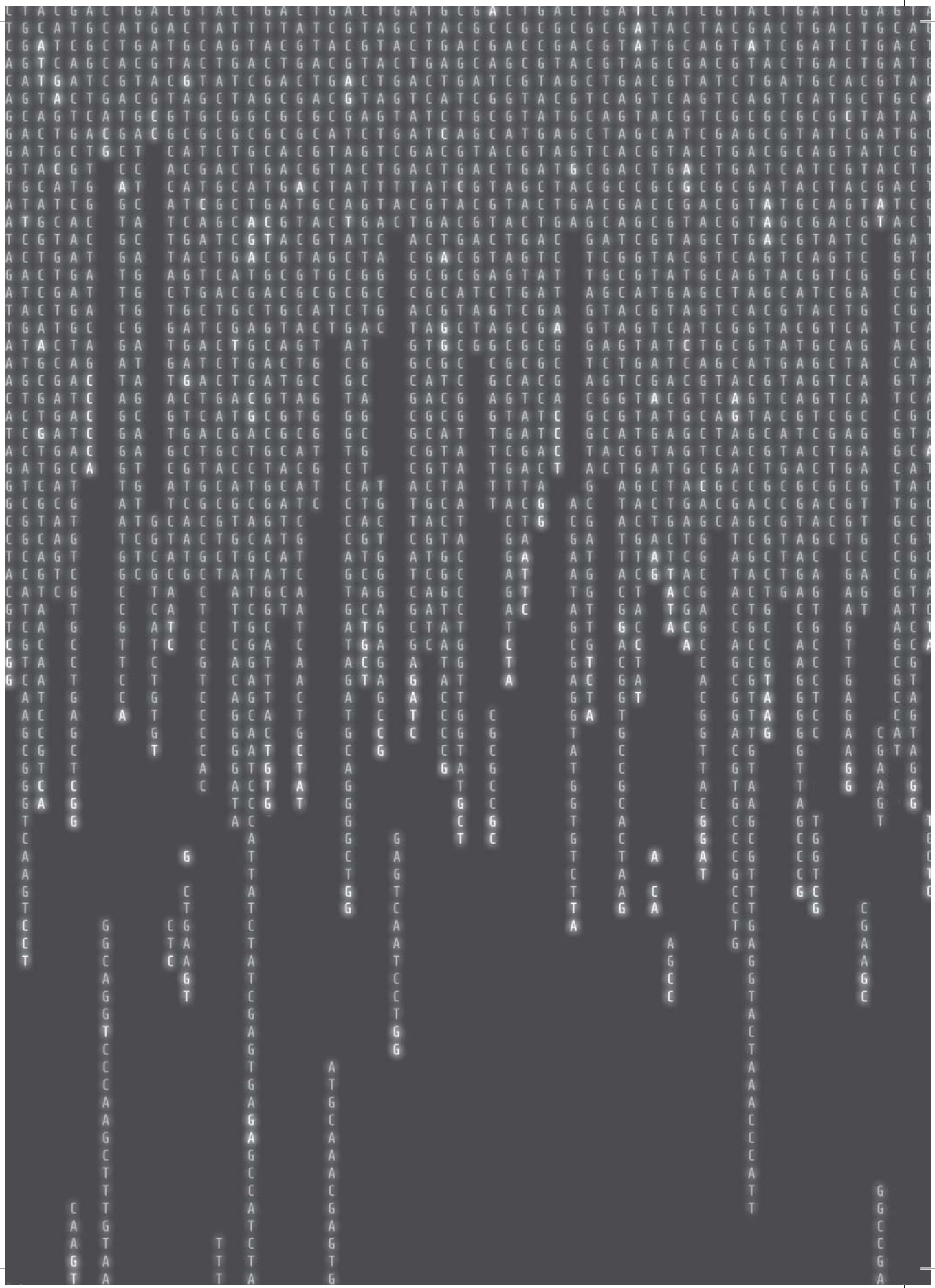
Cracking the code-ing sequence for Parkinson’s disease

Jansen, I.E.

2017

***document version***
Publisher's PDF, also known as Version of record



# CHAPTER 3

# ESTABLISHING THE ROLE OF RARE CODING VARIANTS IN KNOWN PARKINSON'S DISEASE RISK LOCI

This chapter is in preparation for publication as: Jansen IE, Gibbs JR, Nalls M, Price TR, Lubbe S, van Rooij J, Uitterlinden A, Kraaij R, Williams NM, Brice A, Hardy J, Wood NW, Morris HR, Gasser T, Singleton AB, Heutink P, Sharma M for International Parkinson's Disease Genomics Consortium. Establishing the role of rare coding variants in known Parkinson's disease risk loci.

## ABSTRACT

**Importance**: Many common genetic factors have been identified to contribute to PD susceptibility, improving our understanding of the related underlying biological mechanisms.

**Objective:** To explore the involvement of rare genetic variants in common sporadic risk loci for Parkinson's disease.

**Design, setting and participants:** Using International Parkinson's Disease Genomics Consortium (IPDGC) datasets, we performed a comprehensive study to determine the impact of rare variants in 26 previously published GWAS loci in PD. These datasets encompassed whole-exome sequencing data of 1,167 cases and 1,685 controls and a genotype-array NeuroX dataset, comprising 6,801 cases and 5,970 controls. We applied Prixfixe to select the putative causal genes underneath the GWAS peaks, which was based on underlying functional similarities. The Sequence Kernel Association Test (SKAT) was used to analyze the joint effect of rare, common or both types of variants on PD susceptibility. All genes were tested simultaneously as a gene-set and each gene individually.

**Main outcome and measures:** Phenotypic information was extracted from clinical information gathered by local institutes participating in IPDGC. The genetic data consists of whole exome sequencing and a genotype array enriched for (neurologic) rare coding variants.

**Results:** Using WES and NeuroX data for the total gene-set, we observed a moderate association of common variants, confirming the involvement of the known PD risk loci within our genetic datasets. Focusing on rare variants we identified significant association signals for *LRRK2, STBD1* and *SPATA19.*

**Conclusions and meaning:** Our study suggests an involvement of rare variants within several putatively causal genes underneath previously identified GWAS peaks associated to PD. However, as genetic replication is currently absent, more detailed genetic studies of these loci in larger genetic cohorts will improve our knowledge on the molecular landscape of PD.

## INTRODUCTION

Genetic factors play an important role in Parkinson's disease (PD) pathogenesis. In addition to the discovery of rare variants using family-based linkage studies, resulting in the identification of for example *SNCA, LRRK2, parkin, DJ-1, PINK1* and *VPS35,* numerous genome-wide association studies (GWAS) have shown that common genetic variants increase PD risk.[1] The most recent and largest PD association study[2] identified over 20 common risk variants, confirming many previously associated risk factors.

Nevertheless, heritability estimates indicate that additional genetic risk factors remain to be discovered since a relatively large fraction of PD heritability cannot be explained by known PD risk loci or Mendelian genes.[3-5] GWAS approaches are primarily designed to identify common risk variants by the usage of genotyping arrays. However, emerging evidence suggests that rare variants may explain part of the missing heritability.[6,7] Rare variants in protein coding regions are more likely to affect the function of a gene than common variants which tag the causal variants via linkage disequilibrium (LD) and are often located in non-coding regions of the genome.[8,9] Therefore, rare variants might be of more importance to complex diseases than predicted by the Common Disease-Common Variant hypothesis.[10-13] In contrast to GWAS, exome sequencing studies aim at systematically analyzing coding regions of the genome to identify causal variants in complex diseases.[14] Exome studies have been proven to be effective for studying familial disease[15] but an increasing number of applications for populations-based studies have been developed.[16,17]

In PD, multiple genes have been shown to harbor both common and rare variants which contribute to disease pathogenesis. *SNCA* and *LRRK2* contain both PD-risk associated rare variants with Mendelian effects as common variants that increase PD risk in sporadic patients.[2,18-23] *GBA*, for which an association was first seen in families with Gaucher disease and parkinsonism,[24] is furthermore shown to play a role in PD by both rare coding variants and common risk variants.[2,5,25] Thus, we hypothesize that rare coding variants in the known risk loci for sporadic PD are involved in the genetic etiology of PD. The combined effect of rare variants within recently identified PD risk loci will likely explain an additional portion of PD heritability. We aim to assess this hypothesis by determining the genetic burden of rare variants in the PD risk loci using two exome cohorts of the International Parkinson's Disease Genomics Consortium (IPDGC).

## METHODS

### Subjects

All PD cases included in this study have given written informed consent. Relevant local ethical committees for medical research approved involvement in genetic studies. The PD patients were diagnosed using the UK Brain Bank criteria.[26]

### Whole exome sequencing dataset

The whole exome sequencing (WES) dataset includes 1,167 PD cases and 1,685 controls (post QC) of European ancestry. The PD patients have a tendency towards a young age of onset with an average of 41.2 years (SD = 10.9). 1,201 controls originate from the Rotterdam Study version 1 (RSX1), as we merged the IPGDC WES data with the RSX1 WES data.[27] The samples were sequenced in different batches with two exome capture kits: EZ Exome Library v2.0 (Roche/Nimblegen) and Truseq Exome Enrichment Kit targeting 44.1 Mb and 62 Mb, respectively (Supplementary Table 1). To account for putative technical differences between the different capture kits, we only considered variants that were targeted by both capture protocols and included preQC individual sample missingness (as a reference to sequencing coverage) as covariates during all genetic analyses.



On average, 94.4% of the exome was covered for at least 10x. The 100-bp paired-end reads were sequenced on a HiSeq2000 and aligned to the human reference genome (build hg19) using Barrow Wheeler Aligner (BWA)-MEM[28]. Genome Analysis Toolkit[29] (GATK) called single nucleotide variants (SNVs) and small insertions/deletions (indels) for each sample, resulting in individual gVCF files. Genotypes of all IPDGC and RSX1 exome samples were then jointly called and recalibrated, allowing to merge the distinct WES datasets in a correct manner. Standard GATK filter steps were applied, together with a minimum genotype quality Phred-score of 20 and depth of 8, to only select high-quality variants. Only bi-allelic calls were considered that were located in regions targeted by both capture kits. Supplementary Table 2 reports the exons that have been excluded due to insufficient coverage within one of the exome capture protocols.

### NeuroX dataset

The NeuroX dataset encompasses 6,801 PD cases and 5,970 controls (post QC) of European ancestry. Overlapping samples with the WES dataset were excluded. The average age of onset of the PD patients is 63.0 years (SD = 12.4). The Exome NeuroX array[30] was used consisting of ~240,000 exonic variants standard to the Illumina HumanExome array v1.1 and ~24,000 variants focused on neurologic and neurodegenerative diseases.

## Quality procedures

For individual QC in both the WES and the NeuroX datasets, samples were removed when showing gender ambiguity, dubious heterozygosity/genotype calls, evidence of relatedness, or being a population outlier. The latter two were calculated with LD-pruned common variants. Variant QC procedures were slightly different for the two different datasets. For the WES dataset, variants passed QC when having a minimum call rate > 85% and being in Hardy-Weinberg equilibrium (HWE *p*-values > 1e-8 based on controls). For the NeuroX dataset, variants were excluded for subsequent analyses with a minimum call rate < 95%, a HWE *p*-value < 1e-6, or with significant differences in missingness rate between cases and controls.

## Causal gene selection within PD risk loci

Based on the most recent and largest GWAS[2,31] we selected 26 loci containing at least one meta-analysis result with P < 5.00e-08 (as reported by pdgene.org). The published SNPs associated with PD are not the causal variants but rather tag the unknown causal variants with which they are in LD. As the causal variant (and therefore also the related gene) has not been determined for most of the PD risk loci, we explored the involvement of rare variants in PD susceptibility by using the PrixFixe strategy, which selects one gene per locus based on functional similarities of genes within the LD blocks from the different loci. The functional similarity is defined as the degree of shared biological function and is determined by overlapping biological features such as protein domains, transcription factor binding sites, gene-expression, phylogenetic profiles and literature-based protein-protein interactions. The most significantly associated SNPs from the recent meta-analysis by Nalls et al.[2] were used as seeding SNPs to define the LD region per PD locus. If a SNP was not applicable to be used as seeding SNP (not present in either the current dbSNP v137 or HapMap public resources), the next strongest associated SNP or a SNP in high LD (r2 > 0.8) within the same locus was used as a seed. We were unable to define a legitimate seeding SNP for 3 loci. LD-regions were based on the CEU phase III population with a minimal R2 of 0.5. The final Prixfixe gene-set consists of 23 genes, which are reported in Table 1.

## Variant selection

To enrich both genetic datasets for deleterious variants we selected multiple subsets of variants, differing in the method and stringency to select pathogenic variants. Based on variant annotation with ANNOVAR,[32] 3 distinct subsets of variants were created, including: 1) all exonic variants (disruptive, splicing, (non)synonymous and (non)frameshift indels), 2) amino acid changing variants (same as previous except for synonymous) 3) amino acid changing (AAchanging) variants that are predicted to be deleterious. The latter subset includes variants that are predicted to be pathogenic (CADD score  12.37[33]) by

Combined Annotation Dependent Depletion (CADD) v1.[34] Figure 1 displays a workflow of the classification of the different variant subsets. The exonic subset was exclusively tested for the gene-set analysis to determine the involvement of common PD risk loci in the WES and NeuroX dataset. The Sequence Kernel Association Test (SKAT)[35,36]was used to perform burden analyses. The MAF threshold, separating the rare and common variants, was based on the total sample size using the formula (T = 1/(√(2n)) ) suggested by SKAT,[36]

**Table 1.** Selected set of genes

| Polymorphism | Location (hg19) | *P*-value | Seeding SNP | Prixfixe gene |
|---|---|---|---|---|
| rs71628662 | chr1:155359992 | $6.86 \times 10^{-28}$ | NA | |
| rs823118 | chr1:205723572 | $1.96 \times 10^{-16}$ | rs823114 | *RAB7L1* |
| rs10797576 | chr1:232664611 | $1.76 \times 10^{-10}$ | rs2182431 | *SIPA1L2* |
| rs6430538 | chr2:135539967 | $3.35 \times 10^{-19}$ | rs6430538 | *ACMSD* |
| rs1955337 | chr2:169129145 | $1.67 \times 10^{-20}$ | rs2390669 | *STK39* |
| rs12637471 | chr3:182762437 | $5.38 \times 10^{-22}$ | rs12637471 | *LAMP3* |
| rs11724635 | chr4:15737101 | $4.26 \times 10^{-17}$ | rs11724635 | *FBXL5* |
| rs6812193 | chr4:77198986 | $1.85 \times 10^{-11}$ | rs6812193 | *STBD1* |
| rs356182 | chr4:90626111 | $1.85 \times 10^{-82}$ | rs356219 | *SNCA* |
| rs34311866 | chr4:951947 | $6.0 \times 10^{-41}$ | rs748483 | *MFSD7* |
| rs9275326 | chr6:32666660 | $5.81 \times 10^{-13}$ | rs9275311 | *HLA-DRB5* |
| rs199347 | chr7:23293746 | $5.62 \times 10^{-14}$ | rs199347 | *GPNMB* |
| rs591323 | chr8:16697091 | $3.17 \times 10^{-8}$ | NA | |
| rs117896735 | chr10:121536327 | $1.21 \times 10^{-11}$ | rs10886515 | *RGS10* |
| rs329648 | chr11:133765367 | $8.05 \times 10^{-12}$ | rs329648 | *SPATA19* |
| rs3793947 | chr11:83544472 | $2.59 \times 10^{-08}$ | rs1400313 | *DLG2* |
| rs11060180 | chr12:123303586 | $3.08 \times 10^{-11}$ | rs11060180 | *HIP1R* |
| rs76904798 | chr12:40614434 | $4.86 \times 10^{-14}$ | rs2708435 | *LRRK2* |
| rs7155501 | chr14:55347827 | $1.25 \times 10^{-10}$ | rs2878174 | *LGALS3* |
| rs1555399 | chr14:67984370 | $5.70 \times 10^{-16}$ | rs7155830 | *ARG2* |
| rs2414739 | chr15:61994134 | $3.59 \times 10^{-12}$ | NA | |
| rs14235 | chr16:31121793 | $3.63 \times 10^{-12}$ | rs14235 | *PRSS8* |
| rs17649553 | chr17:43994648 | $6.11 \times 10^{-49}$ | rs17649553 | *MAPT* |
| rs12456492 | chr18:40673380 | $2.15 \times 10^{-11}$ | rs12456492 | *RIT2* |
| rs62120679 | chr19:2363319 | $2.52 \times 10^{-09}$ | rs2074546 | *PLEKHJ1* |
| rs55785911 | chr20:3153503 | $3.30 \times 10^{-10}$ | rs2295545 | *AVP* |

*P*-value = Meta *p*-value as reported on pdegene.org. Seeding SNP = input SNP for PrixFixe software. Prixfixe gene = genes selected based on underlying functional similarities, which is determined by overlapping biological features such as protein domains, transcription factor binding sites, gene expression, phylogenetic profiles and literature-based protein-protein interactions.

therefore resulting in the MAF thresholds of 0.013 and 0.006 for the WES dataset and NeuroX dataset, respectively. We performed polygenetic burden analyses for exclusively rare variants, exclusively common variants and both types of variants together. The common variants were pruned (PLINK[37] indep settings 50 5 1.5) aiming to only consider independent variants in our genetic analyses. For the gene-sets we performed a two-sided SKAT test allowing variants within a gene-set to have different directions and magnitudes off effects, which is in concordance with both damaging and protective effect estimates observed for the 26 published PD loci. To test individual genes we performed a one-sided burden test, as we hypothesized that variants in individual genes are likely to have the same direction of effect. We also performed a two-sided SKAT analysis per gene in case we were interested which genes accounted mostly for the observed rare variant association in the total gene-set.

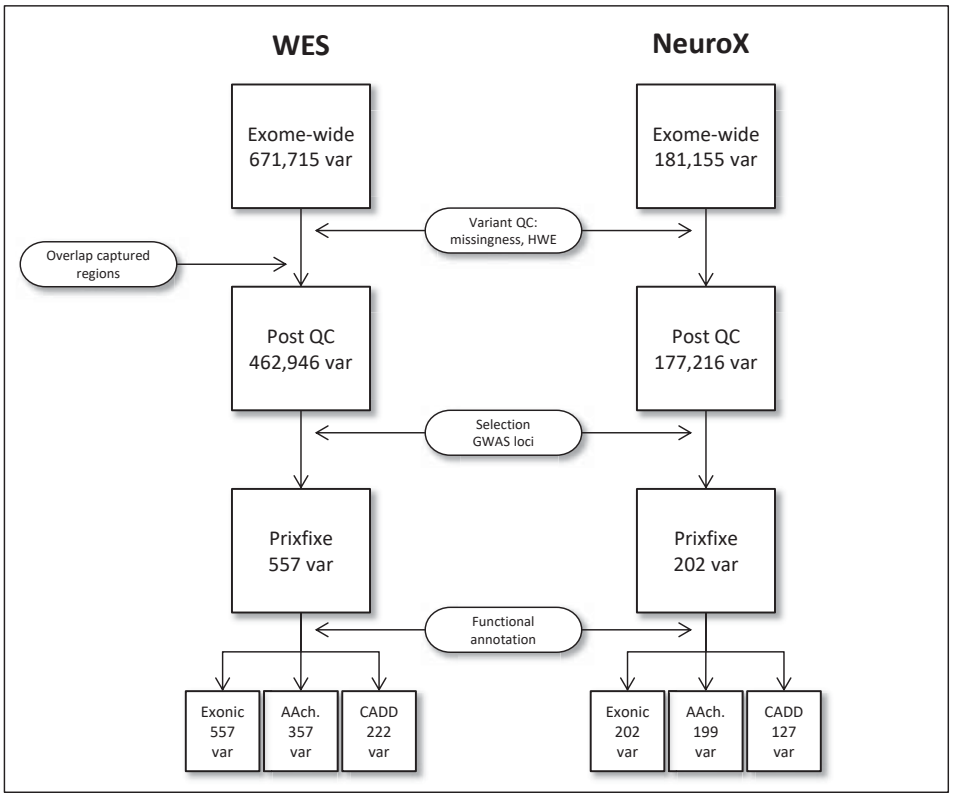



**Figure 1.** Flowchart of variant subset classification. The variants undergo multiple analyses procedures, including quality control, selection of variants within PD loci and functional annotation. Each genetic dataset (WES and NeuroX) is tested for 3 different variant categories, differing in functionality of variants.

To correct for confounding factors (e.g. population stratification and technical artifacts), we included 20 multi-dimensional scaling components, gender and individual missingness rate pre QC (as a reference to the individual WES coverage) for the WES dataset. As the NeuroX dataset is more homogeneous, we corrected for the first 4 MDS components and gender. Empirical *p*-values were calculated for significant sample results ($p < 0.05$). For the gene-set analysis, the original sample *p*-value of the gene-set of interest was compared to *p*-values of 1,000 randomly drawn gene-sets of the same size. For the individual gene associations, empirical *p*-values were calculated using the resampling method implemented by SKAT, by 10,000 permutations of the affection status. Empirical *p*-values are calculated by (n1+1)/(n+1), where n1 = the number of resampling *p*-values smaller than the original sample *p*-value and n = the number of resampling.

### Power calculations

We estimated the power of our study design to detect rare variant associations. Supplementary Table 3 displays the parameters that were chosen for the calculations. For both datasets, the PD prevalence was set to 0.0057[38]. As approximately half of the loci in PDgene.org have an odds ratio below 1, the percentage protective effect was set to 50%. A thousand simulations ($\alpha = 0.025$) were performed on a haplotype matrix of SKAT, mimicking linkage disequilibrium structure of European ancestry, comprising 10,000 haplotypes over 200 kb regions.



## RESULTS

### WES and rare variants

First, we analyzed the WES dataset as it represents all exonic variants, of which the study design has 65% power to detect a rare variant association signal considering individual genes. Testing the aggregated effect of grouped variants within a gene-set has the potential to increase power. Supplementary Table 4 shows the results of the gene-set analyses in the WES dataset. Common exonic variants are moderately associated to PD. The nominal *p*-value is significant, but the empirical *p*-value exceeds 0.05. Although we anticipated a significant association of common variants, we attribute the moderate association to a relatively low sample size (compared to the original GWAS), and the selection of genes (by Prixfixe) with variants in moderate LD with the original highest SNP. The gene-set association is absent when focusing on the common amino acid changing and CADD variants, which is probably due to a decrease in power as the number of variants drops.

No rare variant, or common & rare variant associations were observed for the gene-set in either of the functional variant categories (nominal $p \geq 0.223$; Supplementary Table 4). An alternative approach to study the putative rare variant associations is to test each gene individually within the gene-set. Table 2 displays the 3 strongest associated

**Table 2.** Gene-based rare variant association results for WES dataset.

| Variant type | Gene | *p*-value (emp) | *n* variants | maf cases | maf controls |
|---|---|---|---|---|---|
| AAchanging | *STBD1* | **0.018 (0.046)** | 8 | 0.05% | 0.32% |
| | *HIP1R* | 0.082 | 20 | 0.61% | 0.53% |
| | *STK39* | 0.126 | 4 | 0.20% | 0.00% |
| CADD | *STBD1* | 0.105 | 5 | 0.05% | 0.16% |
| | *SPATA19* | 0.122 | 4 | 0.19% | 0.06% |
| | *GPNMB* | 0.141 | 18 | 0.85% | 0.92% |

*p*-value = nominal *p*-value; (emp.) = empirical *p*-value calculated by comparison to 10,000 permutations of affection status. AAchanging = amino acid changing variants; CADD = variants predicted pathogenic.

genes per variant subset. For the CADD variants, no gene is independently associated to PD. However, investigating the amino-acid changing variant subset resulted in a significant association for *STBD1* (empirical $p = 0.046$).

## NeuroX and rare variants

The NeuroX dataset contains previously identified exonic variants, of which a large proportion is rare.[30] The larger sample size (6,801 cases and 5,970 controls) increases the power (estimated at 96%) to detect a rare variant association signal. Similarly to the WES dataset, a moderate common variant association is detected (nominal $p = 0.031$). In contrast to the WES dataset, we do observe significant associations of the gene-set with PD, even when only considering rare variants (AAchanging = 0.007; CADD = 0.002; Supplementary Table 5a).

To discover whether specific genes drive this observed rare variant association detected with the two-sided SKAT test, the variants were grouped per gene and again two-sided tested for their association to PD. *LRRK2* is the gene driving the association observed in the total gene-set (Supplementary Table 6). Focusing on the CADD subset, this association (nominal $p = 5.17 \times 10^{-13}$) is considerably stronger than the second most significant gene: *SPATA19* (nominal $p = 0.050$). As the NeuroX array is neurology specific, it harbors many variants of the known PD gene *LRRK2*. Furthermore, *LRRK2* is a large gene coding for 2,527 amino acids. As a comparison, NeuroX contains 32 harmful (predicted by CADD) *LRRK2* variants, while only 2 harmful variants are present for *SPATA19*. The variants in *LRRK2* are overrepresented and biasing the results of the total gene-sets. We therefore performed the same gene-set analyses on the NeuroX dataset excluding the variants of *LRRK2* (Supplementary Table 5b), resulting in the absence of a rare variant association in the NeuroX dataset (nominal $p \geq 0.28$). This suggests that the previously observed association of rare variants within the total gene-set to PD was solely driven by *LRRK2.*

The two-sided SKAT analysis per gene aimed at the discovery of genes driving the rare variant association in the total gene-set. Next, we were interested to explore the

genetic burden of rare variants for each gene individually when assuming all rare variants to have the same direction of effect (one-sided BURDEN test). Table 3 shows again that *LRRK2* (empirical $p$ = 0.0005) is the strongest associated gene. Furthermore, *SPATA19* (empirical $p$ = 0.017) is significantly associated when specifically considering rare CADD variants.

**Table 3.** Gene-based rare variant association results for NeuroX dataset.

| Variant type | Gene | *p*-value (emp) | *n* variants | maf cases | maf controls |
|---|---|---|---|---|---|
| AAchanging | *LRRK2* | **0.0004 (0.0005)** | 48 | 1.70% | 1.13% |
| | *RIT2* | 0.051 | 2 | 0.00% | 0.03% |
| | *PRSS8* | 0.098 | 1 | 0.04% | 0.01% |
| CADD | *LRRK2* | **0.0003 (0.0005)** | 32 | 1.38% | 0.86% |
| | *SPATA19* | **0.014 (0.017)** | 2 | 0.05% | 0.00% |
| | *RIT2* | 0.051 | 2 | 0.00% | 0.03% |

*p*-value = nominal *p*-value; (emp.) = empirical *p*-value calculated by comparison to 10,000 permutations of affection status. AAchanging = amino acid changing variants; CADD = variants predicted pathogenic



## Directionality of effect

We further explored the significant individual association signals (empirical $p < 0.05$) for *LRRK2*, *STBD1*, and *SPATA19.* By focusing on the variant level we aimed to comprehend the directions of effects. *LRRK2* showed a significant burden of 32 rare damaging variants in the exome NeuroX dataset. Single-marker association analysis of these variants revealed one extremely strong associated variant ($p = 3.17 \times 10^{-13}$), which we identified as the PD-related pathogenic variant p.G2019S (rs34637584) in 78 cases (MAF = 0.006). Performing the rare variant aggregation test on 31 pathogenic *LRRK2* variants, excluding p.G2019S, resulted in no association ($p$ = 0.98) to PD. This implies that the original observed rare variant association in *LRRK2* was solely driven by the p.G2019S variant. As this variant is only present in 7 cases in the WES dataset (MAF = 0.003) with a single-marker *p*-value of 0.002 (*LRRK2* mutations generally observed in late-onset PD), it explains why *LRRK2* as a gene was not significant in the WES dataset, while it showed a strong association in the NeuroX dataset.

In addition to the rare variant association test in *LRRK2*, we explored the presence of the previously published common *LRRK2* haplotype with a protective effect of 3 exonic variants (p.N551K-p.R1398H-p.K1423K).[39] p.K1423K is not included in the NeuroX genotyping array, but is in high linkage-disequilibrium ($r^2$ = 1.00) with p.R1398H. We therefore tested the p.N551K-p.R1398H (G-A) haplotype and confirmed the protective effect (OR = 0.89, $p$ = 0.027) of this haplotype for the PD cases, showing a minor haplotype

frequency of 6.2% in cases and 6.9% in controls. All 3 variants were detected in the WES dataset, allowing to test the full haplotype (G-A-A). Although the haplotype association was not significant in the WES dataset (OR = 0.81, *p* 0.223), the trend of effect is similar with a minor haplotype frequency of 7.0% in cases and 7.5% in controls. The smaller sample size of the WES dataset is a plausible reason for not obtaining a significant association.

Next, the WES-based *STBD1* and NeuroX-based *SPATA19* were investigated for their variant frequencies. Single-marker association analysis showed no significant individual results for the 8 variants within *STBD1*. It therefore appears that the observed rare variant association is not caused by one exclusive variant but is rather the effect of multiple rare variants. Seven of the 8 variants are control-specific as they are only present in 10 control individuals. In contrast, only 1 variant is present in a single case. The direction of effect of the variants that are generating the *STBD1* gene association is therefore implied to be protective. The significant gene-based association for *SPATA19* is relatively strong considering that it is driven by only 2 CADD variants that are present in 7 cases and 0 controls. The absence of *SPATA19* CADD variants in controls suggests that the association signal is damaging. However, *SPATA19* variants could have been missed as the NeuroX dataset is array-based.



## DISCUSSION

To establish the influence of rare variants in sporadic PD risk loci, we explored two independent PD datasets (WES and NeuroX) enriched for coding rare variants. We used the PrixFixe strategy to select the most likely causal genes underlying the PD loci peaks, which is based on overlapping biological functional similarities. We tested both the effect of rare variants in the gene-set at once, as each gene individually. Aggregating variants simultaneously across a set of genes has the potential to increase power to detect an association signal, given that the selected genes are enriched for a group of genes that are genuinely involved in the disease pathogenesis.

The average age of onset within the case group of the WES dataset (~41 years) is 20 years younger than in the meta-analysis of the most recent PD GWAS (~61 years) where the PD risk loci were based on. As some rare genetic risk factors (*DJ-1, parkin* and *PINK1*)[1] are specific for young onset PD (YOPD), we acknowledge the putative existence of YOPD-specific common genetic risk factors within the WES dataset. However, risk factors related to late onset sporadic PD might also play a role in YOPD. PD risk loci, such as *SNCA* and *GBA*,[2,40] overlap between late and young onset types. We therefore expect that our WES dataset is an adequate dataset to study the rare exonic variants in PD risk loci. Furthermore, YOPD is often genetically explained through rare variants.[1] The YOPD patient group in the WES dataset could therefore be enriched for cases which are genetically

influenced by rare variants, possibly increasing the likelihood of detecting rare variant associations.

Using gene-set approach in the WES dataset, we did not detect a burden of rare variants when comparing PD subjects to controls. However, it remains unclear whether the absence of association is genuine or due to insufficient power or incorrect selection of causal genes. In contrast, with the gene based association test for the genes selected with the Prixfixe strategy we observed a rare variant association for *STBD1,* implying that rare variants in this gene could be a risk factor for PD. *STBD1* has its function in lysosomal-mediated autophagy that has been shown to contribute to PD pathogenesis.[41] Genetic replication or functional validation for *STBD1* is required.

We detected strong associations of rare variants within the gene-set for the NeuroX dataset. However, subsequent analyses showed that these associations were dominated by *LRRK2* variants. Association analysis on variant level revealed that the *LRRK2* gene signal was driven by the known p.G2019S variant. This observation highlights the importance for every aggregated variant study to further investigate variant aggregation association results on variant level to correctly draw conclusions. As shown for the *LRRK2* association and even the total gene-set association, it is driven by only 1 variant, which also could have been detected with the performance of a simple single-marker association test. Besides the pathogenic association signal of rare variant p.G2019S, we observed a significant protective effect of a previously published common haplotype.[39] This observation supports the theory that other variants with opposite effects could interact and potentially influence the penetrance of pathogenic *LRRK2* variants, such as p.G2019S. Besides *LRRK2*, we furthermore detected a NeuroX-based burden of rare CADD variants for *SPATA19* that increases PD risk ($p$ = 0.017). Although the biological function of *SPATA19* is poorly studied, the GTEx portal displays specific high expression for the testis, diminishing the likelihood that defects of this gene would contribute to neurodegeneration.

By selecting genes based on biological similarities we aimed to identify the true causal loci. As we expect that only one gene per locus is the true causal gene, we did not define a gene-set including all the genes underneath the GWAS loci assuming the overrepresentation of non-causal genes would dilute a putative association signal. We acknowledge that the ultimate strategy to test the effect of rare variants in the PD loci would be to sequence all genes in a large cohort, and test the effect of rare variants in each gene individually. Furthermore, sequencing rather than genotyping will define novel rare variants and contribute to cataloguing the influence of rare variants underneath the PD risk loci.

Our study suggests for the first time that, apart from *LRRK2, SNCA* and *GBA,* other common PD risk loci might harbor rare variants that contribute to PD risk. However, as the significant associations for *STBD1* and *SPATA19* were specific for either the WES or NeuroX,

respectively, we are currently unable to draw definite conclusions on their involvement in PD etiology. Although a lack of replication could be explained by a spurious initial result, an alternative reason could be the differences in age of disease onset and applied laboratory techniques, potentially complicating a genetic replication. Future studies could benefit of analyzing more homogenous datasets, yet this scenario is not always feasible. Furthermore, this study emphasizes the current lack of knowledge to interpret GWAS risk loci. Therefore, to enhance more detailed genetic and functional studies of PD risk loci, the genetics research field is in need of valuable bioinformatics approaches that have the ability to generate high-quality predictions for the identification of the genuine causal genes.

## ACKNOWLEDGMENTS

This work was supported in part by the Prinses Beatrix Spierfonds (I.E.J. and P.H.) and the Intramural Research Program of the National Institute on Aging, National Institutes of Health, Department of Health and Human Services; project ZO1 AG000949.

The generation and management of the exome sequencing data for the Rotterdam Study was executed by the Human Genotyping Facility of the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Netherlands. The Exome Sequencing data set was funded by the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO) sponsored Netherlands Consortium for Healthy Aging (NCHA; project nr. 050-060-810), by the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, and by the and by a Complementation Project of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL; www.bbmri.nl ; project number CP2010-41). We thank Mr. Pascal Arp, Ms. Mila Jhamai, BSc and Mr. Marijn Verkerkj for their help in creating the RS-Exome Sequencing database.

The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. The authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists.

This study is also supported by the Courage-PD is an EU Joint Programme - Neurodegenerative Disease Research (JPND) project (M.S, T.G.) The project is supported through the following funding organizations under the aegis of JPND - www.jpnd.eu:

the Medical Research Council, United Kingdom

the French National Research Agency

the German Bundesministerium für Bildung und Forschung

the Italian Ministry of Health/Ministry of Education, Universities and Research
the Israeli Ministry of Health
the Luxembourgian National Research Fund
the Netherlands Organisation for Health Research and Development
the Research Council of Norway
the Portuguese Foundation for Science and Technology
the Spanish National Institute of Health Carlos III

M.S. is also supported by the Michael J Fox Foundation, USA

## REFERENCES


1. Bras J, Guerreiro R, Hardy J. SnapShot: Genetics of Parkinson's disease. *Cell* 2015; 160(3): 570-.e1.
2. Nalls MA, Pankratz N, Lill CM, et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nature genetics* 2014; 46(9): 989-93.
3. Pihlstrom L, Toft M. Parkinson's disease: What remains of the "missing heritability"? *Movement disorders : official journal of the Movement Disorder Society* 2011; 26(11): 1971-3.
4. Keller MF, Saad M, Bras J, et al. Using genome-wide complex trait analysis to quantify 'missing heritability' in Parkinson's disease. *Human molecular genetics* 2012; 21(22): 4996-5009.
5. Do CB, Tung JY, Dorfman E, et al. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS genetics* 2011; 7(6): e1002141.
6. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature* 2009; 461(7265): 747-53.
7. Zuk O, Schaffner SF, Samocha K, et al. Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America* 2014; 111(4): E455-64.
8. Nelson MR, Wegmann D, Ehm MG, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science (New York, NY)* 2012; 337(6090): 100-4.
9. Tennessen JA, Bigham AW, O'Connor TD, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science (New York, NY)* 2012; 337(6090): 64-9.
10. Lander ES. The new genomics: global views of biology. *Science (New York, NY)* 1996; 274(5287): 536-9.
11. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant...or not? *Human molecular genetics* 2002; 11(20): 2417-23.
12. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature genetics* 2003; 33 Suppl: 228-37.
13. Sharma M, Kruger R, Gasser T. From genome-wide association studies to next-generation sequencing: lessons from the past and planning for the future. *JAMA neurology* 2014; 71(1): 5-6.
14. Kiezun A, Garimella K, Do R, et al. Exome sequencing and the genetic basis of complex traits. *Nature genetics* 2012; 44(6): 623-30.
15. Bamshad MJ, Ng SB, Bigham AW, et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature reviews Genetics* 2011; 12(11): 745-55.
16. Purcell SM, Moran JL, Fromer M, et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* 2014; 506(7487): 185-90.
17. Cirulli ET, Lasseigne BN, Petrovski S, et al. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science (New York, NY)* 2015; 347(6229): 1436-41.
18. Polymeropoulos MH, Lavedan C, Leroy E, et al. Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science (New York, NY)* 1997; 276(5321): 2045-7.
19. Paisan-Ruiz C, Jain S, Evans EW, et al. Cloning

of the gene containing mutations that cause PARK8-linked Parkinson's disease. *Neuron* 2004; 44(4): 595-600.
20. Zimprich A, Biskup S, Leitner P, et al. Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. *Neuron* 2004; 44(4): 601-7.
21. Simon-Sanchez J, Schulte C, Bras JM, et al. Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nature genetics* 2009; 41(12): 1308-12.
22. Edwards TL, Scott WK, Almonte C, et al. Genome-wide association study confirms SNPs in SNCA and the MAPT region as common risk factors for Parkinson disease. *Annals of human genetics* 2010; 74(2): 97-109.
23. Nalls MA, Plagnol V, Hernandez DG, et al. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet* 2011; 377(9766): 641-9.
24. Goker-Alpan O, Schiffmann R, LaMarca ME, Nussbaum RL, McInerney-Leo A, Sidransky E. Parkinsonism among Gaucher disease carriers. *Journal of medical genetics* 2004; 41(12): 937-40.
25. Pankratz N, Beecham GW, DeStefano AL, et al. Meta-analysis of Parkinson's disease: identification of a novel locus, RIT2. *Annals of neurology* 2012; 71(3): 370-84.
26. Hughes AJ, Daniel SE, Kilford L, Lees AJ. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. *Journal of neurology, neurosurgery, and psychiatry* 1992; 55(3): 181-4.
27. Hofman A, Brusselle GG, Darwish Murad S, et al. The Rotterdam Study: 2016 objectives and design update. *European journal of epidemiology* 2015; 30(8): 661-708.
28. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* 2009; 25(14): 1754-60.
29. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* 2010; 20(9): 1297-303.
30. Nalls MA, Bras J, Hernandez DG, et al. NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases. *Neurobiology of aging* 2014.
31. Lill CM, Roehr JT, McQueen MB, et al. Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: The PDGene database. *PLoS genetics* 2012; 8(3): e1002548.
32. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* 2010; 38(16): e164.
33. Amendola LM, Dorschner MO, Robertson PD, et al. Actionable exomic incidental findings in 6503 participants: challenges of variant classification. *Genome research* 2015; 25(3): 305-15.
34. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM. A general framework for estimating the relative pathogenicity of human genetic variants. 2014; 46(3): 310-5.
35. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *American journal of human genetics* 2011; 89(1): 82-93.
36. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. *American journal of human genetics* 2013; 92(6): 841-53.
37. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association

and population-based linkage analyses. *American journal of human genetics* 2007; 81(3): 559-75.
38. Pringsheim T, Jette N, Frolkis A, Steeves TD. The prevalence of Parkinson's disease: a systematic review and meta-analysis. *Movement disorders : official journal of the Movement Disorder Society* 2014; 29(13): 1583-90.
39. Ross OA, Soto-Ortolaza AI, Heckman MG, et al. Association of LRRK2 exonic variants with susceptibility to Parkinson's disease: a case-control study. *Lancet neurology* 2011; 10(10): 898-908.
40. Klein C, Westenberger A. Genetics of Parkinson's disease. *Cold Spring Harbor perspectives in medicine* 2012; 2(1): a008888.
41. Trinh J, Farrer M. Advances in the genetics of Parkinson disease. *Nature reviews Neurology* 2013; 9(8): 445-54.

## SUPPLEMENTAL DATA

**Table 1.** WES capture protocols

| | cases | controls | |
|---|---|---|---|
| | **IPDGC** | **IPDGC** | **RSX1** |
| Nimblegenv2 | 252 | 37 | 1201 |
| Truseq | 912 | 446 | 0 |
| Mixed | 3 | 1 | 0 |
| | | | |
| Total | 1167 | 484 | 1201 |

Mixed = samples that have been captured using the 2 distinct capture kits.

**Table 2.** Exclusion of exons based on capture inconsistencies.

| | gene | exon | Source |
|---|---|---|---|
| PD meta | *ASH1L* | 21 | Truseq |
| | *DLG2* | 1+2 | Nimblegenv2 |
| | *TMEM229B* | 1+2 | Nimblegenv2 |
| | *TMEM175* | 1 | Nimblegenv2 |

**Table 3.** Parameters for power calculations.

| Arguments | WES | NeuroX |
|---|---|---|
| Subreg. Length | 3205 | 3205 |
| Prevalence PD | 0.0057 | 0.0057 |
| % protective effect | 50 | 50 |
| n samples | 2852 | 12771 |
| Case proportion | 0.41 | 0.53 |
| Causal MAF cutoff | 0.013 | 0.006 |
| % causal variants | 40 | 52 |

Subregion length = the average length of transcripts corresponding to the genes included in the gene-sets. % protective effect = % of causal variants with a negative coefficient. Causal MAF cutoff is similar to common/rare variant cut-off. % causal variants = % of CADD variants.

**Table 4.** Gene-set association results of WES dataset.

| | | Rare | | Common | | Common & rare | |
|---|---|---|---|---|---|---|---|
| Gene-set | Variant type | *p*-value (emp.) | *n* variants | *p*-value (emp.) | *n* variants | *p*-value (emp.) | *n* variants |
| Prixfixe | exonic | | | 0.014 (0.074) | 29 | | |
| | AAchanging | 0.227 | 343 | 0.319 | 14 | 0.223 | 357 |
| | CADD | 0.189 | 212 | 0.414 | 10 | 0.247 | 222 |

*p*-value = nominal *p*-value; (emp.) = empirical *p*-value calculated by comparison to 1,000 randomly drawn gene-sets of same size. *P*-values in bold are significant. MAF cut-off to separate rare and common variants is 0.013 (based on sample size).

**Table 5.** Gene-set association results of neuroX dataset.

| | | Rare | | Common | | Common & rare | |
|---|---|---|---|---|---|---|---|
| Gene-set | Variant type | *p*-value (emp.) | *n* variants | *p*-value (emp.) | *n* variants | *p*-value (emp.) | *n* variants |
| a. *LRRK2* included | exonic | | | 0.031 (0.101) | 18 | | |
| | AAchanging | **$1.06 \times 10^{-5}$ (0.007)** | 176 | 0.0084 (0.053) | 23 | **$8.45 \times 10^{-7}$ (0.026)** | 199 |
| | CADD | **$5.99 \times 10^{-7}$ (0.002)** | 114 | **0.0032 (0.034)** | 13 | **$8.58 \times 10^{-8}$ (0.020)** | 127 |
| b. *LRRK2* excluded | exonic | | | 0.243 | 13 | | |
| | AAchanging | 0.28 | 128 | 0.154 | 16 | 0.367 | 144 |
| | CADD | 0.70 | 82 | 0.197 | 8 | 0.411 | 90 |

*p*-value = theoretical *p*-value; (emp.) = empirical *p*-value calculated by comparison to 1,000 randomly drawn gene-sets of same size. Boldfaced *p*-values are significant. MAF cut-off to separate rare and common variants is 0.006 (based on sample size).

**Table 6.** Gene-based rare variant association results for NeuroX dataset.

| Variant type | Gene | *p*-value | *n* variants |
|---|---|---|---|
| AAchanging | *LRRK2* | $4.32 \times 10^{-13}$ | 48 |
| | *PRSS8* | 0.098 | 1 |
| | *RIT2* | 0.129 | 2 |
| CADD | *LRRK2* | $5.17 \times 10^{-13}$ | 32 |
| | *SPATA19* | 0.050 | 2 |
| | *HIP1R* | 0.091 | 9 |

*p*-value = nominal *p*-value; AAchanging = amino acid changing variants; CADD = variants predicted pathogenic.