samples with call rates of <95% and whose genetically determined sex from X-chromosome heterogeneity did not match that from clinical data were excluded from the analysis. Samples exhibiting excess heterozygosity estimated by an F statistic > ± 0.25 were also excluded. Once preliminary sample-level QC was completed, SNPs with minor allele frequency (MAF) <0.01, Hardy-Weinberg equilibrium *P* value <1E-5, and missingness rates >5% were excluded. Genetic variants passing QC numbered 433,768 SNPs. Genotyped SNPs thought to be in linkage disequilibrium (LD) in a sliding window of 50 adjacent SNPs, which scrolled through the genome at a rate of five overlapping SNPs, were also removed from the following analyses (as were palin-dromic SNPs). Next, samples were clustered using principal component analysis (PCA) to evaluate European ancestry as compared to the HapMap3 CEU/TSI populations (International HapMap Consortium, 2003; Supporting Information Fig. S1). Confirmed European-ancestry samples were extracted and principal components (PCs) 1 to 20 were used as covariates in all analysis. Samples closely related (sharing proportionally more than 18.5% of alleles) were dropped from the following analysis. After filtering, 4,639 cases and 2,949 controls remained. The data were then merged by overlapping variants between both platforms and imputed using the *Haplotype Reference Consortium r1.1 2016* (http://www.haplotype-reference-consortium.org), under default settings with phasing using the EAGLE option. Imputed variants numbered 9,911,207 after filtering for MAF >1% and imputation quality (RSQ) >0.3.

GWAS versus PD, Age at Onset, and Risk Haplotype Structure

To estimate risk associated with PD, imputed dosages (meaning genotype probabilities for a variant to be A/A, A/B, or B/B from 0 to 2) and provided by the Michigan Imputation Server (https://imputationserver.sph.umich.edu/index.html#!) were analyzed using a logistic regression model adjusted for sex, AAO for cases or examination for controls, and the first 20 PCs as covariates. AAO was used as the age covariate in the PD GWAS because the covariate "age at recruitment" was not available for some Spanish cohorts. Summary statistics were generated using the RVTESTS package[8] and filtered for inclusion after meeting a minimum imputation quality of 0.30 and MAF >1%. To explore the influence of genetic variation on the AAO of PD cases, a linear regression model, adjusted for the same covariates, was performed (using AAO as the outcome instead of as a covariate).

Compared to the current large-scale GWAS meta-analysis[2] that identified 90 loci for PD risk, this study has comparatively lower power. Many of the new loci recently identified are in the 1.1 to 1.2 odds ratio (OR) range. Carrying out power calculations using the current sample size in our Spanish GWAS with a fixed OR of 1.2 and an alpha of 5e-08 at a prevalence of 0.001, our highest possible power across the allele frequency spectrum would be 35%, and not reaching 80% power until we increase the OR to ~1.3. In addition, this level of power only persists for very common variants at a frequency of >20%.[9] We see a similar lack of power compared to the most powered AAO GWAS to date, which used 28,568 PD cases to only identify two loci associated with AAO.[10]

Given the history of admixture in Spain, we hypothe-sized that there might be smaller haplotype blocks at risk loci than in less admixed populations.[11] We compared the size of the 90 independent risk haplotype blocks in Spanish cases with a British ancestry PD cohort composed of 1,478 cases (32.5% female; 67.5% male; see an earlier work[12] for further details). After standardizing both data sets with the same genotyped SNPs passing identical QC in both data sets, we determined the size of the haplotype blocks in both populations by using PLINK 1.9.[13] Using default parameters, PLINK estimates haplotype blocks by Haploview's interpretation of the block definition suggested by Gabriel and colleagues.[14] By default, only pairs of variants within 200 kilobases (kb) of each other are considered. Two variants are considered by this procedure to be in strong LD if the bottom of the 90% D-prime confidence interval (CI) is >0.70, and the top of the CI is at least 0.98.

In an attempt to prioritize putative causal variants within risk haplotypes, we performed fine-mapping analyses across the LD blocks where the genome-wide significant identified signals are located by using the "Approximate Bayes Factor fine mapping under a single causal variant assumption" method provided by R package coloc (https://CRAN.R-project.org/package=coloc). This analysis assesses the posterior probability of each SNP being the causal variant within a locus. We derived posterior probabilities (PPH) for each region and considered PPH > 0.95 with a default prior probability of 1e-4 as strong evidence for fine-mapping under the assumption of a single causative variant per locus.

Whole-Genome Sequencing Analysis

To dissect the novel GWAS signal associated with AAO identified in the Spanish population, we performed whole-genome sequencing (WGS) analyses in 5 of 37 homozygous carriers of the *PARK2* signal. DNA concentration was determined by Qubit fluorescence and normalized to 20 ng/uL. One microgram of total genomic DNA was sheared to a target size of 450 base pairs (bp) using the Covaris LE220 ultrasonicator. Library preparation was achieved using the TruSeq DNA PCR-Free High Throughput Library Prep Kit and IDT for Illumina TruSeq DNA UD Indexes (96 Indexes, 96 Samples). Sequencing libraries were assessed for size distribution, absence of free adapters, and adapter dimers on a Fragment Analyzer. Library quantitation was performed by quantitative polymerase chain reaction using the KAPA Library Quantification Kit subsequent normalization to 4 nM. Libraries were clustered on v2.5 flowcell using the Illumina cBot 2 System before sequencing on the Illumina HiSeq X System using paired-end 150-bp reads. BCL files processed with alignment by ISAAC on HAS 2.2 and BAMs were used for QC assessment of mean coverage, percent duplicates, percent bases >20× coverage, and percent noise sites.

Genetic Risk Profiling

Polygenic risk score analysis for PD and AAO was performed as described in detail elsewhere.[15] Briefly, a cumulative genetic risk score was calculated by using the R package, PRSice2 (R Foundation for Statistical Computing, Vienna, Austria).[15] Permutation testing and *P* value after LD pruning was used to identify best *P* thresholds in external GWAS data (training data set derived from summary statistics from Nalls and colleagues,[2] excluding Spanish samples) to construct the PRS, allowing us to utilize variants below the common GWAS significance threshold of 5E-08. LD clumping was implemented under default settings (window size = 250 kb; $r^2 > 0.1$), and using the Spanish data set (testing data set), 10,000 permutations were applied to generate empirical *P*-estimates–derived *P* threshold ranging from 5E-08 to 0.5, at a minimum increment of 5E-08. Each permutation test provided a Nagelkerke's pseudo r^2 after adjustment for an estimated prevalence of 0.5%, study-specific eigenvectors 1–20, AAO for cases or examination for controls, and sex as covariates. GWAS-derived *P* threshold with the highest pseudo r^2 was selected for further analysis.

Machine Learning to Predict Disease Status

Summary statistics from the most recent meta-analysis excluding Spanish cohort samples[2] were used for initial SNP selection in our machine learning (ML) analyses. This analysis utilizes an upcoming software package (GenoML; https://genoml.github.io), an automated ML tool that optimizes basic ML pipelines for genomic data. We used PD GWAS full summary statistics in the ML feature selection process from which we removed samples present in the Spanish PD cohort to avoid any circularity. Before analyses, we filtered the Spanish cohort genotype data for MAF > 1% and imputation quality > 0.8. Next, we randomly sampled without replacement 70% of the subjects for training the classifier and the other 30% for its validation. The sample was stratified with the same proportion between cases and controls as in the whole cohort. We opted to use PRSice[15] to prefilter variants under default settings yielding 1,521 candidate variants from GWAS at a *P*-value threshold

of 0.0005. In both the training and test sets, the dosage of each of the 1,521 variants per individuals is weighted by the GWAS beta. We then used a Caret ML (https://github.com/topepo/caret) framework with glmnet, xgbTree, xgbDART, xgbLinear, and Random forest, with a grid search size of 30- and 10-fold cross-validation on the training set. The algorithm maximizing mean area under the curve across iterations of cross-validation in the training set was then fit to the validation set to generate predictions and summary statistics.

Heritability Estimates and Genetic Correlations

The heritability of PD in the Spanish population was calculated using linkage disequilibrium score regression (LDSC).[16] This method has the ability to detect the contributions to disease risk of variants which do not reach genome significance, but does not identify the specific variants contributing to disease risk. Summary statistics used for these analyses were generated based on imputed variants (numbered 9,911,207 after filtering for MAF > 1% and imputation quality [RSQ] >0.3). Using the same software, genetic correlations between PD and other catalogued GWAS studies were evaluated. The database, *LD Hub*, was used to screen overlapping genetic etiologies across 757 diseases/traits gathered from publicly available resources.[16] In order to compare the Spanish PD GWAS to other PD GWAS, we performed genetic correlations versus the latest PD GWAS meta-analysis for which no Spanish samples were included.[17] Default settings were used in the analyses, and final results were adjusted for multiple testing by using Bonferroni correction.

Runs of Homozygosity

Based on an LD-pruned data set (using previously described parameters), runs of homozygosity (ROHs) were defined using PLINK 1.9.[13] We explored ROHs containing at least 10 SNPs and a total length ≥1,000 kb, with a rate of scanning windows of at least 0.05 (not containing >1 heterozygous call or 10 missing calls). In order to explore overall homozygosity between cases and controls, three metrics were assessed, including the number of homozygous segments spread across the genome, total kilobase distance spanned by those segments, and average segment size (autosomes only). Subsequently, ROHs were further investigated for known PD risk gene regions (Supporting Information Table S2) and PD significant loci from GWAS[2] with a window of ±1 Mb upstream or downstream. The test of equal or given proportions was used to test the null that the proportions (probabilities of success) in the group of cases and the group of controls were the same.[18] In these analyses, cryptically related PD individuals removed in previous steps were included to identify over-represented sharing of recessive regions among cases.

Burden Analyses

We examined the contribution of rare variation on disease risk by collapsing the cumulative effect of multiple genetic variants at a per-gene level. To do so, we incorporated as part of this analysis, imputed low frequent variants with an MAF <0.05 calculated in the combined data set according to RVTEST default parameters, with an imputation quality of RSQ >0.8. We performed the Sequence Kernel Association Test in imputed data after classifying variants into nested categories based on two maximum MAF thresholds, (a) <MAF 1% or (b) <MAF 5%, and three functional filters, (1) noncoding variants, (2) only coding, and (3) Combined Annotation Dependent Depletion (CADD) likely damaging.

Burden analyses were adjusted for the first 20 PCs, AAO (cases) or examination (controls), and sex. These were run using default settings as part of the RVTESTS package.[8] To adjust for multiple comparisons, we applied Bonferroni correction. Predictions of variant pathogenicity were obtained from ANNOVAR,[19] based on the CADD algorithm (v1.3; http://cadd.gs.washington.edu).[20] In accord with previous reports,[21] we selected a stringent CADD C-score threshold ≥12.37, representing the top ~2% most damaging of all possible nucleotide changes in the genome.

Quantitative Trait Loci Mendelian Randomization

Two-sample Mendelian randomization (MR) was performed to investigate possible functional genomic associations between PD known genes (Supporting Information Table S2) and nominated loci[2] and expression or methylation quantitative trait loci (QTL) using summary statistics from this GWAS of the Spanish population to represent the outcome. Brain and blood QTL association summary statistics from well-curated methylation and expression data sets available through the SMR website (http://cnsgenomics.com/software/smr)[22] were considered possible exposures in the MR models. These include estimates for methylation and cis-expression across multiple brain regions.[23] We also studied expression patterns in blood from the meta-analyses as described.[24] Multi-SNP summary-data–based MR was utilized to generate association estimates by MR between each QTL and local PD risk SNPs that contained two or more SNPs under default settings.[22] The current Spanish data set was used as a reference for LD in this analysis. For each reference QTL data set, false discovery rate (FDR) correction to *P* values was applied (Supporting Information Table S10).

Known Genetic Factors in PD and Atypical Parkinsonism-Related Genes

Considering both related and unrelated samples, we screened for relevant variants in known PD or parkinsonism-related genes. Although we aimed at including only idiopathic PD cases, we further screened atypical PD syndromes-related genes in order to reassess these cases and provide clinicians with genetic information that could be indicative of misdiagnosis. After annotating the enriched customized content of variants within the PD- and parkinsonism-related genes included on NeuroChip,[7] we selected variants that were tagged as “disease causing mutation,” “possibly disease causing mutation,” “probably disease-associated polymorphism,” “disease-associated polymorphism with additional supporting functional evidence,” “pathogenic,” or “possibly pathogenic” in ClinVar[25] or Human Gene Mutation Database.[26] We followed two models: a putative dominant model (either the allele is present only in cases and absent in controls or with a lower frequency in controls) and a putative recessive model (two copies of the allele present only in cases and absent or with a lower frequency in controls). Genes screened are highlighted in Supporting Information Table S2. We further screened putative structural genomic variation associated with PD in *PARK2* and *SNCA*, given that notable genomic arrangements have been detected in these genes.[27,28] Two metrics were assessed and visualized with R software (version 3.5.1)[29]: B allele frequency and log R ratio. These two statistics allow visualization of copy number changes and are described in detail elsewhere.[30]

Results

Genome-Wide Association for PD risk, Age at Onset, and Population-Specific Risk Haplotype Structure Analyses

We identified four genome-wide independent signals associated with PD risk, including *SNCA*, *LRRK2*, *KANSL1/MAPT*, and *HLA-DQB1* (Fig. 1; Table 1; Supporting Information Fig. S2).

In an attempt to prioritize functional variants, fine-mapping analyses were performed considering the LD blocks where the four GWAS hits are located. We failed at identifying any reliable causal variant within loci at a derived PPH > 0.95 (Supporting Information Table S3).

We further assessed expression and methylation changes within these loci that could be associated with PD through MR analyses. *KANSL1* rs2532233 showed functional consequence with decreased expression in brain, which, in turn, was linked to PD. Additionally, 11 SNPs in *KANSL1* demonstrated changes associated with methylation levels linked to PD. No MR significant associations were found for *SNCA*, *LRRK2*, or *HLA-DQB1*, not surprising given the often underpowered nature of this statistical method.

Additionally, we show an association with PD risk at an uncorrected *P* value <0.05 with 39 of the 90 loci previously identified in the largest PD meta-analysis performed to date[2] (Supporting Information Table S4). Although only 39 loci were associated with PD risk at an uncorrected nominal *P* value in this particular GWAS, we assume this is most likely because of a limited statistical power to detect the small effect sizes of the remaining loci. In this regard, a similar directionality of effect was observed for 82 of the 90 loci originally described in Nalls and colleagues[2] (see Supporting information Table S4 for directionality comparisons).

GWA for AAO of PD revealed a genome-wide significant association signal at *PARK2* (rs9356013, beta = −4.11; standard error [SE] = 0.56, corrected *P* value = 4.44 E-13; Fig. 2 and Supporting Information Figs. 3 and 4; Supporting Information Table S5). The AAO association signal was only observed in the Spanish population at a genome-wide significant level and not in other International Parkinson's Disease Genomics Consortium (IPDGC) data sets. We screened IPDGC GWAS data for which rs9356013 genotypes were available with the limitation that this particular variant was either not genotyped or poorly imputed in some of the IPDGC data sets. Of 19,249 PD cases from the IPDGC, only 4,720 non-Spanish PD cases were accurately genotyped or imputed for rs9356013. Of them, 31 PD cases that were homozygous carriers for the rare allele at rs9356013 had available AAO data. The large amount of missing data is likely a sign of array bias attributable to the difficulty of imputing this variant when using older chip versions, which would have diluted power for this association in other populations. Additionally, the causal *PARK2* variant (c.155delA) that this SNP imperfectly tags is enriched in Spain in comparison with other populations, and therefore we would expect to have sufficient statistical power to detect this association in only the Spanish cohorts. A total of 37 cases carried the SNP in the homozygous state and 440 cases in the heterozygous state, which represents 9.5% of the Spanish PD cases. After

removing rs9356013 homozygous carriers from the linear regression analysis, the signal dropped substantially (beta = −2.64; SE = 0.74; *P* value = 3.41 E-4). The mean AAO for the rs9356013 homozygous carriers was 42.67 ± 14.58 years whereas for the heterozygous carriers it was 60.07 ± 12.64 as compared to the AAO of the overall case series, which was 61.23 ± 11.47. WGS analyses performed in 5 of the 37 homozygous rs9356013 carriers revealed that all of them carried the deleterious frameshift mutation, *PARK2* c.155delA (p.Asn52Metfs), which is located in exon 2 and 39 kb away from the GWAS signal. Among them, 4 individuals carried the variant in the homozygous state and 1 in the heterozygous state (Supporting Information Fig. S5). There were no substantial differences in AAO between the homozygous (average AAO = 47.25 ± 13.9) and the heterozygous (likely compound) carrier whose AAO was 31.

Sanger sequencing of the *PARK2* c.155delA variant (p. Asn52Metfs) in 1,275 PD cases followed by conditional analyses was performed to dissect whether the GWAS signal was tagging the indel. Conditional analyses for rs9356013 conditioning on c.155delA suggested that the common variant, rs9356013, and the rare indel, c.155delA, were most likely dependent signals (rs9356013 linear model *P* value = 1.13E-05; c.155delA linear model *P* value = 3.77E-08; rs9356013 *P* value after conditional analysis on c.155delA = 0.02; beta = −3.38; SE = 1.46). The mean AAO for the 14 c.155delA homozygous carriers detected by Sanger sequencing was 32.42 ± 11.41 years (beta = −16.05; SE = 4.04; *P* value = 7.90e-05) whereas for the 20 heterozygous carriers was 31.78 ± 11.4 (beta = −20.80; SE = 4.03; *P* value = 3.04e-07). Among the 20 heterozygous carriers, 4 were confirmed compound heterozygous for a second pathogenic *PARK2* variant. Given the early-onset nature of these cases and the recessive pattern of inheritance for *PARK2*, we assume that there should be additional compound heterozygotes for a second variant in *PARK2* that we were not able to assess or detect.

In our report, we identify a functional variant (*PARK2* c.155delA) as putatively being associated with AAO. We were able to identify this association because of the increased frequency of this variant in Spanish cases versus non-Spanish cases. As evidenced by power calculations in Blauwendraat and colleagues, AAO is a difficult trait to measure and harmonize across study sites and clinics, thus limiting the power in a hard-to-quantify manner. Noteworthy, we do see a subtop hit toward genome-wide significance level of the previously reported AAO association at the *SNCA* locus (rs356203; *P* value = 2.44E-0.7).

Haplotype size at risk loci in a cohort of British ancestry cases was 13.02 (± 7.72) kb larger than the Spanish at overlapping loci. At consensus genotyped variants, a total of 11 risk haplotype blocks were smaller in the Spanish population than in the British, 23 were the same size, and two were larger (Supporting Information Table S6). Other risk loci did not have multi-SNP genotyped haplotypes spanning top risk variants from external GWAS in both cohorts for comparison. Additionally, in order to explore how sample size might influence the haplotype resolution and thereby the size of the defined haplotype blocks, we randomly sampled 1,478 Spanish cases and reran these analyses on the Spanish and British cohorts at the same N in each subset. Haplotype size at risk loci in the British population were still 8.76 kb (±5.12) larger than the Spanish at overlapping loci.

Genetic Risk Profiling Versus Disease and AAO

After adjusting for appropriate covariates and estimated PD prevalence, an overall pseudo r^2 between PRS and PD was approximately $r^2 = 0.026$. For each standard deviation from the population mean of the PRS, risk was estimated to be an OR of 1.667 (beta = 0.511; SE = 0.027; *P* value = 3.63E-79; empirical *P* after permutation = 1.00E-4). This model incorporated a total of 665 SNPs up to *P* value <5.99E-05 in the current GWAS (Fig. 3A,B).

Similar models were generated for AAO. We observed an association between a 1 standard deviation increase in the age at onset polygenic score and a near 1-year earlier onset of disease (beta = −0.944; SE = 0.346; *P* value = 0.006; empirical *P* after permutation = 0.031). This model utilized all unlinked variants of interest, pruned down to 271,191 SNPs (Fig. 3C,D).

ML to Predict Disease Status

The xgbDART algorithm[31] (extreme gradient boosting approach with additive linear regression trees regularized by dropout) yields the best area under the curve (AUC) as predictor of the probability PD in this Spanish GWAS cohort. The model in the validation set had an AUC of 0.6205 with a sensitivity and specificity of 0.86 and 0.24. This is a ~1% improvement over using simply polygenic risk scores prediction with PRSice alone, even after using a two-stage design compared to a previous single phase that could be more overfit than what is presented here. We also report a balanced accuracy of 0.554%. This measure as well as AUC are more robust indicators of model performance in an unbalanced data set than just simple accuracy.

Heritability Estimates and Genetic Correlations

SNP heritability estimates by LDSC were estimated to be 28.67 ± 6.65%. We analyzed cross-trait genetic correlations between PD and 750 other GWAS data sets of interest curated by *LD hub*.[16] No genetic correlations remained significant after adjusting for multiple testing by FDR. However, when considering an unadjusted *P* value <0.05, negative correlations were found for body mass index–related traits, smoking, and alcohol intake, and positive correlations were identified for allergies and physical activity, among others (Supporting Information Table S7). Furthermore, we analyzed cross-trait genetic correlations between the Spanish PD GWAS and Chang and colleagues.[17] We observed a positive correlation at 85.66% with Chang and colleagues (rg = 0.8566; SE = 0.0979; *P* value = 2.1037e-18).

ROHs

PD cases in our data set were shown to have longer ROHs than controls, both with regard to the percentage of the genome within these runs and the average run size. For every 10-Mb increase in ROHs per sample, we noted an OR of 1.02 (95% CI = 1.036–1.004; *P* value = 0.0097), a small but significant increase. Average run size was also associated with PD risk at an OR of 1.244 per 1-Mb increase in average run size (beta = 0.218; SE = 0.068; *P* value = 0.0013). The total numbers of these ROHs were not significantly different between cases and controls. This suggests that fewer large ROHs might be more closely associated with disease risk than many small ROHs.

We further explored extended runs of homozygosity in known Mendelian PD genes (Supporting Information Table S2) and the 90 nominated risk loci in the last meta-analysis.[2] Homozygosity was found enriched in cases versus controls at 28 genes/loci (Supporting Information Table S8). However, only a ROH in *PARK2* surpassed Bonferroni correction (*P* value threshold = 0.001).

Burden Analyses

By using imputed data, we explored the cumulative effect of multiple rare variants at a gene level by grouping them in different categories based on frequency and functionality (Supporting Information Table S9). We found a significant enrichment of coding variants in the *LRRK2* gene in PD cases compared to controls (*P* value = 4.51E-16). When we excluded p.G2019S carried by 2.8% of the PD subjects from the analysis, the risk of PD conferred by *LRRK2* is not significant (*P* value = 0.0611), suggesting that this variant is the main driver of the association. Rare coding variants in *PARK2* were found overrepresented in cases versus controls (*P* value = 0.008), although this association did not surpass Bonferroni correction (*P*-value threshold = 0.0002). When focusing only on noncoding variants, *GBA* displayed an association at a *P* value = 0.0003, which in this case did not reach multiple testing correction (*P*-value threshold = 3.15E-6). Finally, when grouping the variants by CADD score, we did not find any prospective novel gene associated with PD in the Spanish population.

Quantitative Trait Loci MR

After adjustment for FDR, 17 PD-related genes/loci showed functional consequence by two-sample MR in expression and methylation data sets (Supporting Information Table S10). Increased expression of *NSF* and *BST1* in blood and *KANSL1*, *WNT3*, *KAT8*, *CD38*, *HLA-DRB6*, *TMEM175*, *HLA-DRB6*, and *CTSB* in brain were found to be inversely associated with PD risk, whereas a positive risk association was found for *TMEM163*, *GAK*, and *HLA-DQA1* expression in brain. Disparate results were found across different probes tagging *DGK1*.

Methylation QTL MR analyses revealed 56 CpG sites linked to PD risk in brain after multiple test correction. Increased methylation of *ARHGAP27*, *TMEM175*, *CRHR1*, and *GAK* was found to be positively associated with disease whereas *HLA-DRB5*, *IGSF9B*, *TMEM163*, and *DGKQ* showed a negative directionality versus PD risk. Disparate results were found across different probes tagging *KANSL1* and *HLA-DRB5*.

Known Genetic Factors in PD and Atypical Parkinsonism-Related Genes

A total of 73 PD or parkinsonism variants annotated as possibly disease associated were identified with higher frequencies in the PD patient sample (see Supporting Information Table S11; Supporting Information Fig. S6). Of the identified variants, 28.76% (21 of 73) were detected in genes responsible for autosomal-dominant PD. A total of 19 variants were detected in *LRRK2*; 2.8% (134 of 4,783) of the screened PD patients and 0.3% (12 of 3,066) of the controls carry the *LRRK2* p.G2019S mutation in the heterozygous state, whereas 1 case carried the variant in the homozygous state ($P = 2.73 \times 10^{-15}$; OR = 8.05; SE = 0.26).

The *LRRK2* p.Arg1441Gly variant was identified in a case and the *LRRK2* p. Met1869Thr was found in 5 cases and 1 control.

Examining the correlation between population substructure and p.G2019S genotype in Spain showed consistent associations, but small effect estimates. We extracted 20 eigenvectors from PCA for all 7,849 samples and ran a single linear regression to explore whether these estimates of population substructure predicted p.G2019S dosage. Whereas PCs 2, 6, 7, 9, 10, 11, 13, 14, 15, 18, and 20 were all features with individual parameter estimate $P < 0.05$, the overall adjusted r-squared of the model was only 2.5%, suggesting a minor impact on allele frequency within Spain.

Of the variants, 36.9 % (27 of 73) were identified within autosomal-recessive PD genes, including 19 variants in *PARK2* and eight variants in *PINK1*. Overall, 15.78% of the *PARK2* cases carriers (18 of 114) and 5.5% of the *PINK1* carriers were found in the homozygous state. Although NeuroChip was not able to detect any *PARK2* or *PINK1* compound heterozygous carriers, exonic rearrangements were detected in 0.96% of the screened patients (Supporting Information Fig. S7A, B). *PARK2* deletions were identified among 34 patients; 26 in the heterozygous state and 8 in the homozygous. Duplications were identified in 12 patients.

A total of 17.8% (13 of 73) of the variants were found in PD risk genes. Eleven *GBA* variants were detected among (175 of 4,783) patients. The *GBA* p.His490Arg, p.Val437Ile, p.Gly234Glu, p.Val54Leu, p.Lys13Arg, p.Leu29Alafs*18, and p.Leu363Pro variants were found over-represented in cases versus controls, but the association analysis did not reach statistical significance (Supporting Information Table S11). A 1.1% (53 of 4,783) of the patients and 0.3% (12 of 3,066) of the controls under study carry the *GBA* p.Asn409Ser mutation in the heterozygous state. A total of 2% (97 of 4,783) of the cases and 1.01% (31 of 3,066) of the controls harbored the *GBA* p.Glu365Lys heterozygous polymorphism ($P = 0.0006$; OR = 2.005; SE = 0.2). The *GBA* p.Asp448His variant was identified in 14 cases and 3 controls, respectively ($P = 0.07$; OR = 2.99; SE = 0.6). Furthermore, the previously reported p.Asn613del in *MAPT* was identified in an 87-year-old patient with an onset of rest tremor at 62 years and a notable family history of PD and PSP.[32] Finally, 14 variants were found in four atypical parkinsonism-related genes, including *FBXO7* and *POLG1* (autosomal recessive), as well as *ATP13A2* and *DCTN1* (autosomal dominant); however, their disease significance is uncertain.

Discussion

As part of a Spanish multicenter massive collaborative effort, we have gathered the largest collection of PD patients and controls from a single country to comprehensively assess the genetics of PD on a genome-wide scale. We have used the same genotyping platform, thus reducing possible batch effects. Here, we dissect population-specific differences in risk and AAO from a genetic perspective and highlight the utility of the Spanish risk haplotype substructure for future fine-mapping efforts.

In concordance with other populations,[2] our Spanish GWAS on PD risk replicated four loci linked to disease, strengthening once again the role of *SNCA*, *LRRK2*, *KANSL1/MAPT*, and *HLA-DQB1* in disease etiology. The sample size under study limited us to replicate only 4 of the 90 PD risk loci reported to date.[2] Our cohort was underpowered, as demonstrated by power calculations. Although we were not able to detect any of the previously loci for PD AAO[10] at a GWAS level, we do see a trend toward genome-wide significance level of the previously reported AAO association at the *SNCA* locus (rs356203, *P* value = 2.44E-0.7).

Of note, we identified, for the first time, in a European population an intronic signal in *PARK2* as a modifier of AAO. Conditional analysis showed a likely dependent effect with c.155delA, highlighting a higher frequency of this deleterious mutation in Spain compared to other populations. Given the shared ancestry, c.155delA has been described at high frequencies in the Iberian Peninsula[33,34] and has also been reported to be more common in the Latino population (GnomAD allele count = 39 of 35,440; frequency = 0.11%) versus non-Finnish Europeans (GnomAD allele count =31 of 129,038; frequency = 0.024%).

This signal did not show up as genome-wide significant in other IPDGC European populations; we believe this is a population-specific association attributable to an enrichment of *PARK2* c.155delA cases.

Genetic testing can help to design an optimized trial with the highest likelihood of providing meaningful and actionable answers. Our study shows that Spain is a valuable resource for identifying and tracking *PARK2* c.155delA carriers to accelerate enrollment for target-specific PD clinical trials.

The fact that risk PD haplotypes are smaller in the Spanish population, in comparison to the less admixed British population, brings to light the importance of exhaustively studying diverse populations. The investigation of admixed populations in GWA studies has significant potential to accelerate the mapping of PD loci. As we have shown through fine-mapping efforts in this population, we assume the limitation that GWAS might not be the best approach to nominate and prioritize causal functional variants. We believe that target resequencing and WGS approaches might be the best way to further delineate risk loci.

Importantly, we revealed an overall excess of homozygosity in PD cases versus controls and identified 28 genes/loci exhibiting ROH overrepresented in cases, pointing out the possible existence of disease-causing recessive variants that might be uncovered by future sequencing analysis. Additionally, burden analysis reinforced the contribution of both common and rare variants in *LRRK2*, making Spain an important candidate population for specific *LRRK2* clinical trials. Not surprisingly, we found that p.G2019S is a common *LRRK2* mutation among PD patients from Spain with a frequency of 2.8% (135 genotyped carriers +2 imputed carriers/4,783). Although p.G2019S is responsible for 0.5% to 4% of idiopathic PD cases among Europeans, its prevalence has been found much higher among Ashkenazi Jews and North African Arabs.[35,36] However, the frequency of p.G2019S-carriers in PD cases from Spain suggests that it is likely not enriched among PD cases in concordance to what

has been previously reported,[37,38] even though there are close connections between Spain and North Africa historically.

Similar to *LRRK2*, the frequency rate of *GBA* mutations varies considerably depending on the population ethnicity, with a remarkably high frequency in individuals of Ashkenazi Jewish descent. Our results show a frequency of *GBA* variants of 3.6% (175 of 4,783) among PD patients, significant lower in comparison to another study carried out in the Spanish population which found *GBA* mutations with a frequency of 9.8%.[39] This discrepancy could be attributed, in part, to the limited sensitivity of NeuroChip to screen for *GBA* variants as compared with sequencing technologies. Previous studies in other European populations reported higher frequencies too; 6.4% in Greeks,[40] 4.2% in British,[41] and 8.3% in the Portuguese population.[42]

In an effort to explore functional consequences associated to PD risk in the Spanish population, we performed quantitative trait loci MR analyses using expression and methylation data and suggest that biological pathways underlying the nominated genes warrant further study.

Recent research has begun to demonstrate the utility of polygenic risk profiling to identify individuals who could benefit from the knowledge of their probabilistic susceptibility to disease, an aspect that is central to clinical decision making and early disease detection.[43] Here, we assessed the overall cumulative contribution of common SNPs on disease risk and age at onset. Our PRS-derived model for disease risk and age at onset showed expected trends comparable to previous literature.[10,43]

Although we have made progress in assessing genetic risk factors for PD in a population-specific manner, there are a number of limitations to our study. First, although all the available PD cases and controls from Spain have been assessed, we are aware of the caveats driven by sample size. Dissection of additional susceptibility genetic risk and phenotypic relationships would have been possible if a larger cohort had been analyzed. In fact, the heritability estimate, of ~28.67 % in this population, indicates that there is a large component of genetic risk yet to be uncovered. We assume that there are a considerable number of variants that impact risk for disease outside the limits of what can be accurately detected with a genotyping platform. This could explain the lower observed frequency of certain well-established pathogenic variants and exonic rearrangements when comparing other sequencing studies previously performed in the Spanish population.[44,45]

We have applied a state-of-the-art ML approach in an effort to predict disease status. Our results show that genetic data are not sufficient to accurately predict disease status in a clinical setting by itself when used alone, although this may change in the future when combining genetic with other biomarker data. Entering the era of personalized medicine in which an individual’s genetic makeup will help determine the most suitable therapy, we envisage our collaborative initiative will expand toward identifying, refining, and predicting heritable risk in the Spanish population by combining future large-scale WGS approaches, multiomics, and detailed longitudinal clinical data for translational approaches. We conclude

by saying that this is the starting point of a collaborative network of Spanish clinicians and scientists that will continue to pave the road toward future therapeutic interventions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Sara Bandres-Ciga, PhD[1,2,*], Sarah Ahmed, BSc[1,3], Marya S. Sabir, BSc[1,3], Cornelis Blauwendraat, PhD[1], Astrid D. Adarmes-Gómez, MD[4,5], Inmaculada Bernal-Bernal, MSc[4,5], Marta Bonilla-Toribio, MSc[4,5], Dolores Buiza-Rueda, MSc[4,5], Fátima Carrillo, MD, PhD[4,5], Mario Carrión-Claro, MSc[4,5], Pilar Gómez-Garre, PhD[4,5], Silvia Jesús, MD, PhD[4,5], Miguel A. Labrador-Espinosa, MSc[4,5], Daniel Macias, MD[4,5], Carlota Méndez-del-Barrio, MD[4,5], Teresa Periñán-Tocino, MSc[4,5], Cristina Tejera-Parrado, MSc[4,5], Laura Vargas-González, RN[4,5], Monica Diez-Fairen, MSc[6], Ignacio Alvarez, MSc[6], Juan Pablo Tartari, MD[6], Mariateresa Buongiorno, MD[6], Miquel Aguilar, MD[6], Ana Gorostidi, PhD[7,8,9], Jesús Alberto Bergareche, MD, PhD[7,8,10], Elisabet Mondragon, MD[7,8,10], Ana Vinagre-Aragon, MD[10], Ioana Croitoru[7], Javier Ruiz-Martínez, MD, PhD[7,8,10], Oriol Dols-Icardo, PhD[5,11], Jaime Kulisevsky, MD, PhD[5,12], Juan Marín-Lahoz, MD[5,12], Javier Pagonabarraga, MD, PhD[5,12], Berta Pascual-Sedano, MD[5,12], Mario Ezquerra, PhD[5,13,14], Ana Cámara, BSc[5,13,14], Yaroslau Compta, MD, PhD[5,13,14], Manel Fernández, BSc[5,13,14], Rubén Fernández-Santiago, PhD[5,13,14], Esteban Muñoz, MD, PhD[5,13,14], Eduard Tolosa, MD, PhD[5,13,14], Francesc Valldeoriola, MD, PhD[5,13,14], Isabel Gonzalez-Aramburu, MD, PhD[5,15], Antonio Sanchez Rodriguez, MD[5,15], María Sierra, MD, PhD[5,15], Manuel Menéndez-González, MD, PhD[16,17], Marta Blazquez, MD, PhD[16,17], Ciara Garcia, MD[16,17], Esther Suarez-San Martin, MD[16,17], Pedro García-Ruiz, MD, PhD[18], Juan Carlos Martínez-Castrillo, PhD[19], Lydia Vela-Desojo, PhD[20], Clara Ruz, BSc[2,21], Francisco Javier Barrero, MD, PhD[2,22], Francisco Escamilla-Sevilla, MD, PhD[2,23], Adolfo Mínguez-Castellanos, MD, PhD[2,23], Debora Cerdan, MD, PhD[24], Cesar Tabernero, MD[24], Maria Jose Gomez Heredia, MD[25], Francisco Perez Errazquin, MD[25], Manolo Romero-Acebal, MD[25], Cici Feliz, MD[18], Jose Luis Lopez-Sendon, MD[19], Marina Mata, MD[26], Irene Martínez Torres, PhD[27], Jonggeol Jeffrey Kim, BSc[1], Clifton L. Dalgard, PhD, CLS[28,29], The American Genome Center, Janet Brooks, BSc[1], Sara Saez-Atienzar, PhD[30], J. Raphael Gibbs, PhD[31], Rafael Jorda, BSc[32], Juan A. Botia, PhD[32,33], Luis Bonet-Ponce, PhD[1], Karen E. Morrison, BMBCh, DPhil[34], Carl Clarke, MD[35,36], Manuela Tan, BSc[37], Huw Morris, MD, PhD[37], Connor Edsall, BSc[1], Dena Hernandez, PhD[1], Javier Simon-Sanchez, PhD[38], Mike A. Nalls, PhD[1,39], Sonja W. Scholz, MD, PhD[3,40], Adriano Jimenez-Escrig, MD[19], Jacinto Duarte, MD, PhD[24], Francisco Vives, MD, PhD[2,21], Raquel Duran, PhD[2,21], Janet Hoenicka, PhD[41,42], Victoria Alvarez, PhD[17,43], Jon Infante, MD, PhD[5,15], Maria José Marti, MD, PhD[5,13,14], Jordi Clarimón, PhD[5,11], Adolfo López de Munain, MD, PhD[7,8,44], Pau Pastor, MD, PhD[6], Pablo Mir, MD, PhD[4,5], Andrew Singleton, PhD[1,*]

on behalf of the International Parkinson Disease Genomics Consortium

Affiliations

[1]Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, Maryland, USA

[2]Instituto de Investigación Biosanitaria de Granada (ibs.GRANADA), Granada, Spain

[3]Neurodegenerative Diseases Research Unit, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland, USA

[4]Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Unidad de Trastornos del Movimiento, Servicio de Neurología y Neurofisiología Clínica, Instituto de Biomedicina de Sevilla, Seville, Spain

[5]Centro de Investigación Biomédica en Red sobre Enfermedades Neurodegenerativas (CIBERNED), Spain

[6]Fundació Docència i Recerca Mútua de Terrassa and Movement Disorders Unit, Department of Neurology, University Hospital Mútua de Terrassa, Terrassa, Barcelona, Spain

[7]Neurodegenerative Disorders Area, Biodonostia Health Research Institute, San Sebastián, Spain

[8]Centro de Investigación Biomédica en Red sobre Enfermedades Neurodegenerativas (CIBERNED), Madrid, Spain

[9]Plataforma de Genomica, Instituto de Investigacion Biodonostia, San Sebastián, Spain

[10]Unidad de Trastornos de Movimiento, Departamento de Neurologia, Hospital Universitario de Donostia, San Sebastián, Spain

[11]Genetics of Neurodegenerative Disorders Unit, IIB Sant Pau, and Universitat Autònoma de Barcelona, Barcelona, Catalonia, Spain

[12]Movement Disorders Unit, Neurology Department, Sant Pau Hospital, Universitat Autònoma de Barcelona, Barcelona, Catalonia, Spain

[13]Lab. of Parkinson disease and Other Neurodegenerative Movement Disorders, IDIBAPS-Institut d'Investigacions Biomèdiques, Barcelona, Catalonia, Spain

[14]Unitat de Parkinson i Trastorns del Moviment. Servicio de Neurologia, Hospital Clínic de Barcelona and Institut de Neurociencies de la Universitat de Barcelona (Maria de Maetzu Center), Catalonia, Spain

[15]Servicio de Neurología, Hospital Universitario Marqués de Valdecilla (IDIVAL) and Universidad de Cantabria, Santander, Spain

[16]Servicio de Neurología, Hospital Universitario Central de Asturias, Asturias, Spain

[17]Instituto de Investigación Sanitaria del Principado de Asturias (ISPA), Asturias, Spain

[18]Departamento de Neurologia, Instituto de Investigación Sanitaria Fundación Jiménez Díaz, Madrid, Spain

[19]Departamento de Neurologia, Instituto Ramón y Cajal de Investigación Sanitaria, Hospital Universitario Ramón y Cajal, Madrid, Spain

[20]Servicio de Neurologia, Hospital Universitario Fundación Alcorcón, Madrid, Spain

[21]Centro de Investigacion Biomedica and Departamento de Fisiologia, Facultad de Medicina, Universidad de Granada, Granada, Spain

[22]Servicio de Neurología, Hospital Universitario San Cecilio, Granada, Universidad de Granada, Spain

[23]Servicio de Neurología, Hospital Universitario Virgen de las Nieves, Granada, Spain

[24]Servicio de Neurología, Hospital General de Segovia, Segovia, Spain

[25]Servicio de Neurología, Hospital Universitario Virgen de la Victoria, Malaga, Spain

[26]Departamento de Neurologia, Hospital Universitario Infanta Sofía, Madrid, Spain

[27]Departamento de Neurologia, Instituto de Investigación Sanitaria La Fe, Hospital Universitario y Politécnico La Fe, Valencia, Spain

[28]Department of Anatomy, Physiology & Genetics, Uniformed Services University of the Health Sciences, Bethesda, Maryland, USA

[29]The American Genome Center, Collaborative Health Initiative Research Program, Uniformed Services University of the Health Sciences, Bethesda, Maryland, USA

[30]Neuromuscular Diseases Research Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, Maryland, USA

[31]Computational Biology Group, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, Maryland, USA

[32]Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, Murcia, Spain

[33]Department of Molecular Neuroscience, UCL, Institute of Neurology, London, United Kingdom

[34]Department of Neurology, Faculty of Medicine, University of Southampton, Southampton, United Kingdom

[35]University of Birmingham, Birmingham, United Kingdom

[36]Sandwell and West Birmingham Hospitals NHS Trust, Birmingham, United Kingdom

[37]Department of Clinical Neuroscience, University College London, London, United Kingdom

[38]Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tübingen, and DZNE, German Center for Neurodegenerative Diseases, Tübingen, Germany

[39]Data Tecnica International, Glen Echo, Maryland, USA

[40]Department of Neurology, Johns Hopkins Medical Center, Baltimore, Maryland, USA

[41]Laboratorio de Neurogenética y Medicina Molecular, Institut de Recerca Sant Joan de Déu, Barcelona, Spain

[42]Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Madrid, Spain

[43]Laboratorio de Genética, Hospital Universitario Central de Asturias, Asturias, Spain

[44]Departamento de Neurociencias. UPV-EHU, Servicio de Neurología, Hospital Universitario Donostia, San Sebastián, Spain

Acknowledgments:

We are grateful to the participants in this study without whom this work would not have been possible. We thank HUMV-IDIVAL biobank for providing the biological samples and associated data included in this study.

Funding agencies:

This research was supported, in part, by the Intra-mural Research Program of the National Institutes of Health (National Institute on Aging, National Institute of Neurological Disorders and Stroke; project numbers: 1ZIA-NS003154-03, Z01-AG000949-02, and Z01-ES101986). In addition, this work was supported by the Department of Defense (award W81XWH-09-2-0128), The Michael J Fox Foundation for Parkinson's Research, and the ISCIII Grants PI 15/0878 (Fondos Feder) to V.A. and PI 15/01013 to J,H. This study was supported by grants from the Spanish Ministry of Economy and Competitiveness (PI14/01823, PI16/01575, PI18/01898, [SAF2006-10126 (2006-2009), SAF2010-22329-C02-01 (2010-2012), and SAF2013-47939-R (2013-2018)]), co-founded by ISCIII (Subdirección General de Evaluación y Fomento de la Investigación) and by Fondo Europeo de Desarrollo Regional (FEDER), the Consejería de Economía, Innovación, Ciencia y Empleo de la Junta de Andalucía (CVI-02526, CTS-7685), the Consejería de Salud y Bienestar Social de la Junta de Andalucía (PI-0437-2012, PI-0471-2013), the Sociedad Andaluza de Neurología, the Jacques and Gloria Gossweiler Foundation, the Fundación Alicia Koplowitz, and the Fundación Mutua Madrileña. Pilar Gómez-Garre was supported by the "Miguel Servet" (from ISCIII16 FEDER) and "Nicolás Monardes" (from Andalusian Ministry of Health) programmes. Silvia Jesús Maestre was supported by the "Juan Rodés" programme, and Daniel Macías-García was supported by the "Río Hortega" programme (both from ISCIII-FEDER). Cristina Tejera Parrado was supported by VPPI-US from the Universidad de Sevilla. This research has been conducted using samples from the HUVR-IBiS Biobank (Andalusian Public Health System Biobank and ISCIII-Red de Biobancos PT13/0010/0056). This work was also supported by the grant PSI2014-57643 from the Junta de Andalucía to the CTS-438 group and a research award from the Andalusian Society of Neurology.

Relevant conflicts of interest/financial disclosures:

Mike A. Nalls' participation is supported by a consulting contract between Data Tecnica International and the National Institute on Aging, NIH (Bethesda, MD) as a possible conflict of interest. Dr. Nalls also consults for Neuron23 Inc., Lysosomal Therapeutics Inc., Illumina Inc., the Michael J. Fox Foundation, and Vivid Genomics, among others.

References

1. Billingsley KJ, Bandres-Ciga S, Saez-Atienzar S, Singleton AB. Genetic risk factors in Parkinson's disease. Cell Tissue Res2018; 373:9–20. [PubMed: 29536161]

2. Nalls MA, Blauwendraat C, Vallerga CL, et al.Expanding Parkinson's disease genetics: novel risk loci, genomic context, causal insights and heritable risk. bioRxiv201889. doi: 10.1101/388165. [Epub ahead of print]

3. Botigué LR, Henn BM, Gravel S, et al.Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. Proc Natl Acad Sci U S A2013;110:11791–11796. [PubMed: 23733930]

4. Gayán J, Galan JJ, González-Pérez A, et al.Genetic structure of the Spanish population. BMC Genomics2010;11:326. [PubMed: 20500880]

5. Bycroft C, Fernández-Rozadilla C, Ruiz-Ponte C, et al.Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. bioRxiv2018;arXiv:250191.

6. Gelb DJ, Oliver E, Gilman S. Diagnostic criteria for Parkinson disease. Arch Neurol1999;56:33–39. [PubMed: 9923759]

7. Blauwendraat C, Faghri F, Pihlstrom L, et al.NeuroChip, an updated version of the NeuroX genotyping platform to rapidly screen for variants associated with neurological diseases. Neurobiol Aging2017;57:247.e9–e247.e13.

8. Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics2016;32:1423–1426. [PubMed: 27153000]

9. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet2006;38:209–213. [PubMed: 16415888]

10. Blauwendraat C, Heilbron K, Vallerga CL, et al.Parkinson disease age of onset GWAS: defining heritability, genetic loci and a-synuclein mechanisms. 20181011. doi: 10.1101/424010. [Epub ahead of print]

11. Guan YDetecting structure of haplotypes and local ancestry. Genetics2014;196:625–642. [PubMed: 24388880]

12. PD Med Collaborative Group, Gray R, Ives N, et al.Long-term effectiveness of dopamine agonists and monoamine oxidase B inhibitors compared with levodopa as initial treatment for Parkinson's disease (PD MED): a large, open-label, pragmatic randomised trial. Lancet2014;384:1196–1205. [PubMed: 24928805]

13. Purcell S, Neale B, Todd-Brown K, et al.PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet2007;81:559–575. [PubMed: 17701901]

14. Gabriel SB, Schaffner SF, Nguyen H, et al.The structure of haplotype blocks in the human genome. Science2002;296: 2225–2229. [PubMed: 12029063]

15. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. Bioinformatics2014;31:1466–1468. [PubMed: 25550326]

16. Zheng J, Erzurumluoglu AM, Elsworth BL, et al.LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics2017;33:272–279. [PubMed: 27663502]

17. Chang D, Nalls MA, Hallgrímsdóttir IB, et al.A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. Nat Genet2017;49:1511–1516. [PubMed: 28892059]

18. Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc1927;22:209–212.

19. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res2010;38:e164. [PubMed: 20601685]

20. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet2014;46: 310–315. [PubMed: 24487276]

21. Robak LA, Jansen IE, van Rooij J, et al.Excessive burden of lysosomal storage disorder gene variants in Parkinson's disease. Brain2017;140:3191–3203. [PubMed: 29140481]

22. Zhu Z, Zhang F, Hu H, et al.Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet2016;48:481–487. [PubMed: 27019110]

23. Qi T, Wu Y, Zeng J, et al.Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. Nat Commun2018;9:2282. [PubMed: 29891976]
24. Westra H-J, Peters MJ, Esko T, et al.Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet2013;45:1238–1243. [PubMed: 24013639]
25. Harrison SM, Riggs ER, Maglott DR, et al.Using ClinVar as a Resource to Support Variant Interpretation. Curr Protoc Hum Genet2016;89:8.16.1–8.16.23.
26. Stenson PD, Ball E, Howells K, Phillips A, Mort M, Cooper DN. Human Gene Mutation Database: towards a comprehensive central mutation database. J Med Genet2007;45:124–126.
27. Singleton AB, Farrer M, Johnson J, et al.alpha-Synuclein locus triplication causes Parkinson's disease. Science2003;302:841. [PubMed: 14593171]
28. Leroy E, Anastasopoulos D, Konitsiotis S, Lavedan C, Polymeropoulos MH. Deletions in the Parkin gene and genetic heterogeneity in a Greek family with early onset Parkinson's disease. Hum Genet1998;103:424–427. [PubMed: 9856485]
29. Horton N, Kleinman K. Using R and RStudio for Data Management, Statistical Analysis, and Graphics, 2nd ed. Boca Raton, FL: CRC; 2015.
30. Matarin M, Simon-Sanchez J, Fung HC, et al.Structural genomic variation in ischemic stroke. Neurogenetics2008;9:101–108. [PubMed: 18288507]
31. Rashmi K, Gilad-Bachrach R. Dart: dropouts meet multiple additive regression trees. arXiv preprint2015;arXiv:150501866.
32. Pastor P, Pastor E, Carnero C, et al.Familial atypical progressive supranuclear palsy associated with homozigosity for the delN296 mutation in the tau gene. Ann Neurol2001;49:263–267. [PubMed: 11220749]
33. Muñoz E, Tolosa E, Pastor P, et al.Relative high frequency of the c.255delA parkin gene mutation in Spanish patients with autosomal recessive parkinsonism. J Neurol Neurosurg Psychiatry2002;73: 582–584. [PubMed: 12397156]
34. Morais S, Bastos-Ferreira R, Sequeiros J, Alonso I. Genomic mechanisms underlying PARK2 large deletions identified in a cohort of patients with PD. Neurol Genet2016;2:e73. [PubMed: 27182553]
35. Healy DG, Falchi M, O'Sullivan SS, et al.Phenotype, genotype, and worldwide genetic penetrance of LRRK2-associated Parkinson's disease: a case-control study. Lancet Neurol2008;7:583–590. [PubMed: 18539534]
36. Lesage S, Dürr A, Tazir M, et al.LRRK2 G2019S as a cause of Parkinson's disease in North African Arabs. N Engl J Med2006; 354:422–423. [PubMed: 16436781]
37. Gao L, Gómez-Garre P, Díaz-Corrales FJ, et al.Prevalence and clinical features of LRRK2 mutations in patients with Parkinson's disease in southern Spain. Eur J Neurol2009;16:957–960. [PubMed: 19473361]
38. Bandrés-Ciga S, Mencacci NE, Durán R, et al.Analysis of the genetic variability in Parkinson's disease from Southern Spain. Neurobiol Aging2016;37:210.e1–e210.e5. [PubMed: 26518746]
39. Setó-Salvia N, Clarimón J, Pagonabarraga J, et al.Dementia risk in Parkinson disease. Arch Neurol2011;68:359–364. [PubMed: 21403021]
40. Kalinderi K, Bostantjopoulou S, Paisan-Ruiz C, Katsarou Z, Hardy J, Fidani L. Complete screening for glucocerebrosidase mutations in Parkinson disease patients from Greece. Neurosci Lett2009;452:87–89. [PubMed: 19383421]
41. Neumann J, Bras J, Deas E, et al.Glucocerebrosidase mutations in clinical and pathologically proven Parkinson's disease. Brain2009; 132(Pt 7):1783–1794. [PubMed: 19286695]
42. Bras J, Paisan-Ruiz C, Guerreiro R, et al.Complete screening for glucocerebrosidase mutations in Parkinson disease patients from Portugal. Neurobiol Aging2009;30:1515–1517. [PubMed: 18160183]
43. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet2018;19: 581–590. [PubMed: 29789686]
44. Diez-Fairen M, Benitez BA, Ortega-Cubero S, et al.Pooled-DNA target sequencing of Parkinson genes reveals novel phenotypic associations in Spanish population. Neurobiol Aging2018;70:325.e1–e325.e5.

45. Gómez-Garre P, Jesús S, Carrillo F, et al.Systematic mutational analysis of FBXO7 in a Parkinson's disease population from southern Spain. Neurobiol Aging2014;35:727.e5–e727.e7.

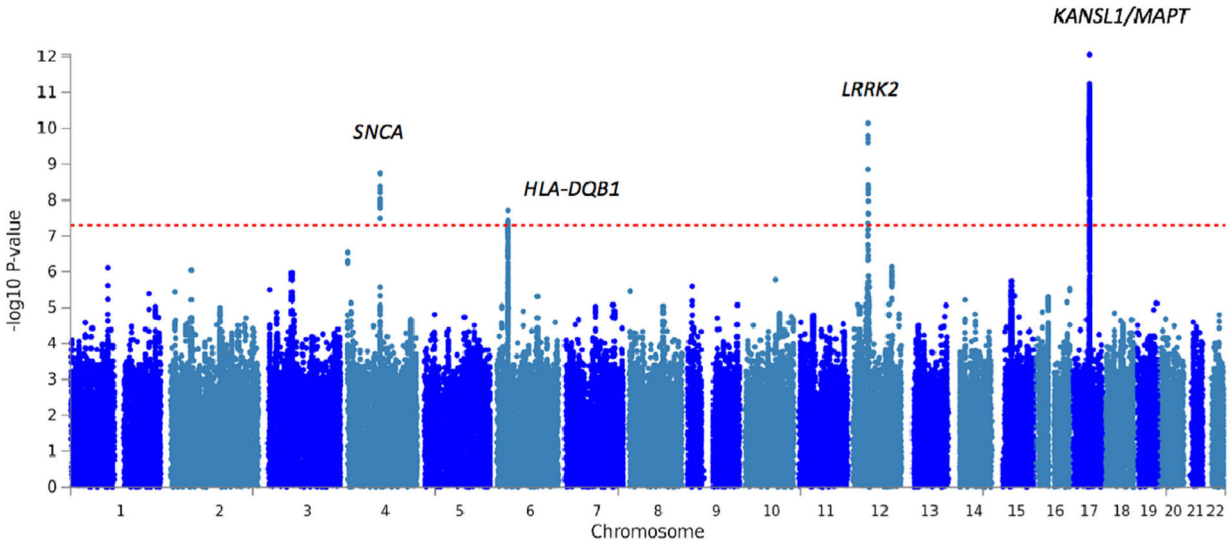

FIG. 1.

Manhattan plot showing results of PD GWA testing. Based on unrelated individuals (4,639 cases and 2,949 controls) using 9,945,565 SNPs. Four genome-wide significant loci were identified: *SNCA*, *LRRK2*, *HLA-DQB1*, and *MAPT*.

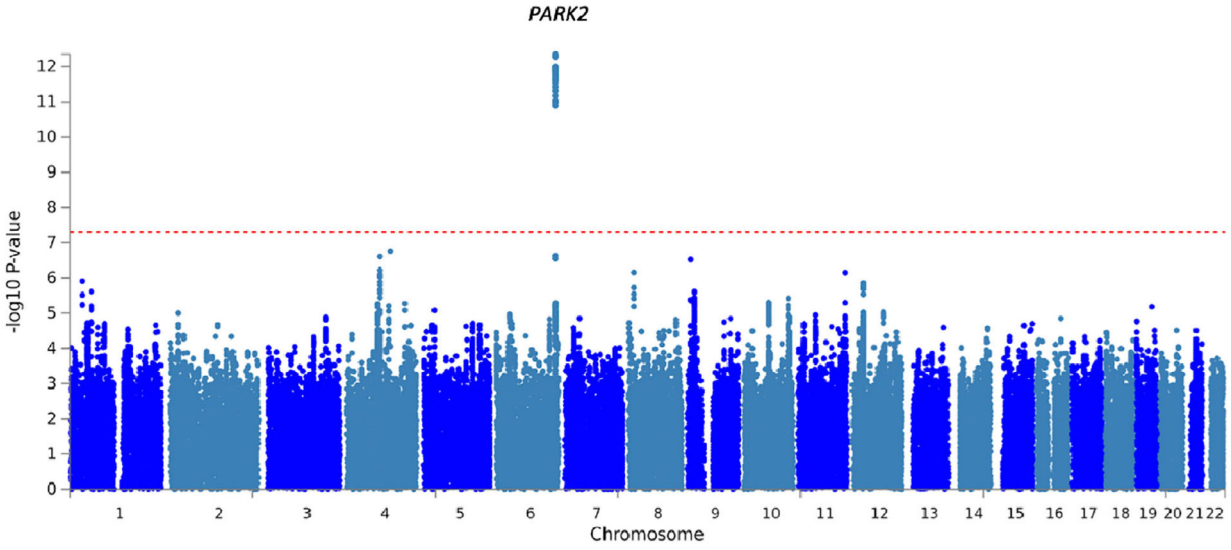

FIG. 2.

Manhattan plot showing results of PD GWA with AAO testing. Based on 3,997 unrelated cases with available age at onset information using 9,945,565 SNPs. One genome-wide significant loci was identified: *PARK2*.

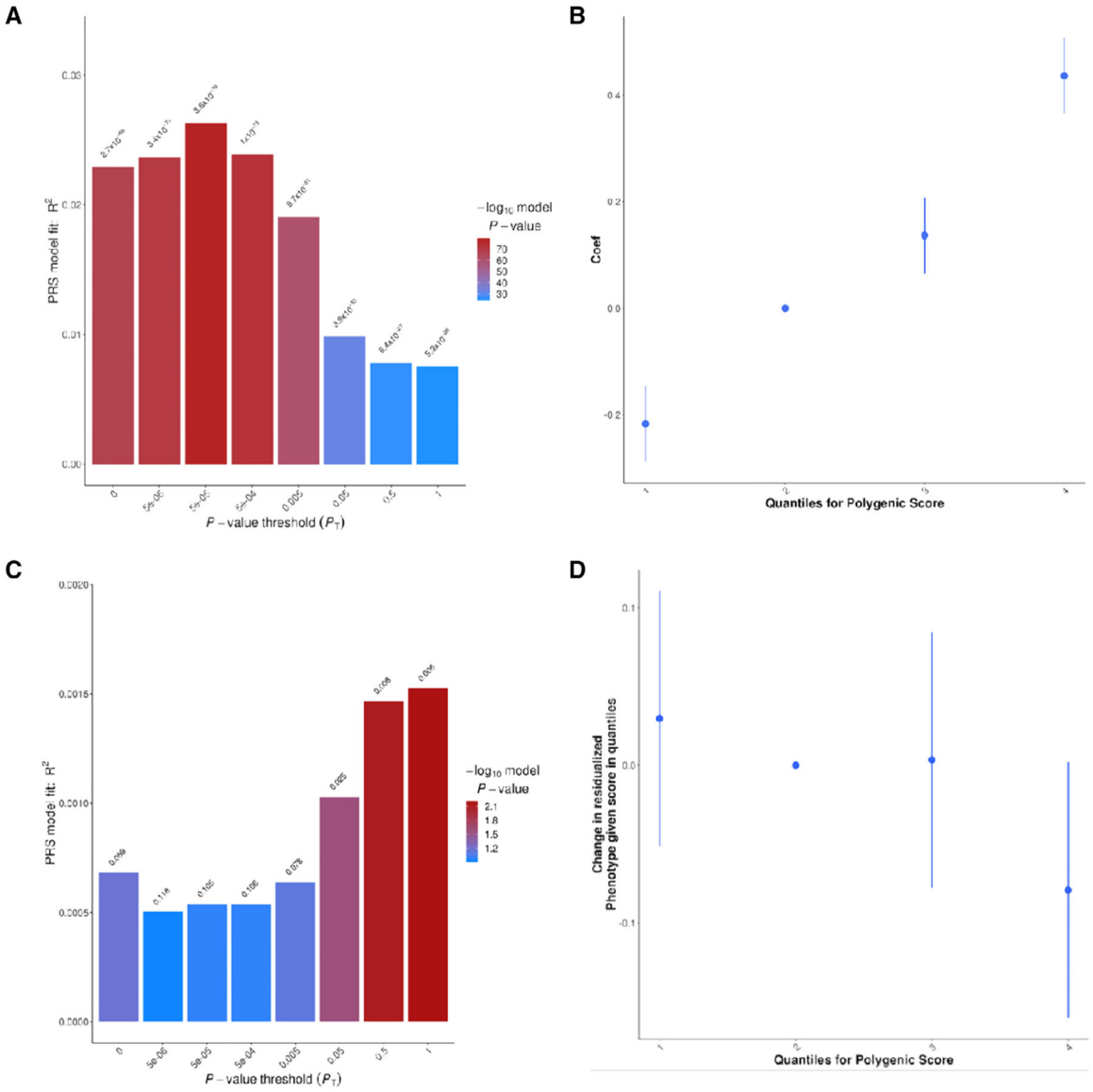

FIG. 3.

Polygenic risk score versus disease status and AAO. (A) Polygenic risk score versus disease status. R^2 estimates at various *P*-value thresholds. (B) ORs by quantile of PD polygenic risk score. (C) Polygenic risk score versus AAo. R^2 estimates at various *P*-value thresholds. (D) ORs by quantile of AAO polygenic risk score.

TABLE 1.

Genome-wide significant loci associated with PD risk in the Spanish population

SNP	Nearest Gene	Chr	Position	A1	A2	Beta	SE	MAF	*P* Value
rs356182	*SNCA*	4	90626111	G	A	0.223036	0.0370765	0.34261	1.79E-09
rs190807041	*LRRK2*	12	40773684	G	A	2.03403	0.312096	0.01044	7.16E-11
rs113434679	*KANSL1/MAPT*	17	44126765	A	C	−0.311323	0.0435317	0.22119	8.57E-13
rs9275152	*HLA-DQB1*	6	32652196	C	T	−0.376551	0.0670165	0.08263	1.92E-08

A1, minor allele; A2, major allele; Chr, chromosome.