

## *Supplemental Information*

### **Haploinsufficiency of *ITSN1* is associated with Parkinson's disease**

Thomas P. Spargo, PhD<sup>1\*</sup>, Chloe F. Sands, BS<sup>2\*</sup>, Isabella R. Juan, BS<sup>3</sup>, Jonathan Mitchell, PhD<sup>1</sup>, Vida Ravanmehr, PhD<sup>2</sup>, Jessica C. Butts, PhD<sup>4</sup>, Ruth B. De-Paula, MSc<sup>5</sup>, Youngdo Kim, PhD<sup>3</sup>, Fengyuan Hu, PhD<sup>1</sup>, Quanli Wang, MS<sup>6</sup>, Dimitrios Vitsios, PhD<sup>1</sup>, Manik Garg, PhD<sup>1</sup>, Mirko Messa, PhD<sup>7</sup>, Guillermo del Angel, PhD<sup>8</sup>, Daniel G. Calame, MD PhD<sup>3,9</sup>, Hiba Saade, MD<sup>10,11</sup>, Laurie Robak, MD PhD<sup>3,11</sup>, Ben Hollis, PhD<sup>1</sup>, Huda Y. Zoghbi, MD<sup>3,9,11,12,13,14</sup>, Joshua Shulman, MD PhD<sup>3,10,11,13,14</sup>, Slavé Petrovski, PhD<sup>1,15</sup>, Ismael Al-Ramahi, Ph.D.<sup>3,11,14</sup>, Ioanna Tachmazidou, PhD<sup>1</sup>, Ryan S. Dhindsa, MD, PhD<sup>2,3,6,11</sup>

*\* These authors contributed equally*

1. Centre for Genomics Research, Discovery Sciences, R&D, AstraZeneca, Cambridge, UK
2. Department of Pathology and Immunology, Baylor College of Medicine, Houston, TX, USA
3. Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA
4. Department of Bioengineering, George R. Brown School of Engineering, Rice University, Houston, TX, USA
5. Quantitative and Computational Biology Program, Baylor College of Medicine, Houston, TX, USA
6. Centre for Genomics Research, Discovery Sciences, R&D, AstraZeneca, Waltham, MA, USA
7. Translational Genomics, Centre for Genomics Research, Discovery Sciences BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden
8. Centre for Genomics Research, Discovery Sciences, R&D, AstraZeneca, Boston, MA, USA
9. Department of Pediatrics, Texas Children's Hospital, Houston, TX, USA
10. Department of Neurology, Baylor College of Medicine, Houston, TX, USA
11. Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, 1250 Moursund St., Suite N.1150, Houston, TX, 77030, USA
12. Howard Hughes Medical Institute, Baylor College of Medicine, Houston, TX, USA
13. Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA
14. Center for Alzheimer's and Neurodegenerative Diseases, Baylor College of Medicine, Houston, TX, USA
15. Department of Medicine, Austin Health, University of Melbourne, Melbourne, Australia

#### **Correspondence:**

Ryan Dhindsa, MD PhD

[ryan.dhindsa@bcm.edu](mailto:ryan.dhindsa@bcm.edu)

+1 (832) 824-8633

1250 Moursund St.,

Houston, TX 77027

Slavé Petrovski, PhD

[slav.petrovski@astrazeneca.com](mailto:slav.petrovski@astrazeneca.com)

+44 (0) 7384 239426

1 Francis Crick Ave,

Trumpington, Cambridge CB2 0AA, United Kingdom

## Table of Contents

Supplementary Methods.....	3
UK Biobank dataset.....	3
AMP-PD dataset.....	4
The 100,000 Genomes Project (100kGP) dataset.....	6
All of Us dataset .....	6
Power analysis for <i>ITSN1</i> replication .....	8
Supplementary Table Legends .....	9
Supplementary Figures.....	13
References .....	24

## Supplementary Methods

### UK Biobank dataset

#### *Sequencing and variant calling*

Whole-genome sequencing (WGS) data of the UKB participants were generated by deCODE Genetics and the Wellcome Trust Sanger Institute as part of a public-private partnership involving AstraZeneca, Amgen, GlaxoSmithKline, Johnson & Johnson, Wellcome Trust Sanger, UK Research and Innovation, and the UKB. The WGS sequencing methods have been previously described.<sup>1,2</sup> Briefly, genomic DNA underwent paired-end sequencing on Illumina NovaSeq6000 instruments with a read length of 2x151 and an average coverage of 32.5x. Conversion of sequencing data in BCL format to FASTQ format and the assignments of paired-end sequence reads to samples were based on 10-base barcodes, using `bcl2fastq v2.19.0`. Initial quality control was performed by deCODE and Wellcome Sanger, which included sex discordance, contamination, unresolved duplicate sequences, and discordance with microarray genotyping data checks.

UK Biobank genomes were processed at AstraZeneca using the provided CRAM format files. A custom-built Amazon Web Services cloud compute platform running Illumina DRAGEN Bio-IT Platform Germline Pipeline v3.7.8 was used to align the reads to the GRCh38 genome reference and to call small variants. Variants were annotated against the transcript for which they are most damaging using `SnEff v4.3`<sup>3</sup> and Ensembl Build 38.92.<sup>4</sup>

#### *Sample QC*

Genome sequencing data from 490,560 people were downloaded from the UK Biobank in April 2023. Phenotypic data used to identify people with Parkinson's disease (cases) and non-neurological controls (see **Table 2**) were accessed in April 2022.

Sample-level filters were applied as previously described to derive a subset of UKB suitable for analysis.<sup>5</sup> We removed any person withdrawn from UKB and without linked WGS data. Where available, WGS data were checked for concordance with previous exome sequencing and genotyping array data releases using the KING relatedness software v2.2.3,<sup>6</sup> excluding people with  $<0.49$  estimated kinship with their exome-sequencing or  $<0.465$  with the array data.

Of the remaining samples, we retained only those with `VerifyBamID FREEMIX`<sup>7</sup> contamination  $<0.04$  and  $\geq 10x$  coverage across  $\geq 94.5\%$  of Consensus Coding Sequence (CCDS) database (release 22) bases.<sup>8</sup> The cohort was filtered further using the `ukb_gen_samples_to_remove` function from the `ukbtools (v0.11.3)`<sup>9</sup> R package to select the largest subset of individuals with  $\leq 0.3536$  kinship estimate from KING, removing any first-degree relatives.

The data were next stratified by inferred ancestry across four superpopulation cohorts: African (AFR), East Asian (EAS), European (EUR), and South Asian (SAS). Ancestry was inferred based on the 1000 Genomes phase 3 cohort<sup>10</sup> using the `Peddy (v0.4.2)` software package<sup>11</sup> and we retained people with  $\geq 0.90$  probability of belonging to an ancestry group. For EUR ancestry, which represents  $>90\%$  of the dataset, we further restricted to people  $\leq 4$

standard deviations (SD) from the mean of the first four genetic principal components (PCs). To avoid analytical confounding, female controls were pseudo-randomly downsampled if there was a significant difference in the odds of being male across cases and controls from a given ancestry cohort (Fisher's Exact Test p-value <0.05). **Table S1** overviews per-ancestry cohorts before and after sex rebalancing, where applied.

#### *Variant-level (ExWAS) analysis*

The following in-sample variant QC metrics were applied to select variants for inclusion in the ExWAS analysis: coverage  $\geq 10x$ ;  $\geq 0.20$  of reads are for the alternate allele for heterozygous genotype calls; binomial test of alternate allele proportion departure from 50% in heterozygous state  $p \geq 1 \times 10^{-6}$ ; GQ  $\geq 20$ ; Fisher Strand Bias (FS)  $\leq 200$  for indels and  $\leq 60$  for SNVs; root mean square mapping quality (MQ)  $\geq 40$ ; QUAL  $\geq 30$ ; read position rank sum score (RPRS)  $\geq -2$ ; mapping quality rank score (MQRS)  $\geq -8$ ; DRAGEN variant status = PASS;  $\leq 10\%$  of the cohort have missing genotypes. Additional out-of-sample QC filters were also imposed based on the gnomAD v2.1.1 exomes (GRCh38 liftover) dataset.<sup>12</sup> The site of all variants should have  $\geq 10x$  coverage in  $\geq 30\%$  of gnomAD exomes and, if present, the variant should have an allele count  $\geq 50\%$  of the raw allele count (before removing low-quality genotypes) in the dataset. Variants with missing values on a particular filter were treated as passing that filter, and therefore retained unless failing another metric. Those present in fewer than 6 people from UKB EUR ancestry population, or failing QC in  $>20,000$  people, were also removed.

#### *Gene-level collapsing analysis*

Ten collapsing models were defined to test gene-level associations with different types of coding-sequence variation, including a synonymous variant model as an empirical negative control. Per-model filters for selecting qualifying variants are described in **Table S3**. The following in-sample variant QC metrics were applied universally across the collapsing analysis models: coverage  $\geq 10x$ ; present in CCDS (release 22);<sup>8</sup>  $\geq 0.8$  of reads are for the alternate allele among homozygous genotype calls;  $[0.25, 0.8]$  of reads are for the alternate allele for heterozygous genotype calls; binomial test of alternate allele proportion departure from 50% in heterozygous state  $p \geq 1 \times 10^{-6}$ ; GQ  $\geq 20$ ; FS  $\leq 200$  for indels and  $\leq 60$  for SNVs; MQ  $\geq 40$ ; QUAL  $\geq 30$ ; RPRS  $\geq -2$ ; MQRS  $\geq -8$ ; DRAGEN variant status = PASS. Additional out-of-sample QC filters were also imposed based on the gnomAD v2.1.1 exomes (GRCh38 liftover) dataset.<sup>12</sup> For all variants, the site should have  $\geq 10x$  coverage in  $\geq 25\%$  of gnomAD exomes. For variants present in gnomAD, we retained those with RPRS  $\geq -2$  and MQ  $\geq 30$  in that cohort. Variants with missing values on a particular filter were treated as passing that filter and retained unless failing another metric.

### **AMP-PD dataset**

#### *Sequencing and variant calling*

WGS data for cohorts contributing to AMP-PD were generated by Macrogen and the Uniformed Services University of Health Sciences using the Illumina HiSeq X Ten sequencer

with samples coming from whole blood. Data processing was performed on the Google Cloud Platform and in Terra (<https://app.terra.bio/>). All data processing was performed against Build 38 of the Human Genome reference (GRCh38DH, 1000 Genomes Project version). Analysis was based on the AMP-PD joint-genotyping v4 release which consists of samples from the BioFIND, HBS, LBD, LCC, PDBP, PPMI, Steady PD, and SURE PD cohorts. These data were accessed in April 2024.

Single sample alignment and variant calling was performed using the Broad Institute's Functional Equivalent Workflow (<https://github.com/amp-pd/amp-pd-workflows/tree/master/wgs/wgs-germline-snps-indels>). Joint genotyping was performed using the Broad Institute's joint discovery workflow (<https://github.com/amp-pd/amp-pd-workflows/tree/master/wgs/joint-discovery>). Joint-genotyped VCFs were processed through the Variant Transforms tool from Google Cloud (<https://github.com/googlegenomics/gcp-variant-transforms>), annotating variants against multiple transcripts with the Ensembl Variant Effect Predictor (VEP)<sup>13</sup> and converting to Variant Transforms format.

### *Sample QC*

AMP-PD WGS samples passing the following QC protocol, which has also been described previously,<sup>14</sup> were used in joint-genotyping. Samples were retained in accordance with the following inclusion criteria: VerifyBamID FREEMIX  $\leq 0.03$ , Mean coverage  $\geq 25$  reads per variant, joint calling missingness  $< 0.05$ , Transition transversion ratio (TiTv)  $\geq 2$  for variants in dbSNP (v138), deviation from expected autosomal homozygous genotype counts (Absolute method-of-moments F  $\leq 0.15$ ). Sex was inferred from genetic data using plink, and samples with an explicit mismatch to self-reported sex ('male' or 'female') were removed.

The KING relatedness software was used for estimation of relatedness in the cohort and to infer ancestry. Concordance between WGS and NeuroX genotyping array<sup>15</sup> samples was tested where available, excluding people with matched samples predicted to be less closely related than duplicates / MZ twins. A deduplication step was performed across the WGS samples, excluding the lower-quality WGS sample from each pair predicted to be duplicates / MZ twins.

Ancestry was inferred based on principal component analysis against the HapMap 3 reference panel.<sup>16</sup> Continental ancestries were assigned to people within  $\pm 6$  standard deviations of the mean of PC1 and PC2 for that population in the HapMap reference. Conflicts between self-reported race and genetically-inferred ancestry were resolved by removing people inferred to be 'African' or 'Asian' and clinically-reported 'white' and those inferred to be 'European' and clinically-reported as 'African' or 'Asian' from joint-genotyping. For the current study, we prioritized analysis of a homogenous European ancestry cohort and so retained only people reporting 'white' race that were not flagged as having admixed ancestry during the prior ancestry inference step.

### *Gene-level analysis*

Analysis was performed based on gene-level aggregation of variants called during the central AMP-PD joint-genotyping protocol that passed GATK QC filters and occurring at sites present

in CCDS (release 22).<sup>8</sup> The *ptv* collapsing analysis model was implemented in AMP-PD (see **Table S3**). Qualifying variants were included on the basis of consequences for any transcript of *ITSN1*.

## The 100,000 Genomes Project (100kGP) dataset

### *Sequencing and variant calling*

Whole-genome sequencing data were generated, as previously described, using TruSeq DNA polymerase-chain-reaction (PCR)-free sample preparation kit (Illumina) on the HiSeq2500 platform. Reads were aligned using the Isaac Genome Alignment Software, and the Platypus variant caller was used for small variant calling.<sup>17</sup> Variants were annotated using VEP (v105) with the gnomAD plugin included.<sup>13</sup>

### *Sample QC*

Samples from 100kGP that were retained for analysis were determined in-line with the following QC protocol inclusion criteria: VerifyBamID FREEMIX contamination  $\leq 0.03$ ; aligned reads have  $\geq 15x$  coverage across 95% of the genome with MQ  $> 10$ ;  $> 90\%$  concordance between variant calls from sequencing and matched genotyping array; median fragment size  $> 250bp$ ;  $< 5\%$  chimeric reads;  $> 60\%$  mapped reads;  $< 10\%$  AT dropout; concordance between self-reported and genetically determined sex.

Ancestry was inferred by training a random forest model upon PCs 1-8 from the 1000 Genomes phase 3 cohort.<sup>10</sup> People with  $> 99\%$  probability of European ancestry and  $< 4SD$  from the mean of PCs 1-4 in the 100kGP inferred-European cohort were retained for analysis. Finally, we removed one from each pair of individuals estimated to have  $\geq 2nd$  degree relatedness (kinship coefficient  $> 0.0442$ ) under the plink2 implementation of the KING-relatedness software. Here we followed a 3-step procedure, removing one from each pair of related cases with the *ukb\_gen\_samples\_to\_remove* R function, removing controls related to remaining cases, and finally using *ukb\_gen\_samples\_to\_remove* to removing one from each related pair of remaining controls. Sex-rebalancing was then performed as for the UK Biobank dataset so that the odds of being male was comparable in cases and controls.

### *Gene-level analysis*

Variants within the 100kGP dataset that were retained for analysis passed the following in-sample inclusion criteria: site missingness  $\leq 0.05$ ; median coverage  $\geq 10x$ ; Median GQ  $\geq 15$ ;  $\geq 0.25$  heterozygous allele calls do not show significant ( $p < 0.01$ ) allele imbalance in binomial test;  $\geq 0.50$  complete genotype data for a variant; Hardy-Weinberg equilibrium mid  $p > 1 \times 10^{-5}$  among unrelated samples of European ancestry.

We tested gene-level associations to *ITSN1* under the *ptv* model (see **Table S3**). Variants were mapped to consequences based on their impact upon the Matched Annotation from NCBI and EMBL-EBI (MANE) transcript for the gene (ENST00000381318).

## All of Us dataset

The All of Us Research Program aims to obtain genetic and health data from one million or more people living in the United States to accelerate biomedical research and improve human health. So far, more than 740,000 people have joined the program. At the time that this study was conducted, All of Us contained short-read WGS data from 245,376 individuals, 123,063 of whom were classified as having European genetic ancestry (see below).

### *Sequencing and variant calling*

Genome sequencing, variant discovery, sample QC and variant QC of All of Us data are explained in the recent study by the All of Us research program genomics investigators.<sup>18</sup> The materials explained here are obtained from that study and All of Us genomic quality report (<https://support.researchallofus.org/hc/en-us/articles/4617899955092-All-of-Us-Genomic-Quality-Report>).

To ensure accurate and consistent DNA sample extraction and sequencing, the All of Us Genome Centers and Biobank implemented standardized laboratory protocols, quality control methods, and validation experiments with known clinical samples and reference standards. PCR-free Barcoded WGS libraries were created using the Illumina Kapa HyperPrep kit. The libraries were then pooled and sequenced on the Illumina NovaSeq 6000 instrument. Following demultiplexing, initial QC analysis is conducted using the Illumina DRAGEN pipeline, which includes metrics at the lane, library, flow cell, barcode, and sample levels. This analysis also evaluates contamination, mapping quality, and concordance with genotyping array data processed separately from a different DNA aliquot. The Genome Centers utilize these metrics to assess if each sample meets program specifications. Once approved, the sequencing data is then sent to the Data and Research Center for additional quality control, joint calling, and distribution to the research community.

### *Sample QC*

All of Us samples passing the following quality control metrics were used in joint calling and released to the research community. The quality control criteria include Mean coverage (threshold  $\geq 30x$ ), Genome coverage (threshold  $\geq 90\%$  at  $20x$ , All of Us Hereditary Disease Risk gene (AoUHDR) coverage (threshold  $\geq 95\%$  at  $20x$ ), Aligned Q30 bases (threshold  $\geq 8 \times 10^{10}$ ) and contamination  $\leq 1\%$ . Additional QC checks include fingerprint mismatch and array mismatch.

Ancestry was inferred across populations defined in gnomAD using a random forest classifier trained using a diverse reference dataset of 3,202 samples and 151,159 autosomal single-nucleotide polymorphisms. By projecting the All of Us samples onto the classifier's 16-dimensional PCA space, categorical ancestry predictions were generated. Since data from AoU were used as a non-neurological control cohort for analysis together with cases of European ancestry from AMP-PD, we restricted to those classified as having European ancestry. The dataset was then restricted to include only those with corresponding sex at birth and genetically-inferred sex; 119,089 samples remained.

AoU used the Hail `pc_relate` function to determine the kinship score to report any pairs with a kinship score  $> 0.1$ . The kinship score is half of the fraction of the genetic material shared

(ranges from 0.0 - 0.5); parent-child or siblings: 0.25, Identical twins: 0.5. We identified pairs of related individuals in the remaining cohort using the relatedness.tsv file provided by AoU, which lists all pairs of samples with a kinship score above 0.2. The first sample of each related pair was removed, leaving 114,415 unrelated individuals. A final dataset of 50,754 people remained after removing samples with any diagnosis under the ICD-10 neurological disease chapter (**Table 2**).

### *Gene-level analysis*

AoU applied the following site-level quality control metrics on variants:  $GQ \geq 20$ ,  $DP \geq 10$ ,  $AB \geq 0.2$  for heterozygotes,  $ExcessHet < 54.69$  and  $QUAL \geq 60$  for SNPs and  $QUAL \geq 69$  for Indels. We restricted to variants in positions overlapping CCDS transcripts (release 22). Analysis was performed based on gene-level aggregation of variants with qualifying effects upon *ITSN1* based on the *ptv* model (see **Table S3**). Variants were annotated using SnpEff v4.3 and included on the basis of consequences for any transcript of *ITSN1*.

### **Power analysis for *ITSN1* replication**

Power analysis was performed to assess the sample requirements for 80% power to replicate the association between PD and *ITSN1* PTVs at  $p=0.05$ . Assuming that qualifying *ITSN1* PTVs occur in 0.0003 of people (**Table S8, Table S9**), we estimated that a total of ~17,500 cases and controls (at 1:1 sampling) would be required to detect an association with an odds ratio of 10, and ~25,000 for an odds ratio of 6 (**Fig. S8**).



## Supplementary Table Legends

### Table S1 - UK Biobank ancestry cohorts and sex-rebalancing statistics.

Two-sided Fisher's Exact Tests were used to compare the difference in odds of being male versus female across cases and control cohorts of UK Biobank. In cohorts with a  $p < 0.05$ , we downsampled the sex overrepresented in controls. The column 'N controls (male: female)' indicates the control cohort sample size before downsampling, and the column 'N rebalanced controls (male: female)' indicates control cohort sample size after downsampling, where performed.

### Table S2 - Variant-level association statistics across all genetic models for variants with associations with $p < 1 \times 10^{-4}$

Association statistics for allelic/dominant/recessive genetic models are shown for all variants, which are aligned to GRCh38 and indicated in the 'variantID' column in the format CHR-POS-REF-ALT (chromosome, position, reference allele, alternate allele). The columns 'Gene\_Symbol', 'hgnc\_id', and 'ensembl\_gene\_id' label the gene for which each variant is identified as most damaging using SnpEff (v4.3). Rows are sorted by p-value from smallest to largest, with corresponding odds ratios and 95% confidence intervals. Minor allele frequencies (MAF) are given overall and for cases and controls individually. Variants with overall MAF  $\geq 0.001$  were clumped into independent loci based on the most significant genetic model for each variant; those with MAF  $< 0.001$  were not entered into clumping; The 'clumping\_info' column labels variants as 'index', 'clumped', or 'LowMAF\_not\_clumped'; For index variants, the 'clumping\_nClumped' column indicates the number of variants clumped within that column. For clumped variants, the 'clumping\_clumpedIntoVariant' column indicates the associated index variant; '†' marks in the 'annotation' column flag variants with associations at  $p < 1 \times 10^{-6}$  that are situated near a large chromosome 17 inversion (GRCh38: chr17:45,495,836-46,707,123)<sup>19</sup> and which may be subject to long-spanning linkage disequilibrium not fully captured during clumping. The "Gene\_known\_in" column labels variants mapped to genes previously implicated in Parkinson's disease, Alzheimer's disease, or amyotrophic lateral sclerosis (ALS). The genotype counts for reference allele 'p' and alternate allele 'q' are given in the columns 'caseGenotypes' and 'controlGenotypes', using the format 'pp/pq/qq' for autosomal and pseudo-autosomal variants and 'pp/pq/qq/p/q' for X-chromosome variants not in the pseudo-autosomal region, representing hemizygous male genotypes separately. The columns 'caseRefCount', 'caseAltCount', 'ctrlRefCount', 'ctrlAltCount' indicate the cell counts for the contingency table analyzed using Fisher's Exact Test; counts are derived from the genotypes indicated in 'caseGenotypes' and 'controlGenotypes' and depend upon to the tested genetic model (see Methods).

### Table S3 - Qualifying variant models.

\* reflects the gnomAD global\_raw minor allele frequency (MAF) unless otherwise specified

^ reflects the maximum percentage of UKB genome sequences permitted to either have  $\leq 10x$

coverage at variant site or carry a low-confidence variant failing at least one of the quality-control thresholds applied for collapsing analysis. The threshold is set with varying stringency for the large-scale European ancestry (EUR) cohort according to the Internal MAF threshold for each model; a default threshold of  $\leq 0.005\%$  is set across all models for the smaller African (AFR), East Asian (EAS), and South Asian (SAS) ancestry cohorts.

† For some models, a more lenient Internal MAF threshold is used in replication analysis because the datasets are substantially smaller than UKB and the analysis specifically targets genes with evidence of association to PD from the initial discovery analyses.

∅ reflects the probability score for a variant from gnomAD's random forest filter; higher probability scores indicate higher confidence of a true allele.

Ω Variant consequences:

Synonymous: `synonymous_variant`.

PTV: `exon_loss_variant`, `frameshift_variant`, `start_lost`, `stop_gained`, `stop_lost`, `splice_acceptor_variant`, `splice_donor_variant`, `gene_fusion`, `bidirectional_gene_fusion`, `rare_amino_acid_variant`, `transcript_ablation`.

Missense: `missense_variant` & `splice_region_variant`, `missense_variant`.

Non-synonymous: `exon_loss_variant`, `frameshift_variant`, `start_lost`, `stop_gained`, `stop_lost`, `splice_acceptor_variant`, `splice_donor_variant`, `gene_fusion`, `bidirectional_gene_fusion`, `rare_amino_acid_variant`, `transcript_ablation`, `conservative_inframe_deletion`, `conservative_inframe_insertion`, `disruptive_inframe_insertion`, `disruptive_inframe_deletion`, `missense_variant` & `splice_region_variant`, `missense_variant`, `protein_altering_variant`

UKB = UK Biobank; 100kGP = 100,000 Genomes project

**Table S4 - Gene-level association statistics in UK Biobank European ancestry cohort across all models for genes with associations at  $p < 1 \times 10^{-4}$ .**

Rows are sorted by p-value from smallest to largest. The 'Gene\_known\_in' column labels whether each gene has been previously implicated in ALS, Alzheimer's disease, or Parkinson's disease.

**Table S5. Breakdown of *GBAI* variants included within the *flexnonsynmtr* collapsing model analysis of UK Biobank Europeans.**

Variants were annotated to consequences against the MANE *GBAI* transcript (ENST00000368373) using SnpEff v4.3t against genome build GRCH38.92. Two-tailed Fisher's exact tests were performed to identify variants most influential upon the *GBAI* gene *flexnonsynmtr* model association. The "Abs\_log10\_pValueChange" column indicates the magnitude of difference from the main association result ( $p = 2.41 \times 10^{-9}$ ) after removing a given variant, letting Abs\_log10\_pValueChange=1 indicate 1 order of magnitude difference from the main result; rows of the table are sorted by Abs\_log10\_pValueChange from largest to smallest.

The columns “n\_UKBEUR\_cases” and “n\_UKBEUR\_ctrls” respectively indicate the number of cases and controls harboring each variant.

**Table S6 - Gene-level association statistics in UK Biobank Pan-ancestry cohort across all models for genes with associations at  $p < 1 \times 10^{-4}$ .**

Rows are sorted by p-value from smallest to largest. The 'Gene\_known\_in' column labels whether each gene has been previously implicated in ALS, Alzheimer's disease, or Parkinson's disease. Counts for "n controls without QV", "n controls with QV", "n cases without QV" and "n cases with QV" are provided for each of the ancestry groups analyzed: African (AFR), East Asian (EAS), European (EUR), and South Asian (SAS). QV = variant qualifying for collapsing model.

**Table S7 - Breakdown of *ITSN1* variants present in each cohort.**

Variants are annotated to consequences against the MANE *ITSN1* transcript (ENST00000381318) and, for variants without a protein-truncating effect upon ENST00000381318, the protein-coding transcript for which they are most damaging. Annotation was performed using SnpEff v4.3t against genome build GRCH38.92. Rows of the table are sorted by variant position.

UKBEUR = UK Biobank European-ancestry cohort; 100kGP = 100,000 Genomes Project; AMP-PD = Accelerating Medicines Partnership - Parkinson's Disease; AoU = All of Us.

**Table S8 - Cumulative frequency of *ITSN1* PTVs qualifying for *ptv* model across PD and control cohorts.**

**Table S9 - Summary of *ITSN1* variants present in gnomAD (v2.1.1 exomes) that pass *ptv* collapsing model filters.**

The cumulative variant frequency in Non-Finnish Europeans is 0.000313, the sum of AC\_nfe/AN\_nfe for each variant. For the non-Finnish European non-neurological control cohort the frequency is 0.000214, a sum of non\_neuro\_AC\_nfe/non\_neuro\_AN\_nfe.

**Table S10 - Transcript-specific analysis of *ITSN1 ptv* gene-level collapsing model in the UK Biobank European ancestry cohort.**

Variants qualifying for the main transcript-agnostic *ITSN1 ptv* model analysis were reannotated against specific *ITSN1* transcripts using SnpEff v4.3t against genome build GRCH38.92. Variants were included in the analysis when they had a PTV effect upon that transcript (Table S3).

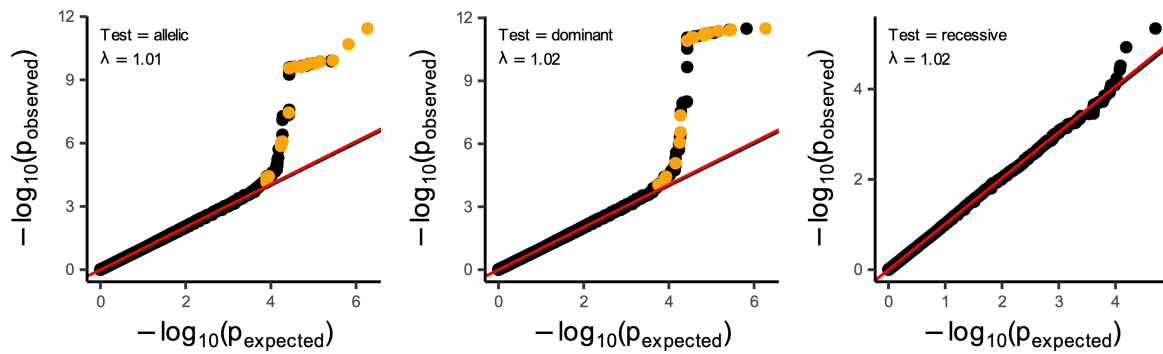
**Table S11 – Mantis-ML v2 results.**

Prioritization of Human Phenotype Ontology terms most likely to be associated with *ITSN1* by Mantis-ML v2

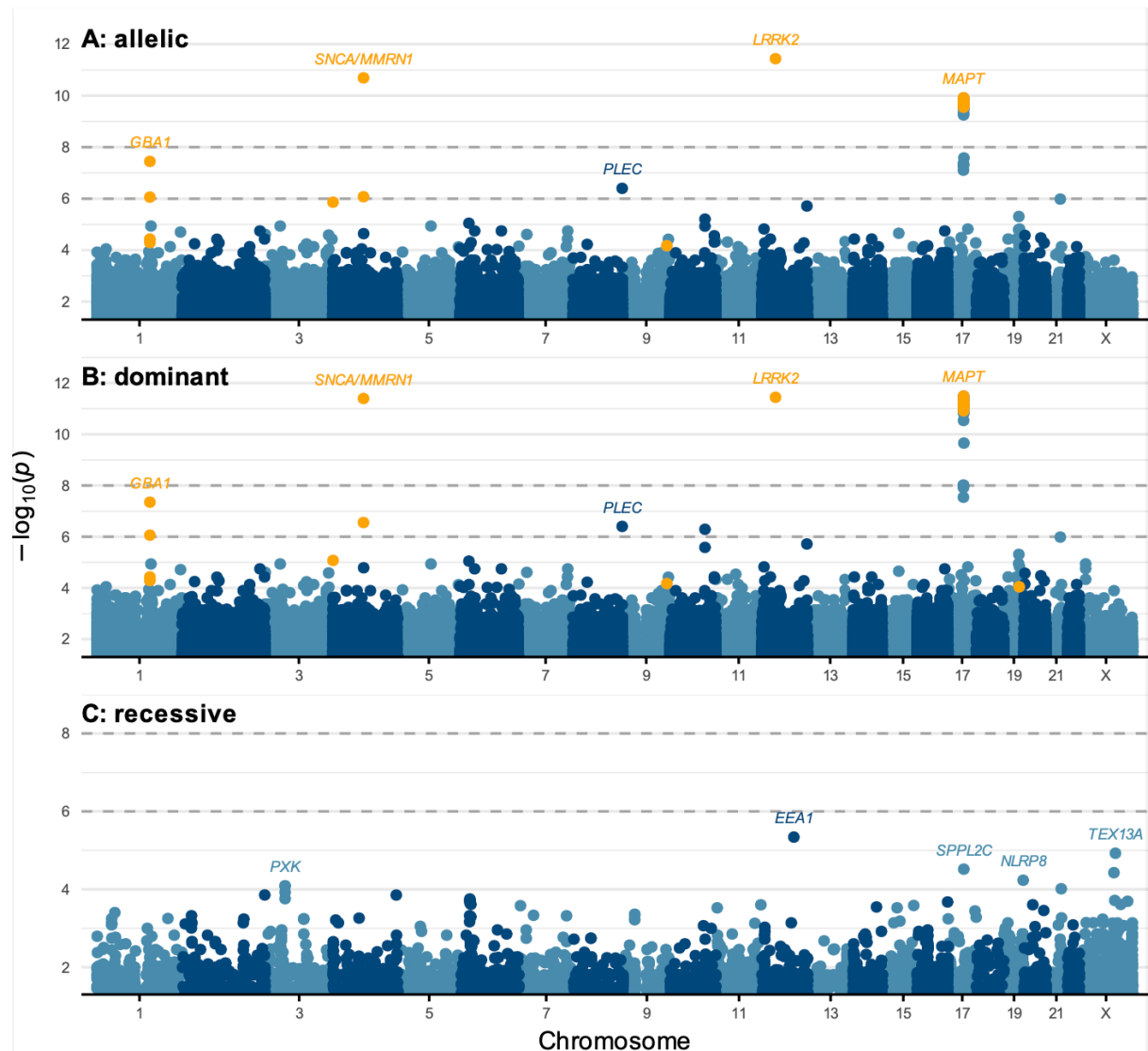
**Table S12 – *Drosophila* genotypes.**

Complete *Drosophila* genotypes for animals used in figure 4.

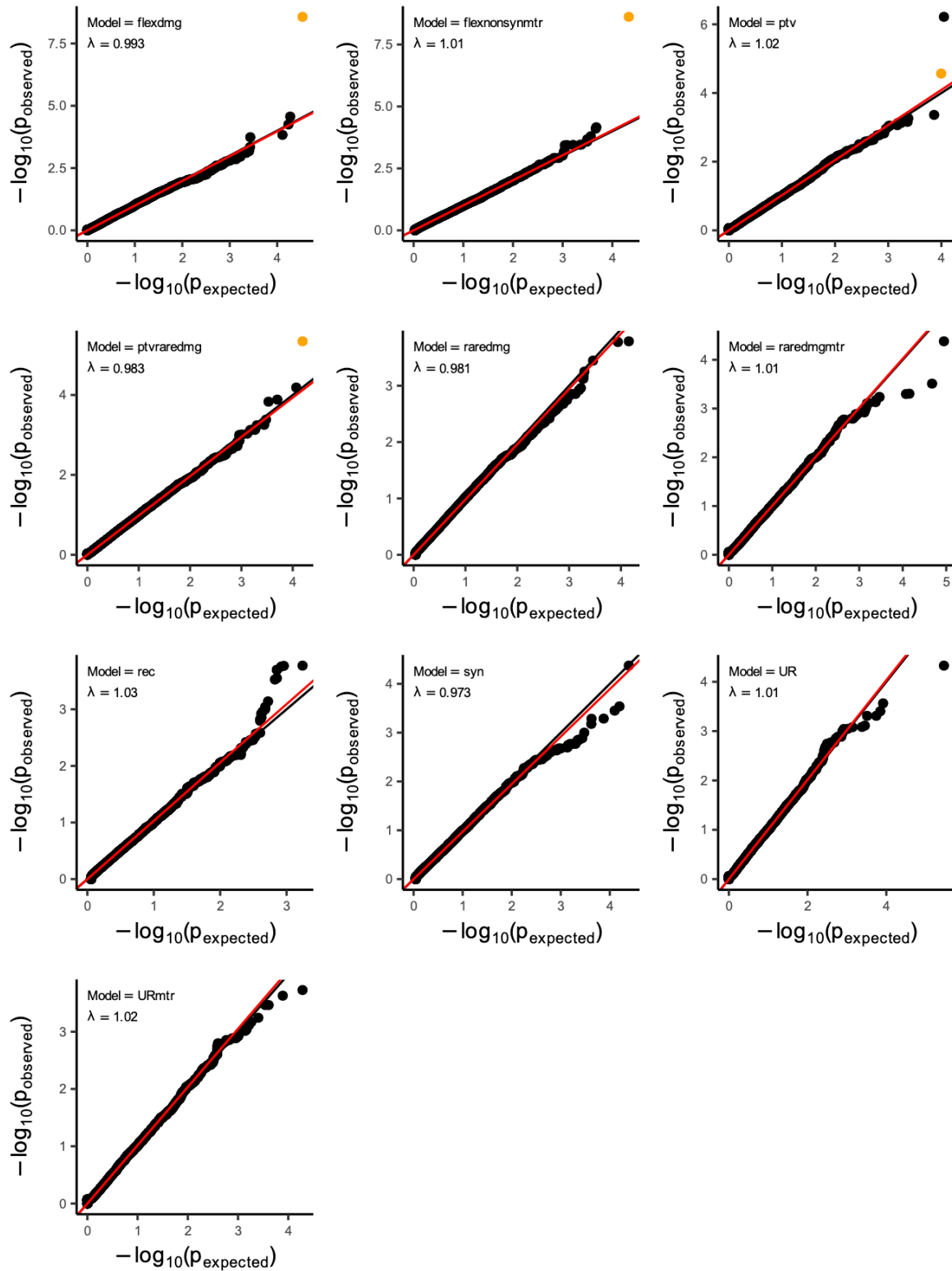
## Supplementary Figures



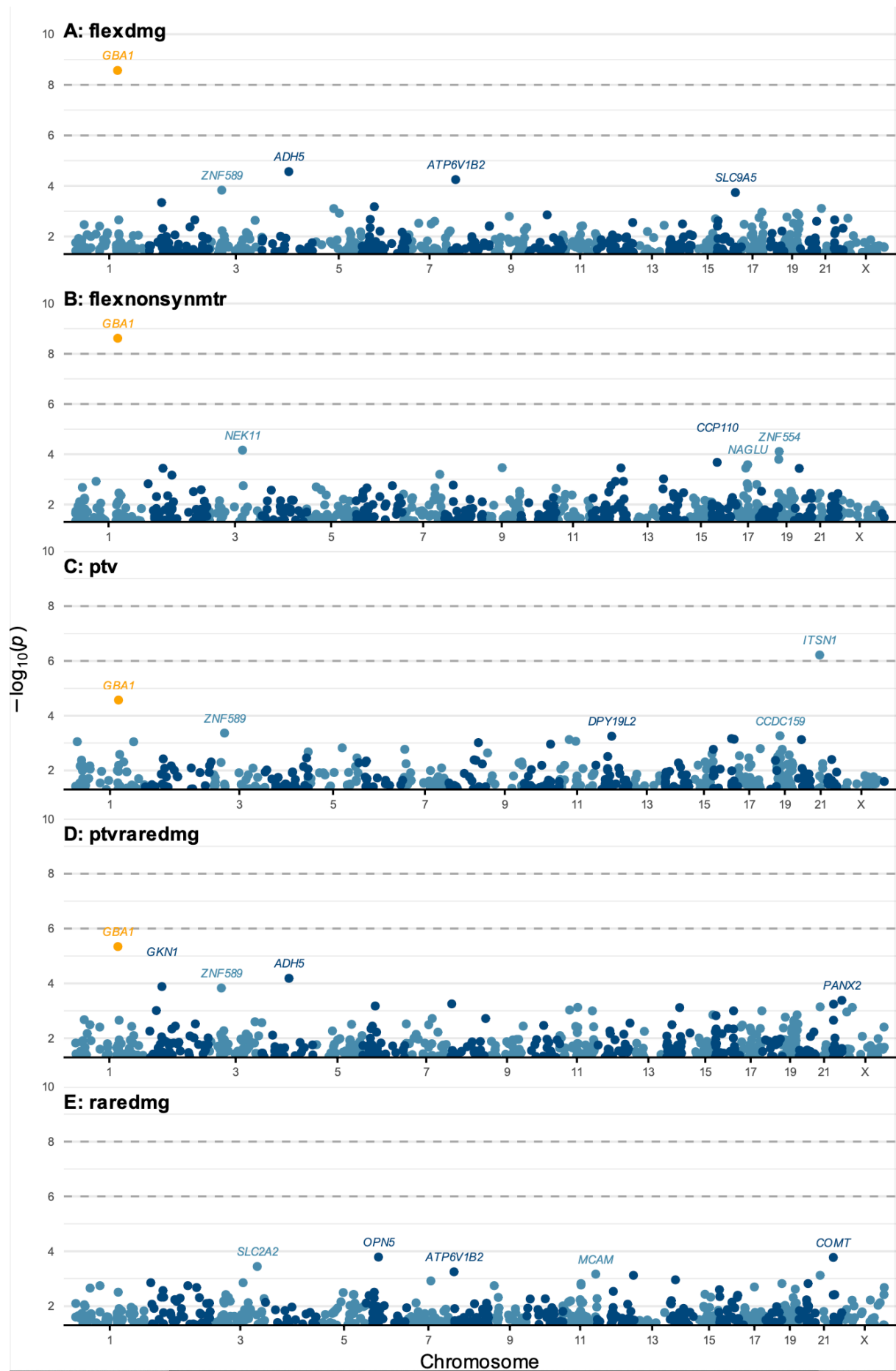
**Figure S1. QQ plots for observed and expected  $-\log_{10}(p)$  values in the variant-level analysis.** Orange point shading labels genes that have previously been associated to Parkinson's disease and have associations at  $p < 1 \times 10^{-4}$ . The null-distribution of expected p-values is defined based on n-of-1 permutation of phenotype group for each genetic test. The lambda ( $\lambda$ ) genomic inflation factor was calculated under a regression-based approach.



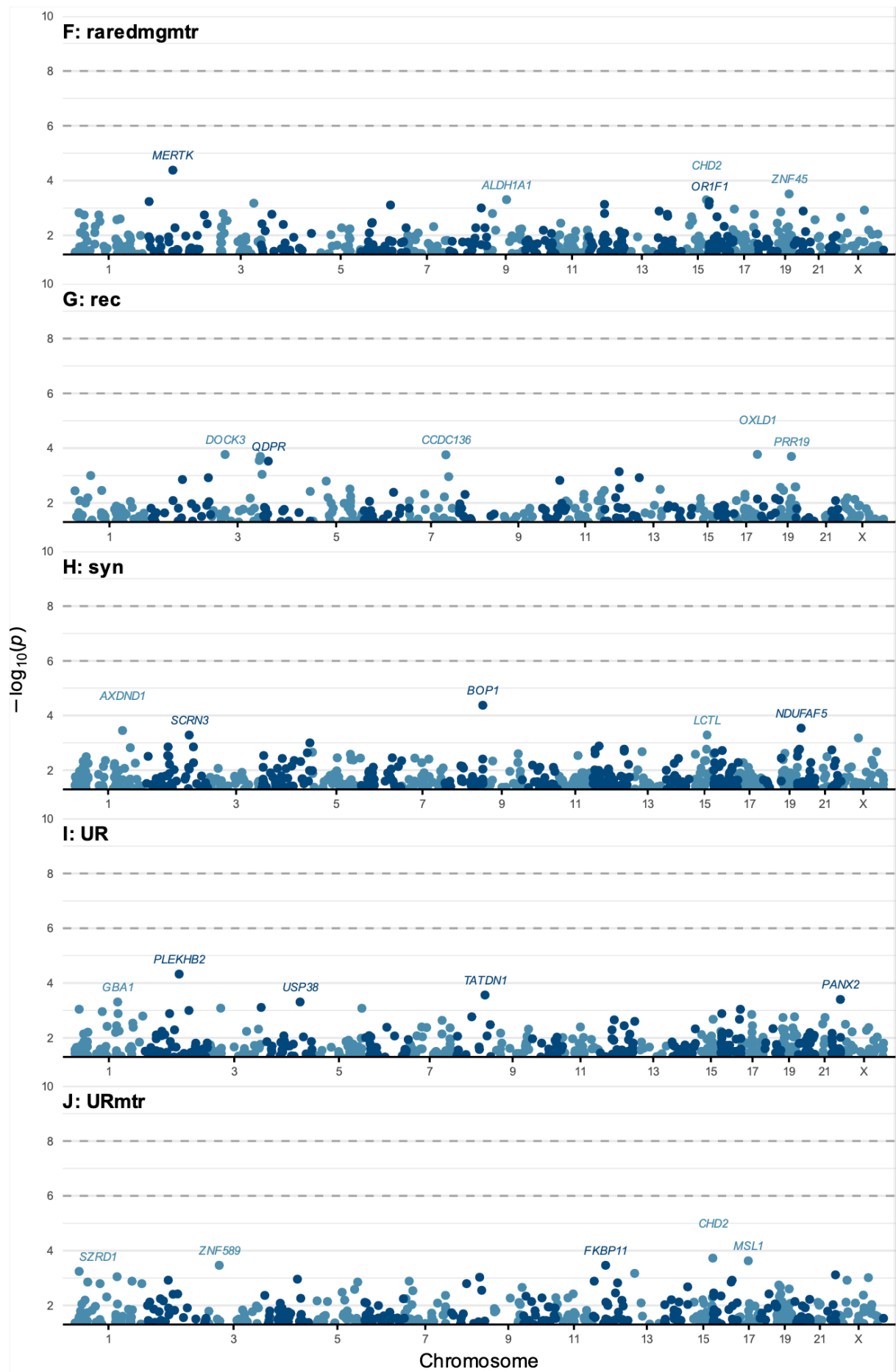
**Figure S2. Manhattan plots of variant-level associations to Parkinson's disease under each genetic model.** Orange point shading labels variants annotated against genes that have previously been associated to PD and have associations at  $p < 1 \times 10^{-4}$ ; Horizontal dashed lines indicate the significance ( $p < 1 \times 10^{-8}$ ) and suggestive ( $p < 1 \times 10^{-6}$ ) thresholds for the analyses.



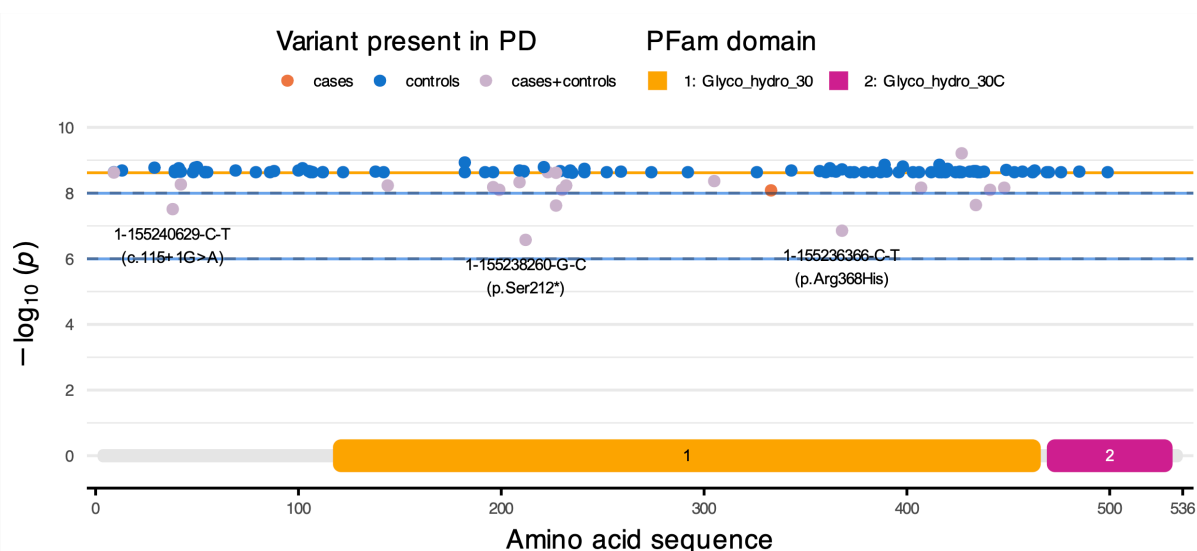
**Figure S3. QQ plots for observed and expected  $-\log_{10}(p)$  values in the gene-level analysis.** Orange point shading labels variants mapped to genes that have previously been associated to Parkinson's disease and have associations at  $p < 1 \times 10^{-4}$ . The null-distribution of expected p-values is defined based on n-of-1 permutation of phenotype group for each genetic test. The lambda ( $\lambda$ ) genomic inflation factor was calculated under a regression-based approach.



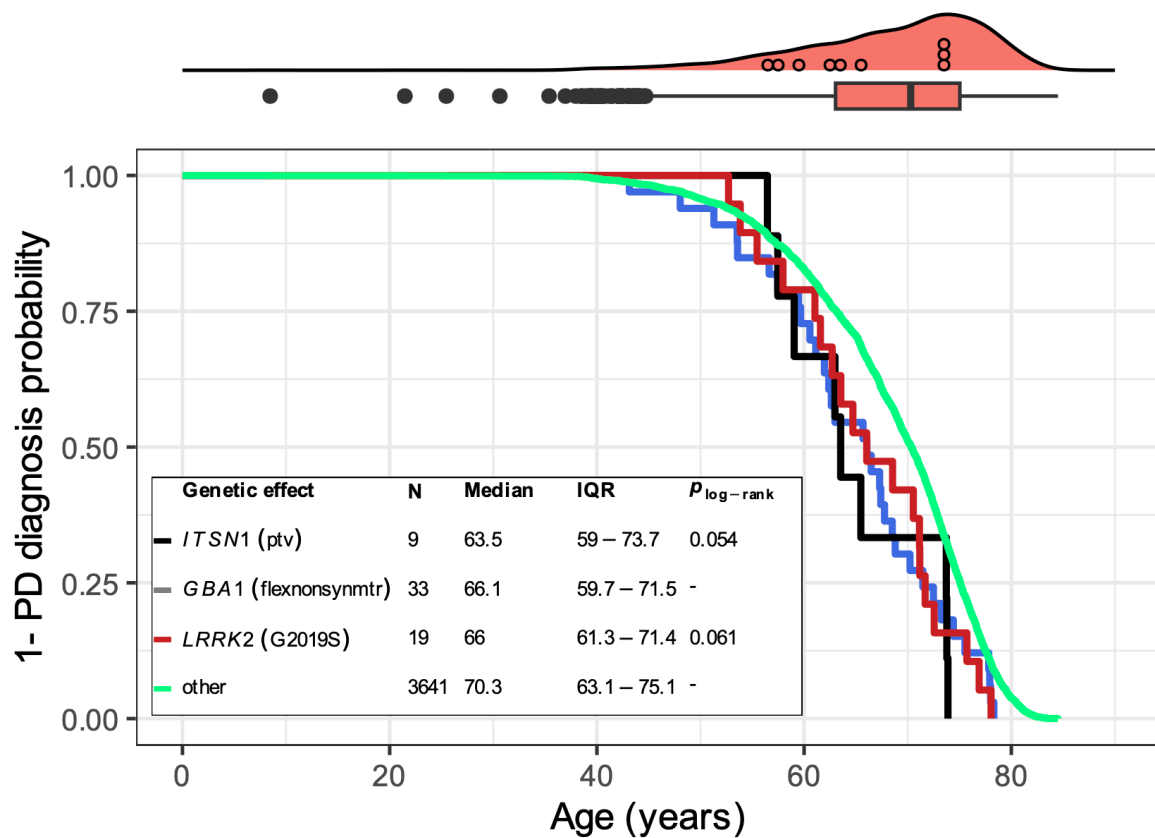




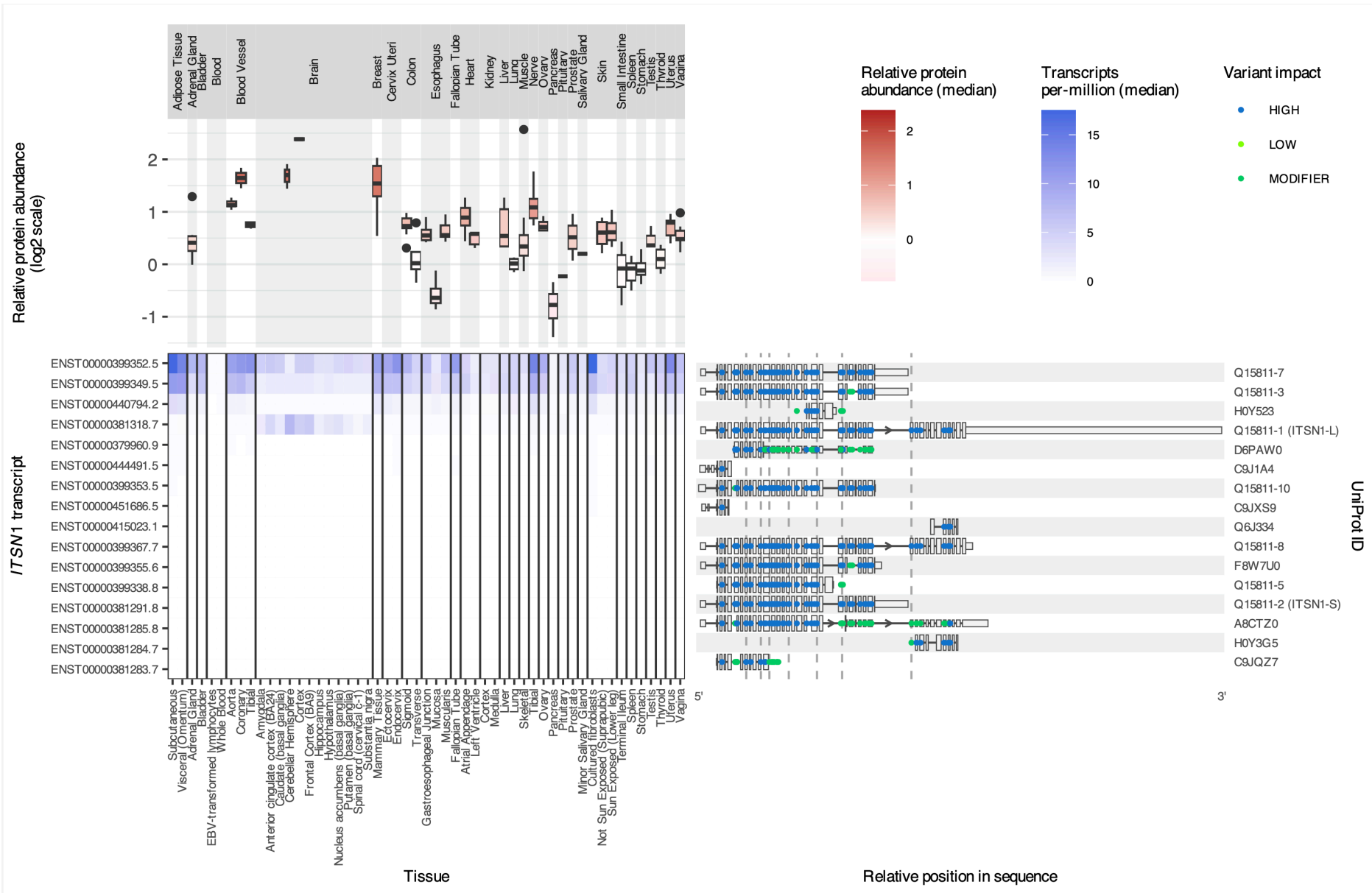
**Figure S4. Manhattan plots of gene-level associations to Parkinson's disease under each collapsing model.** Orange point shading labels genes that have previously been associated to PD and have associations at  $p < 1 \times 10^{-4}$ ; Horizontal dashed lines indicate the significance ( $p < 1 \times 10^{-8}$ ) and suggestive ( $p < 1 \times 10^{-6}$ ) thresholds for the analyses.



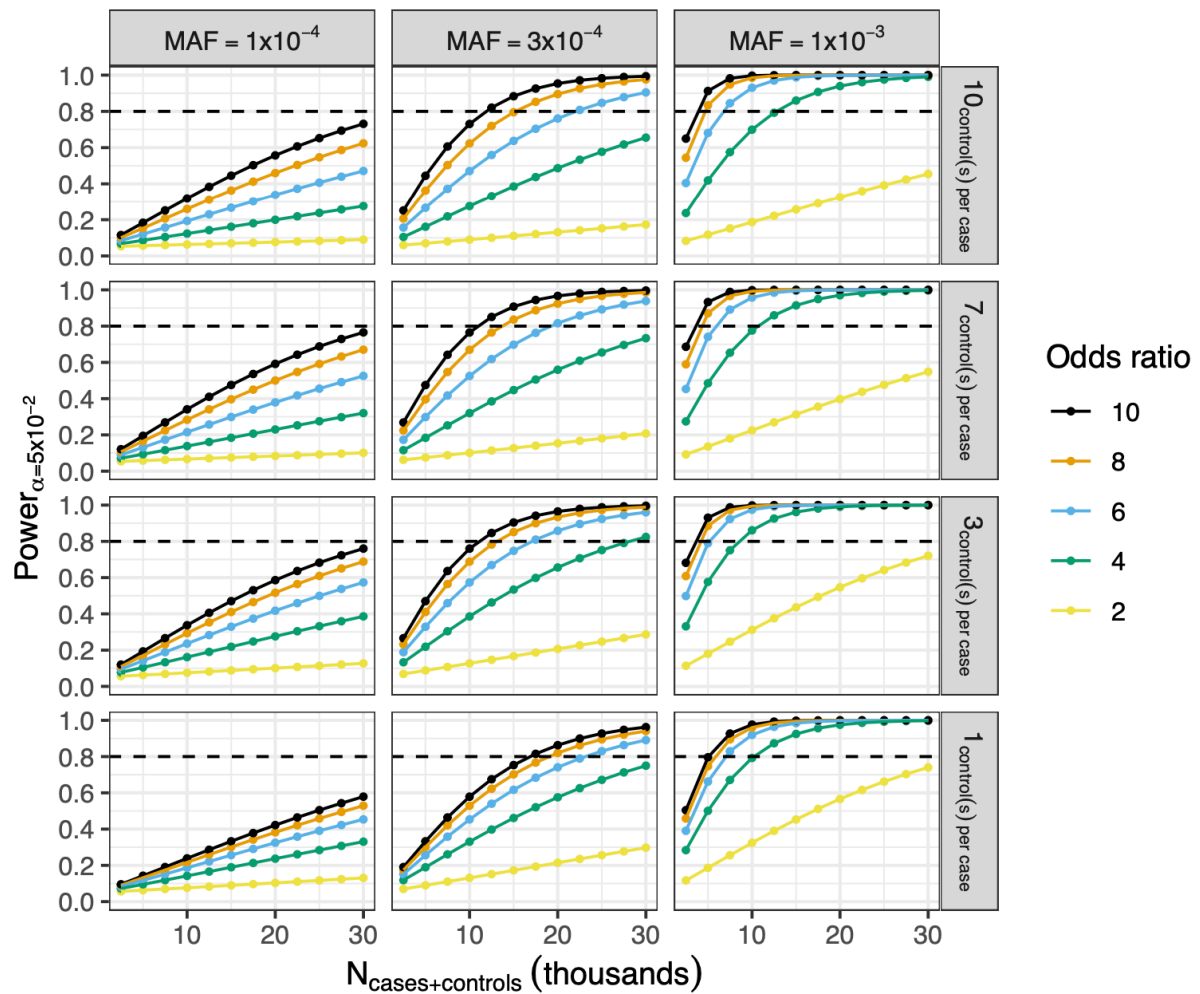
**Figure S5. Gene-level association results in UK Biobank European ancestry cohort 'leave-one-out' analyses.** Two-tailed Fisher's exact tests were performed for the *GBA1* gene under the *flexnonsynmtr* collapsing model after removal of a given qualifying variant. x-axis positions of points indicate the location of the removed variant relative to the amino acid sequence of the *GBA1* MANE transcript (ENST00000368373) and PFam domains of the corresponding protein (P04062). The orange horizontal line on each panel indicates the association result from the main analysis ( $p=2.41 \times 10^{-9}$ ). Greater deviation from this line indicates greater influence of the removed variant upon the association result. Variants with  $>1$  absolute difference in  $-\log_{10}(p)$  between the main and leave-one-out analysis results are labelled in the format 'variant ID (MANE transcript consequence: protein impact [or coding sequence impact for splice variants])'. Gray hatched lines indicate significance ( $p < 1 \times 10^{-8}$ ) and suggestive ( $p < 1 \times 10^{-6}$ ) thresholds for the study (Methods).



**Figure S6. Trends in age of Parkinson's disease diagnosis according to genetic susceptibility.** The density curve at the top of the figure and corresponding boxplot show the distribution in age of PD diagnosis across the UK Biobank European-ancestry cohort; black circular points within the density curve show age of diagnosis among people with a qualifying *ITSN1* PTV, binned by year. The Kaplan-Meier curve shows time from birth until age of PD diagnosis, stratifying by genetic effect; Trends in *ITSN1* and *GBA1* are shown for people harboring qualifying variants captured in the gene-level collapsing model most significant for each gene (see **Table 3**). Trends in *LRRK2* are shown for carriers of the well-established p.Gly2019Ser (12-40340400-G-A) variant; The 'other' group comprises the remaining PD cohort, without these genetic burdens. The inset table provides descriptive statistics for each group, and the ' $p_{\log\text{-rank}}$ ' values are for a log-rank test of the difference in survival for each group relative to the 'other' group; non-significant test results are not surprising since specific genetic effects occur in so few people.



**Figure S7. *ITSN1* PTVs impact multiple protein-coding transcripts.** The heatmap displays expression of different protein-coding *ITSN1* transcripts across tissue samples from GTEx, ordered from most to least highly expressed. The panel to the right illustrates the sequence of each transcript and colored points show the location and impact of each *ITSN1* variant from the *ptv* model collapsing analysis against each transcript; HIGH impact indicates a protein-truncating effect; The positions of qualifying PTVs present in people with Parkinson's disease are marked by vertical hatched lines across this panel; intronic sequences that do not overlap exons of another transcript have been shortened to better visualize exons; bars for 5' and 3' untranslated regions are half the height of bars showing the coding sequence. The boxplots above the heatmap display estimates of relative *ITSN1* protein abundance across 32 tissue sites in a subset of the GTEx cohort.<sup>20</sup>



**Figure S8. Analysis of power to detect a dominant genetic association for a rare variant using a dominant test.** Plot points represent the estimated power for a given N (cases+controls) at the effect size (Odds ratio) indicated by color. Figure panels are split column-wise according to the minor allele frequency (MAF) for the variant tested (which can be approximately interpreted as the cumulative frequency at which people have a qualifying variant in a gene-level collapsing analysis) and row-wise by the number of controls sampled per-case. Analysis was performed with the *genpwr* R package (v1.0.4).<sup>21</sup>

## References

- 1 Li S, Carss KJ, Halldorsson BV, Cortes A, UK Biobank Whole-Genome Sequencing Consortium. Whole-genome sequencing of half-a-million UK Biobank participants. *medRxiv* 2023; published online Dec 8. DOI:10.1101/2023.12.06.23299426.
- 2 Halldorsson BV, Eggertsson HP, Moore KHS, *et al.* The sequences of 150,119 genomes in the UK Biobank. *Nature* 2022; **607**: 732–40.
- 3 Cingolani P, Platts A, Wang LL, *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* 2012; **6**: 80–92.
- 4 Howe KL, Achuthan P, Allen J, *et al.* Ensembl 2021. *Nucleic Acids Res* 2021; **49**: D884–91.
- 5 Wang Q, Dhindsa RS, Carss K, *et al.* Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* 2021; **597**: 527–32.
- 6 Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. *Bioinformatics* 2010; **26**: 2867–73.
- 7 Jun G, Flickinger M, Hetrick KN, *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* 2012; **91**: 839–48.
- 8 Pujar S, O’Leary NA, Farrell CM, *et al.* Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. *Nucleic Acids Res* 2018; **46**: D221–8.
- 9 Hanscombe KB, Coleman JRI, Traylor M, Lewis CM. ukbtools: An R package to manage and query UK Biobank data. *PLOS ONE* 2019; **14**: e0214311.
- 10 Auton A, Brooks LD, Durbin RM, *et al.* A global reference for human genetic variation. *Nature* 2015; **526**: 68–74.
- 11 Pedersen BS, Quinlan AR. Who’s Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. *Am J Hum Genet* 2017; **100**: 406–13.
- 12 Karczewski KJ, Francioli LC, Tiao G, *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020; **581**: 434–43.
- 13 McLaren W, Gil L, Hunt SE, *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* 2016; **17**: 122.
- 14 Iwaki H, Leonard HL, Makarios MB, *et al.* Accelerating Medicines Partnership: Parkinson’s Disease. Genetic Resource. *Mov Disord* 2021; **36**: 1795–804.
- 15 Nalls MA, Bras J, Hernandez DG, *et al.* NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases. *Neurobol Aging* 2015; **36**: 1605.e7-1605.e12.



- 16 Altshuler DM, Gibbs RA, Peltonen L, *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* 2010; **467**: 52–8.
- 17 Rimmer A, Phan H, Mathieson I, *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* 2014; **46**: 912–8.
- 18 Bick AG, Metcalf GA, Mayo KR, *et al.* Genomic data in the All of Us Research Program. *Nature* 2024; **627**: 340–6.
- 19 Campoy E, Puig M, Yakymenko I, Lerga-Jaso J, Cáceres M. Genomic architecture and functional effects of potential human inversion supergenes. *Philos Trans R Soc Lond B Biol Sci* 2022; **377**: 20210209.
- 20 Jiang L, Wang M, Lin S, *et al.* A Quantitative Proteome Map of the Human Body. *Cell* 2020; **183**: 269-283.e19.
- 21 Moore CM, Jacobson SA, Fingerlin TE. Power and Sample Size Calculations for Genetic Association Studies in the Presence of Genetic Model Misspecification. *Hum Hered* 2020; **84**: 256–71.