Supplementary Information for

Coding mutations in *NUS1* contribute to Parkinson's disease

Ji-feng Guo, Lu Zhang, Kai Li, Jun-pu Mei, Jin Xue, Jia Chen, Xia Tang, Lu Shen, Hong Jiang, Chao Chen, Hui Guo, Xue-li Wu, Si-long Sun, Qian Xu, Qi-ying Sun, Piu Chan, Hui-fang Shang, Tao Wang, Guo-hua Zhao, Jing-yu Liu, Xue-feng Xie, Yi-qi Jiang, Zhen-hua Liu,Yu-wen Zhao, Zuo-bin Zhu, Jia-da Li, Zheng-mao Hu, Xin-xiang Yan, Xiao-dong Fang, Guang-hui Wang, Feng-yu Zhang, Kun Xia, Chun-yu Liu, Xiong-wei Zhu, Zhen-yu Yue, Shuai Cheng Li, Huai-bin Cai, Zhuo-hua Zhang, Ran-hui Duan & Bei-sha Tang*

*Correspondence author. Bei-sha Tang, Email: bstang7398@163.com

This files includes:

Supplementary Information Text

Materials and Methods

Human Subjects

Thirty-nine EOPD families of Han Chinese were recruited for exome sequencing, including 19 trios and 20 quads. The probands have the typical PD phenotypes, and their age at onset are less than 35 years. All the 39 families are from non-consanguineous unions. We ruled out the known genetic and environmental factors of PD, such as known genetic mutations and traumatic brain injury, in all 39 probands using multiple genetic and clinical examinations in our previous work (1-6). None of the first- and second-degree relatives of these 39 probands has PD or Parkinsonism. All the patients and their enrolled family members were subjected to the standard clinical neurological examination. Idiopathic PD was diagnosed according to the United Kingdom PD Brain Bank Criteria(7) by at least two neurologists, and the other healthy family members did not have any nervous or psychiatric system diseases. One family was removed due to its exceeding Mendelian error of 17%. A total of 1,852 sporadic patients (average age at onset=48.57 ± 12.50 year; male=54.37%; and EOPD=45.36%) and 1,565 controls (average age=50.73 ± 16.68 year and male=51.37%) were collected for the first genetic replication. A total of 3,237 sporadic patients (average age at onset=58.05 ± 10.10 year; male=53.85%; and EOPD=9.36%) were collected for the second genetic replication. These subjects were recruited by the Department of Neurology of Xiangya Hospital, Xuanwu

Hospital, West China Hospital, Wuhan Union Hospital, The Second Affiliated Hospital of Zhejiang University School of Medicine and the State Key Laboratory of Medical Genetics of China. The protocol was approved by the Ethics Committee of Central South University, and written informed consent was collected from all the subjects.

Exome capture and sequencing

The genomic DNA for each individual was hybridized by the NimbleGen 2.1M-probe sequence capture array (http://www.nimblegen.com/products/seqcap/) to capture the exonic DNA. We performed whole exome sequencing with 90bp pair-end reads using the Illumina HiSeq 2000 platform. The raw image files were processed using the standard Illumina Pipeline (version 1.3.4) for base calling with the default parameters.

Detection and confirmation of *de novo* mutations

After removing the adapters, the raw reads in the FASTA format were aligned to the human reference genome (hg 19 version) by BWA (version 0.5.9-r16), and the PCR duplicates were marked by Picard (http://broadinstitute.github.io/picard/command-line-overview.html). We used the Genome Analysis Toolkit (GATK)(8) to perform the indel realignment, recalibrate the base quality score, and thereby obtained an ‘Analysis-Ready’ bam file for each individual. The SNVs and indels were jointly called by HaplotypeCaller in GATK for every three or four members per family. We further removed the mutations with a Variant Quality Score logs odds ratio with a tranche sensitivity of less than 99.9% to alleviate other confounding effects.

According to the definition of *de novo* mutation, we selected the heterozygous variants in the offspring and homozygous reference in both parents. We designed the following quality criteria to remove false positive *de novo* mutations: a) all genotype Phred quality scores must be greater than 30, b) only one type of alternative allele was allowed, c) the read coverage of alternative alleles in the offspring was required to be greater than 4, d) more than 30% and less than 5% of the covered reads should be alternative allele for the offspring and parents, e) for the offspring, we required that PL(0/0)≥30, PL(0/1)=0, and PL(1/1)≥30 (PL: Phred-scaled likelihoods for a given genotype), f) for both parents, we required that PL(0/0)=0, PL(0/1)≥30, and PL(1/1)≥30, g) two adjacent SNVs needed to be located at least 10 bp away, h) we removed indels in known structure variation regions, i) the *de novo* mutations were excluded if they were in dbSNP137, the Han Chinese of 1000 Genomes Project, or both of the two offspring in quads.

Sanger sequencing for both the mutation carriers and their parents validated these putative *de novo* mutations. The sequencing of each amplicon was performed with both forward and reverse primers. After revising the validated *de novo* mutations, we found that four criteria (c, d, e, and f) were crucial to achieve accurate results. Next, we relaxed these four criteria to rescue the missing *de novo* mutations due to the stringent parameters. For the offspring, a) the covered reads with alternative alleles required to be no less than 4,

b) the proportion of aligned reads with alternative alleles required to be greater than 25% for offspring, c) we required that PL(0/0)≥20, PL(0/1)=0, and PL(1/1)≥20. For the parents, we required that PL (0/0)=0, PL (0/1)≥20, and PL (1/1)≥20. We eliminated the mutations in intronic or intergenic regions based on the annotation of RefSeqGene (http://www.ncbi.nlm.nih.gov/refseq/rsg/). Four methods (SKIPPY(9), NetGene2(10), SplicePort(11) and Human Splicing Finder(12)) were applied to predict whether a *de novo* mutation could lead to transcript splicing. The *de novo* mutations were thought as the predicted splice sites if at least three of the abovementioned programs supported them.

To examine the effect of *de novo* mutations, we eliminated the influence of the disease susceptible inherited mutations that may lead to PD: a) rare (minor allele frequency <1% in dbSNP) homozygous or compound heterozygous mutations, b) rare heterozygous mutations on the maternal X chromosome and transmitted to the male proband, and c) rare deleterious variants (predicted by PolyPhen-2) inherited from one of the parents. We also collected the private inherited mutations (inherited from either one of the parents and observed in only one family) to compare with those *de novo* mutations. All the extracted inherited variants have a genotype Phred quality score greater than 20.

Brain-expressed genes and co-expression network in human brain

We used the microarray and RNA-Seq data from the Allen Brain Atlas (http://www.brain-map) and BrainSpan (http://www.brainspan.org/static/download.html), respectively. The expressed genes in specific regions of the human brain were defined by the log intensity >6 (microarray) or RPKM >5 (RNA-Seq). We considered the genes if they expressed in at least one of the SNc or STR regions for PD. We calculated the gene co-expression based on the average Pearson correlations of the 12 new candidate genes from the brain developmental expression data in BrainSpan and its statistical significance was evaluated by 100,000 random simulations, each of which contained 12 genes.

Prediction of microRNA targets

The predicted microRNA targets with a good mirSVR were obtained from the microrna.org(13), and we applied a hypergeometric test to evaluate the co-targets of the 12 candidate genes and PD known causal genes by the same microRNAs. We performed 10,000 random simulations to calculate the corrected P values and evaluate the empirical distribution of the selected genes as the targets of hsa-miR-125a-3p. Twelve genes were randomly selected for each simulation, and the number genes were targeted by hsa-miR-125a-3p calculated.

Protein-Protein interaction networks

The protein-protein interaction networks were constructed for both the 12 new candidate genes and PD known causative genes based on DAPPLE (Disease Association Protein-Protein Link Evaluator)(14). We further explored the differential expression of the genes involved in the protein-protein interaction networks by GEO2R (http://www.ncbi.nlm.nih.gov/geo/geo2r/) on the collected gene expression data from PD known genetic mouse models, MPTP-treated mouse models, and control mice from GEO. The smallest P value from the selected gene expression datasets is presented (**Table S6**).

Gene ontology and KEGG pathway enrichment

We annotated the 12 new candidate genes based on Gene ontology (GO) (http://www.geneotology.org), the KEGG pathway database (http://www.genome.jp/kegg/pathway.html), and calculated their functional enrichment by a hypergeometric test. The P values for the enrichment of GO and KEGG were corrected by the Bonferroni correction and False discovery rate, respectively. The nonsynonymous *de novo* mutations in the siblings were used to calculate the enrichment of GO and KEGG.

MIPs design and procedure

MIPgen (https://github.com/shendurelab/MIPGEN)(15) designed MIPs. All the designed MIPs for the candidate genes exons are provided in the **Dataset S5**. Multiplex capture, amplification procedure and high-throughput sequencing data analysis followed the protocol proposed by Nuttle et al.(16, 17). The PCR products were sequenced by Illumina HiSeq 3000 with 150bp paired-end reads.

Statistical analysis for case-control replications

The paired-end reads sequenced from MIPs and exome sequencing were aligned to hg19 human reference genome and followed by variants calling with GATK. We kept the high confident candidate variants (genotype quality≥20, sequencing depth≥6X and the proportion of the reads with alternative alleles≥0.3) for further association analysis. The association of single variants was evaluated by Fisher's exact test in PLINK(18) after removing the variants that satisfied the thresholds in Hardy Weinberg disequilibrium ($P<10^{-4}$), minor allele frequency (<0.01), and genotype missing rate (>0.05). The enrichment of rare nonsynonymous variants (minor allele count ≤3 in controls) in a given gene was calculated by Fisher's exact test. We assumed 100 rare variants (0.05%-1% for minor allele frequency) were involved in the candidate genes and required gene-based p-value surpassing 0.004 (0.05/12). The power reached 99.97% for 5,089 patients by using non-central chi-square approximation in KATSP (19)

Real-time PCR and reverse transcription-PCR

PBMCs were isolated from EDTA blood by density gradient centrifugation (Histopaque, Sigma-Aldrich). Total RNA isolated using Trizol reagent (Invitrogen, 15596–018) was converted to cDNA by the Verso™ cDNA Kit (Thermo Scientific, AB1453B) following the manufacturer's instruction. SYBR Green qPCR Master Mix (2x) (Thermo Scientific, #K0251) was used for quantitative real-time PCR amplification using a CFX96 Real-Time PCR Detection System (BiO-RAD) and corresponding software (Applied Biosystems, Foster City, USA). Primers for *NUS1* were 5’- AGCCTCGTGGTGTGGTGTAT- 3’(forward) and 5’-GCCCAGAAGTTCTTGCTGTT -3’(reverse). PCR was performed with 1 cycle at 50°C for 2 min and 95°C for 10 min, followed by 40 cycles at 95°C, 15 s, and 60°C, 1 min. Gene expression was normalized to actin, and relative mRNA levels were calculated based on the comparative CT method. Reverse transcription PCR primer sequences are 5’- CCGGAAGATGGAAAAGCAGA- 3’(forward) and 5’-TCCTTTCCTCCACAAGCCT -3’(reverse). PCR products were separated on a 1.2% agarose gel. Following electrophoresis, DNA bands were cut out of the agarose gel and sequence the DNA samples.

Drosophila Stocks

Two Tango14 RNAi fly lines were obtained from the Bloomington Drosophila Stock Center (stock number: 31571) and Vienna Drosophila Resource Center (stock number: v42499), respectively. The mRNA expression levels of these two RNAi lines under the driver of pan-neural Elav-GAL4 were quantified by qPCR. The knockdown efficiencies of these two RNAi lines were 64% (31571) and 50% (v42499) respectively. GAL4 flies were obtained from the Bloomington *Drosophila* Stock Center. Flies were raised at 25℃ according to standard procedures.

Climbing ability

Groups of ten 3-day-old and 30-day-old male *NUS1* knockdown flies driven by Elav-Gal4 were gently tapped to the bottom of the container and allowed to climb up the line (15 cm) to assay their climbing ability. The average climbing time for 3 trials (± standard deviation (s.d.) was calculated for each genotype.

Dopamine Measurements

Whole-mount immunohistochemistry for TH staining was performed as described (20). Immunohistochemistry for TH staining used anti-TH antiserum Ab152 (1:100, Millipore), 3-day-old and 30-day-old male flies induced by dopaminergic neuron-specific TH-Gal4， and PPL1, PPM1/2, PPM3 clusters were quantitated from confocal images.

HPLC analysis of dopamine levels was performed as described (21). For sample preparation, 3-day-old and 30-day-old male fly's heads were dissected out and homogenized in 0.1 M perchloric acid. The homogenate was frozen on dry ice and stored at -80°C before HPLC analysis. Mean±s.d. were from n= 3 experiments.

TUNEL assay

Analysis of the apoptotic signal in 30-day-old male flies driven by Elav-Gal4 was previously described by Huang et al. (22). TUNEL analysis was detected using the in situ cell death detection kit (Roche).

Proband clinical description

The characteristics and phenotypic variables of the probands in 39 EOPD families (Patients cohorts) shown in **Dataset S6**. Multiple genetic and clinical examinations in our preliminary work are also shown in the same table.

Sample Cohort

All the PD patients and their family members were subjected to a standard clinical neurological examination. The diagnosis of idiopathic PD was made according to the United Kingdom PD Brain Bank Criteria. The family members did not have any nervous or psychiatric system diseases. All the probands had undergone brain MRI examinations that showed no evident lesions in the brain, and some of them had received a PET scan that showed decreased^{11} C-CFT uptake in the putamen (**Fig. S13**). We excluded those patients with aberrant short tandem repeat expansions in *SCA2*[2], *SCA3*[2], *SCA17*[3], *C9orf72*[4], as well as the rearrangements and point and indels mutations in *Parkin*[5], *PINK1*[5], and *DJ-1*[5], and point and indels mutations in *FBXO7*[6], *PLA2G6*[7], *GCH1*, *TH*, *SPR* and *ATP7B* (unpublished data).

SALSA MLPA kits P051-C3 also assessed SNCA, ATP13A2, GCH1, and TH rearrangements and P099-C2 (MRC Holland, Amsterdam, The Netherlands) (unpublished data). We also excluded individuals carrying rare variants with large odds ratios in *GBA*[8], *LRRK2*[9], and *SMPD1*[10].

All 39 probands were collected from Han Chinese non-consanguineous families, and none of the first- and second-degree relatives of the 39 probands had PD or Parkinsonism. We chose only young patients with an age at onset of at most 35 years (**Dataset S6**), because they may have had fewer chances to be affected by environmental factors, more possibility affected by genetic factors, and more difficulties to getting married than the late onset patients. All the patients involved in our study had no known history of heavy metal/pesticide/carbon monoxide exposure, drug abuse, or antipsychotic drug use, and were not previously diagnosed with diabetes, stroke or encephalitis.

Detection and validation of *de novo* mutations

We applied a two-stage strategy to detect *de novo* mutations. In the first stage, we used stringent quality criteria (**Materials and Methods**) to eliminate false positive mutations. *De novo* SNVs 91.94% (57/62) and indels 28.57% (2/7) were confirmed after validation with Sanger sequencing.

To estimate the number of *de novo* mutations missing in the first stage, we relaxed the quality criteria, which resulted in the addition of 32 *de novo* SNVs and 1 indel. Of these additions, only one *de novo* SNV (3.13%) and zero indels (0%) were validated, suggesting most of the *de novo* mutations were identified.

Known causative PD genes

To date, 20 genes are reported to cause PD/Parkinsonism causative through monogenic inheritance: *LRRK2*, *PARK2* (*Parkin*), *PLA2G6*, *DNAJC13*, *GIGYF2*, *FBXO7*, *SYNJ1*, *HTRA2*, *EIF4G1*, *SNCA*, *DNAJC6*, *VPS35*, *ATP13A2*, *PINK1*, and *PARK7* (*DJ-1*), *UCHL1*, *RAB39B CHCHD2*, *VPS13C* and *TMEM230*[11-16].

Calculation of exome-wide *de novo* mutation rates

The *de novo* mutation rates can fluctuate and are influenced by childbearing age of the parents, sample size, and other factors. Based on the quality thresholds used to explore *de novo* mutations, the number of nucleotides covered at least 8 times (the lowest depth of our validated *de novo* mutation) for all the members of each family were used as denominator. The total number of the confirmed *de novo* SNVs divided by the denominator was the observed mutation rates.

Comparison with private inherited mutations

Private inherited mutations are defined herein as inherited mutations that are unique family and inherited from one of the parents. We identified 22,866 private inherited mutations in the 38 EOPD families; of them, 13,576 were nonsynonymous, 8,612 were synonymous, whereas 285 SNVs were nonsense, 17 were located in canonical splice sites, and 161 were indels.

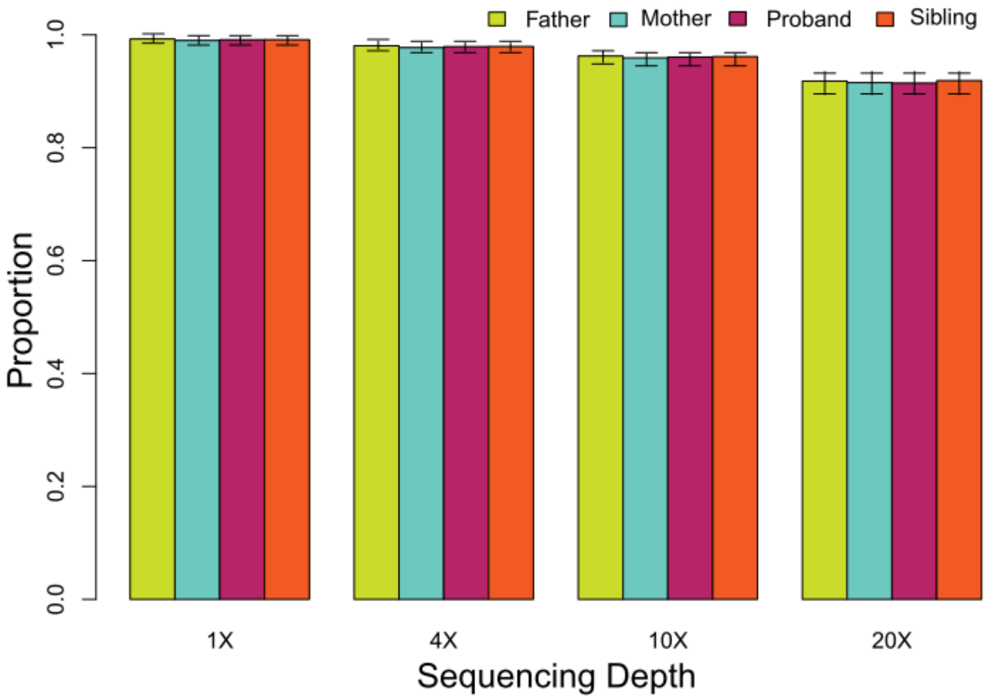
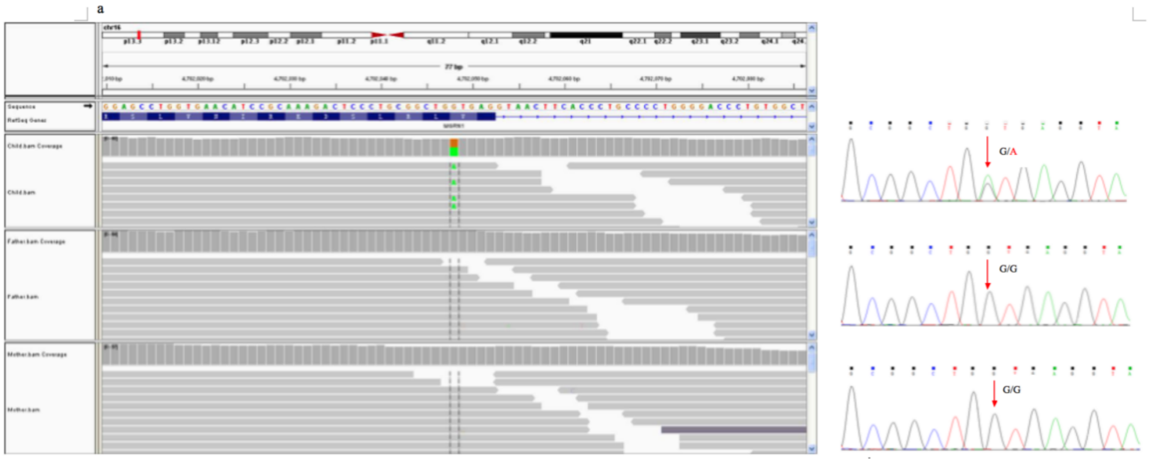
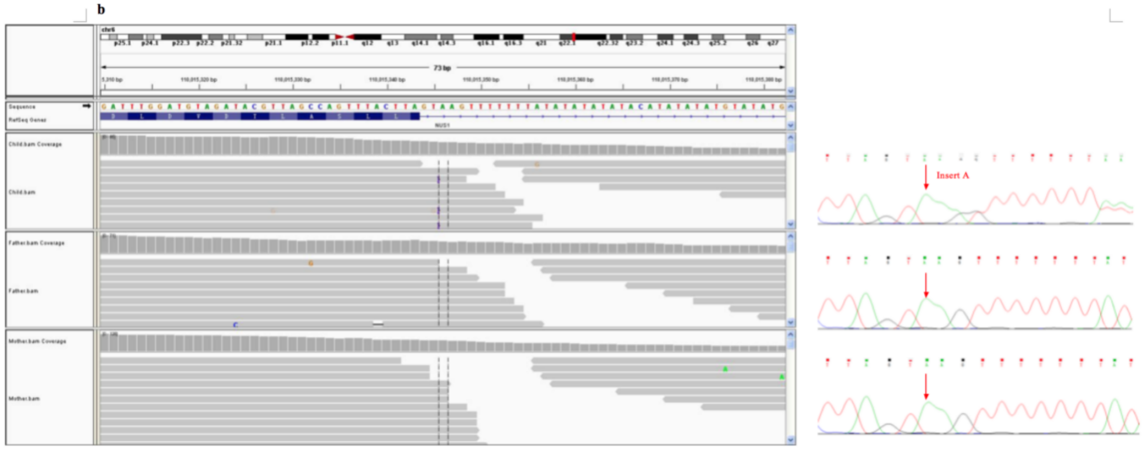


Fig. S1. Sequencing coverage of the exonic target regions. The average proportions of read depth in the target regions at 1X, 4X, 10X, and 20X for all parents, probands and their siblings.


a
chr16
Sequence
RefSeq Genes
Child.bam Coverage
Child.bam
Father.bam Coverage
Father.bam
Mother.bam Coverage
Mother.bam
G/A
G/G
G/G

(a) SNV


b
chr6
Sequence
RefSeq Genes
Child.bam Coverage
Child.bam
Father.bam Coverage
Father.bam
Mother.bam Coverage
Mother.bam
Insert A

(b) Indel

Fig. S2. An example for Sanger sequencing validation of *de novo* SNVs and indels showing confirmation of *de novo* mutations in probands. Left subfigures: IGV (Integrative Genomics Viewer) browser view of *de novo* mutations in (a) *MGRN1* and (b) *NUS1*. Top panel shows the mutation location indicated by a red tag. Middle panel shows the reference sequence and translated amino acids. Bottom panel represents the reads pileup for the proband, father, and mother. Right subfigures: Sanger sequencing traces. Red arrow in top panel indicates *de novo* mutation in probands, the middle and bottom ones are for the parents, respectively.

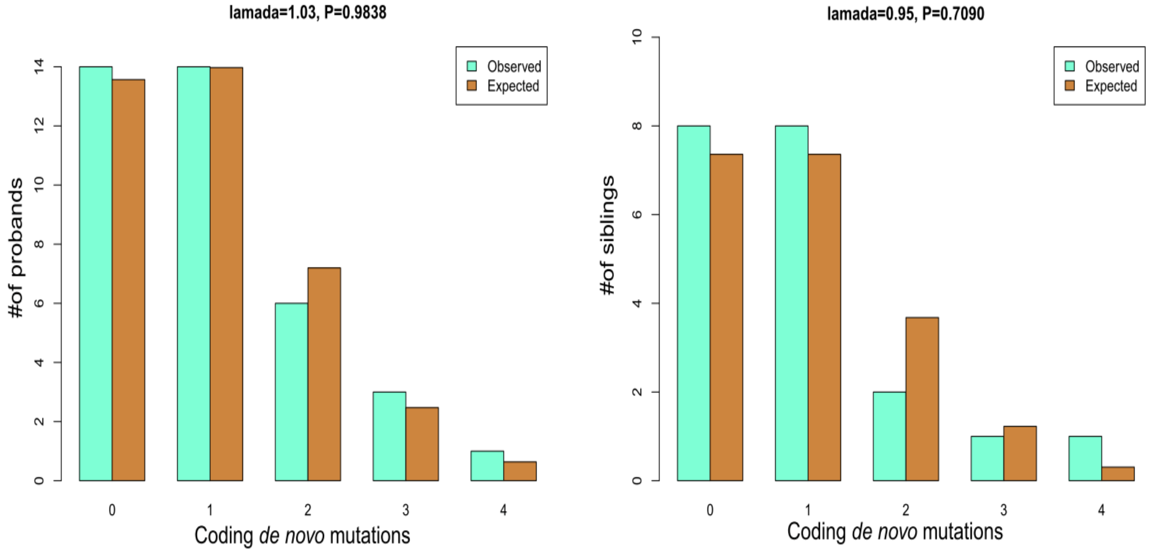


Fig. S3. The number of *de novo* mutations in probands and siblings follows Poisson distribution (Probands: P=0.98, Siblings: P=0.71). The average number of *de novo* mutations for probands and siblings are 1.03 and 0.95, respectively.

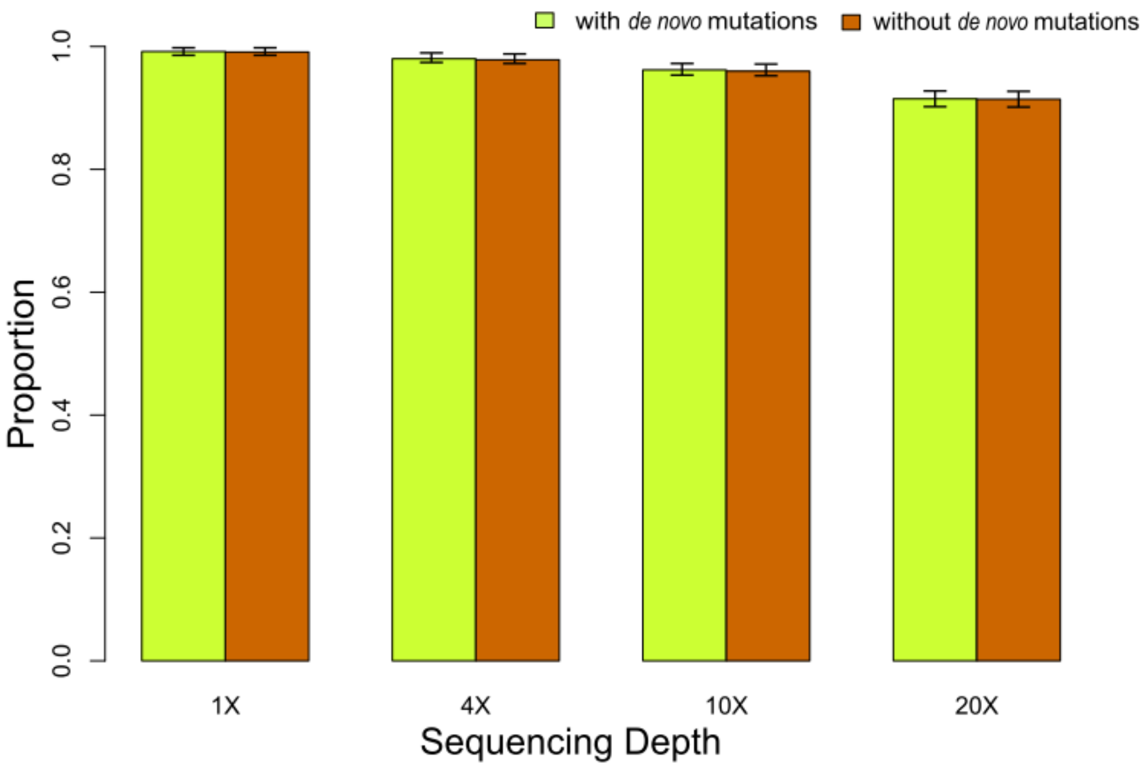
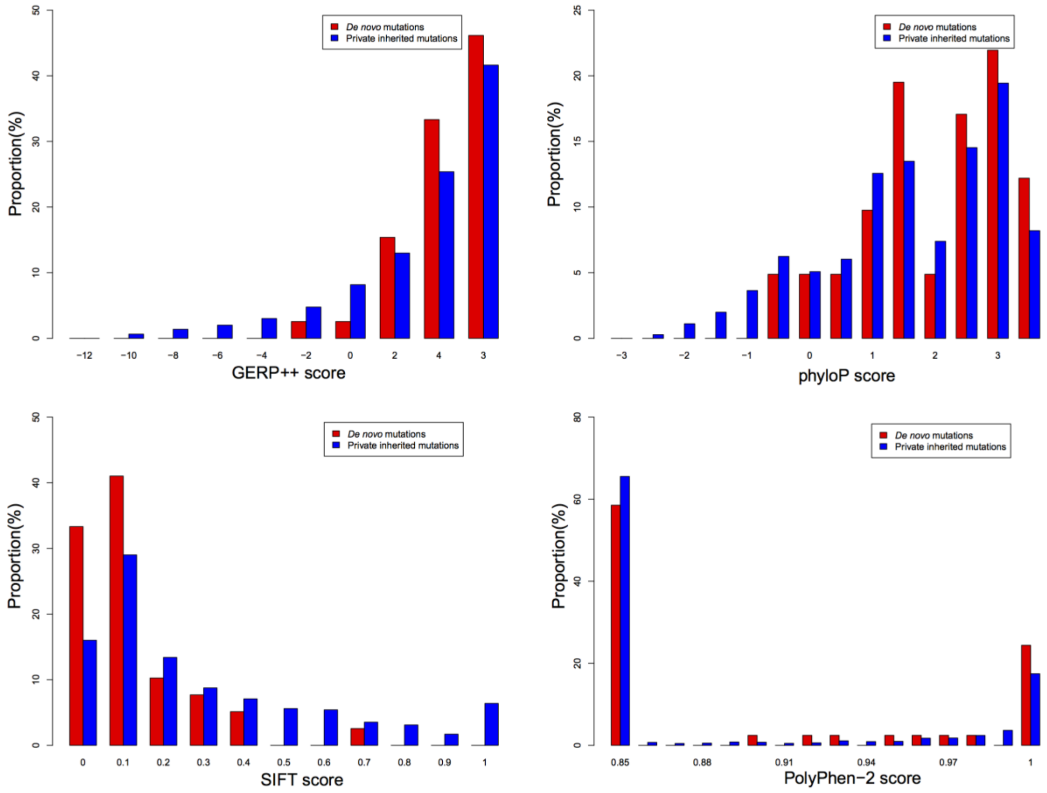
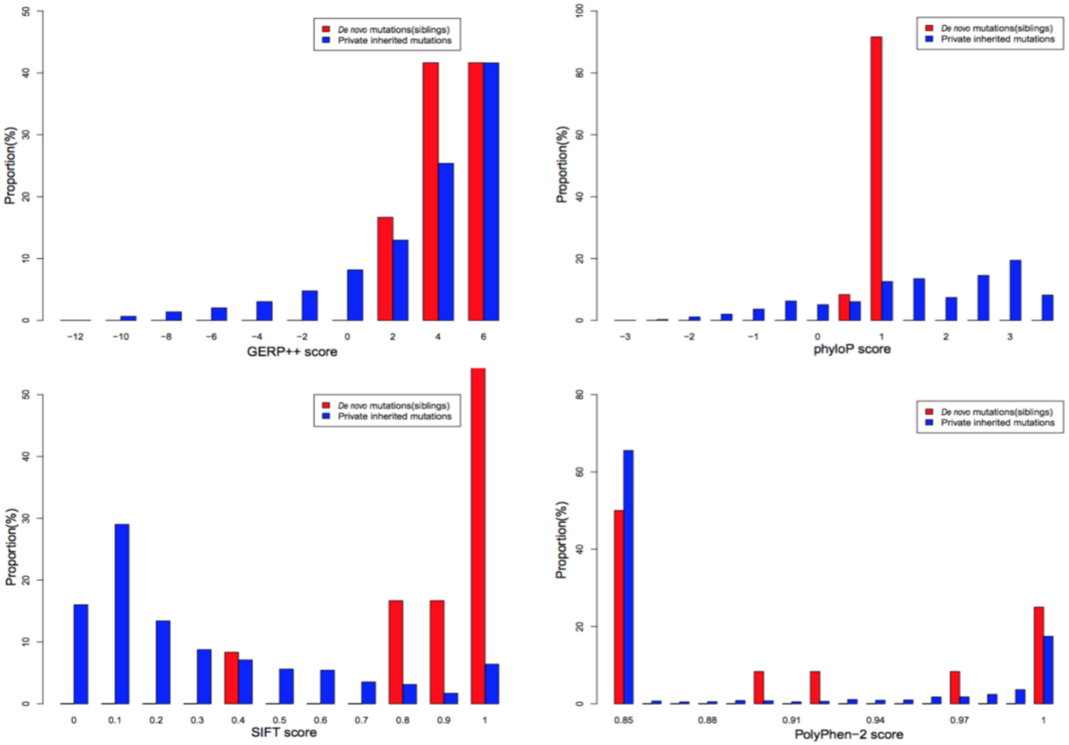


Fig. S4. Sequencing depth comparison for the samples with or without *de novo* mutations in the target regions. No sequencing depth bias was observed between probands and siblings.


De novo mutations
Private inherited mutations
Proportion(%)
0
10
20
30
40
50
−12
−10
−8
−6
−4
−2
0
2
4
3
GERP++ score
De novo mutations
Private inherited mutations
Proportion(%)
0
5
10
15
20
25
−3
−2
−1
0
1
2
3
phyloP score
De novo mutations
Private inherited mutations
Proportion(%)
0
10
20
30
40
50
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
SIFT score
De novo mutations
Private inherited mutations
Proportion(%)
0
20
40
60
80
0.85
0.88
0.91
0.94
0.97
1
PolyPhen−2 score

A. Probands



B. Siblings

Fig. S5. Assessment of the potential pathogenicity of *de novo* mutations identified in the probands and siblings in terms of the conserved and deleterious amino acid changes. **A.** Comparison of the distributions of GERP++ scores (P=0.14), phyloP scores (P=0.04), SIFT scores (P=2.18E-05) and PolyPhen-2 scores (P=0.03) for the *de novo* mutations and private inherited variants found in the probands. **B.** Comparison of the distributions of GERP++ scores (P=0.93), phyloP scores (P=1.00), SIFT scores (P=0.07) and PolyPhen-2 scores

(P=0.99) for the *de novo* mutations and private inherited variants found in the siblings. The P values were calculated by the Wilcoxon rank sum test.

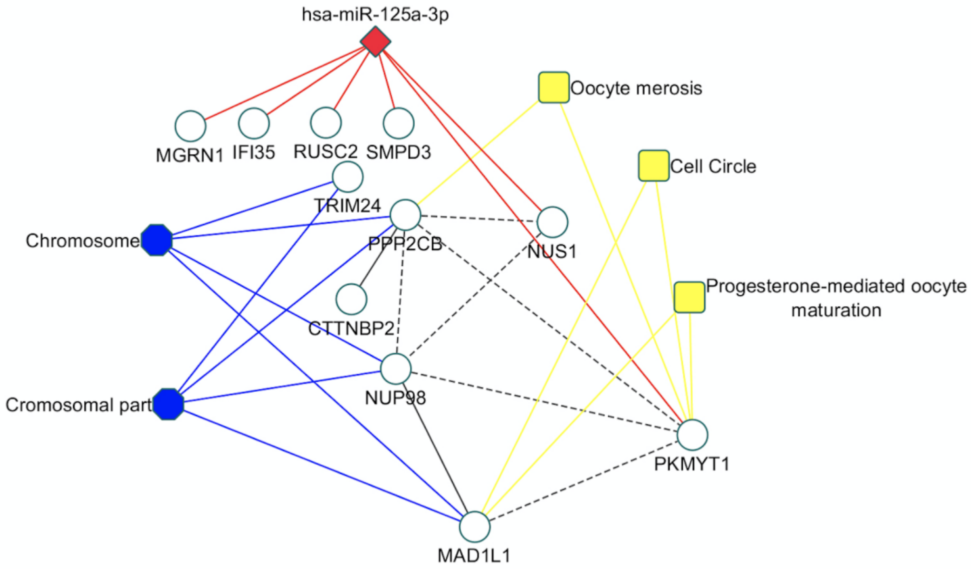


Fig. S6. Connectome of the 12 new candidate genes with *de novo* mutations. Six genes (*NUP98*, *MAD1L1*, *PPP2CB*, *PKMYT1*, *CTTNBP2*, and *NUS1*) are involved in the Protein-Protein interaction network predicted by DAPPLE. The solid black lines represent direct interactions, and the dashed black lines indicate indirect interactions. The candidate genes significantly enriched in two gene ontology terms, as shown by blue lines: chromosome ($P_{corrected}$=6.78E-03) and chromosomal part ($P_{corrected}$=1.15E-02). The KEGG pathway enrichment analysis discovered three significant pathways, shown with yellow lines: progesterone-mediated oocyte maturation

($P_{corrected}$=0.03), cell cycle ($P_{corrected}$=0.03), and oocyte meiosis ($P_{corrected}$=0.03). Overall, 6 genes with protein-altering *de novo* mutations were predicted as the targets of hsa-miR-125a-3p ($P_{corrected}$= 6.50E-03), as shown by red lines.

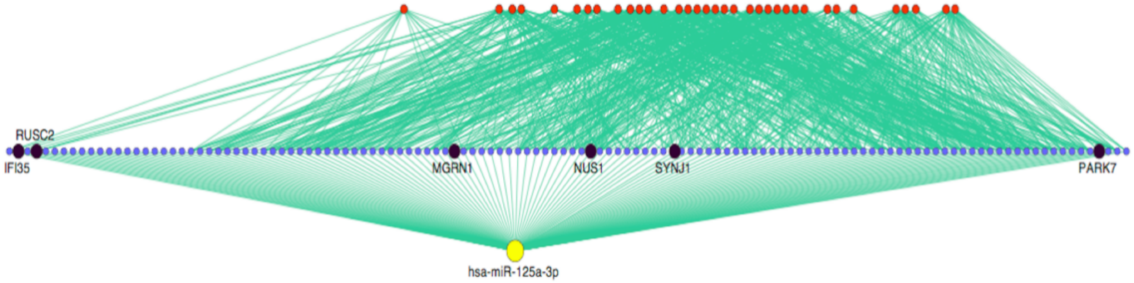
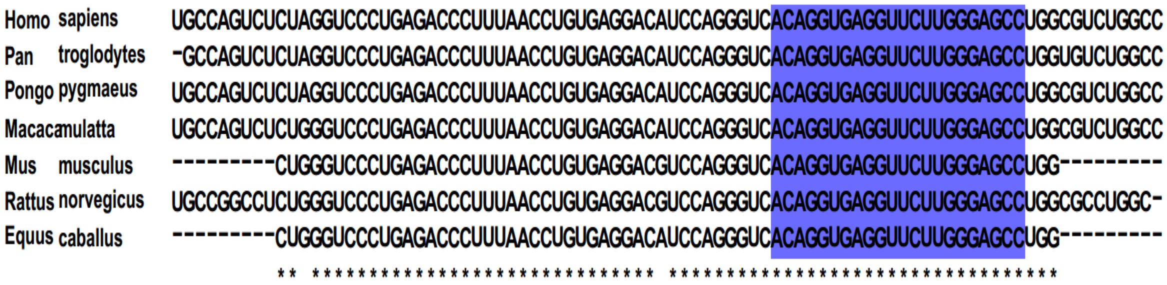


Fig. S7. The interaction network between has-miR125a-3p and its targets. Besides hsa-miR-125a-3p (yellow circle), other known PD-related microRNAs (red circle, **Dataset S3**) were also included. The selected targeted genes (blue and black circles) contained validated targets of hsa-miR-125a-3p and its targets predicted in our study (*PKMYT1*, *NUS1*, *SMPD3*, *MGRN1*, *RUSC2* and *IFI35*). Six genes that were co-targets of has-miR-125a-3p and other know PD-related microRNA were highlighted as black circles.



```
Homo sapiens      UGCCAGUCUCUAGGUCCCUGAGACCCUUUAACCUGUGAGGACAUCCAGGGUCACAGGUGAGGUUCUUGGGAGCCUGGCGUCUGGCC
Pan troglodytes   -GCCAGUCUCUAGGUCCCUGAGACCCUUUAACCUGUGAGGACAUCCAGGGUCACAGGUGAGGUUCUUGGGAGCCUGGUGUCUGGCC
Pongo pygmaeus    UGCCAGUCUCUAGGUCCCUGAGACCCUUUAACCUGUGAGGACAUCCAGGGUCACAGGUGAGGUUCUUGGGAGCCUGGCGUCUGGCC
Macacamulatta     UGCCAGUCUCUGGGUCCCUGAGACCCUUUAACCUGUGAGGACAUCCAGGGUCACAGGUGAGGUUCUUGGGAGCCUGGCGUCUGGCC
Mus musculus      ---------CUGGGUCCCUGAGACCCUUUAACCUGUGAGGACGUCCAGGGUCACAGGUGAGGUUCUUGGGAGCCUGG---------
Rattus norvegicus UGCCGGCCUCUGGGUCCCUGAGACCCUUUAACCUGUGAGGACGUCCAGGGUCACAGGUGAGGUUCUUGGGAGCCUGGCGCCUGGC-
Equus caballus    ---------CUGGGUCCCUGAGACCCUUUAACCUGUGAGGACAUCCAGGGUCACAGGUGAGGUUCUUGGGAGCCUGG---------
                           ** ****************************** **********************************
```

Fig. S8. Multiple sequence alignment and conservation of hsa-miR125a-3p among seven vertebrate species. The region in blue defines the sequence of hsa-miR125a-3p, *denotes conserved positions. The multiple sequence alignment was obtained from miRviewer (http://people.csail.mit.edu/akiezun/microRNAviewer/).

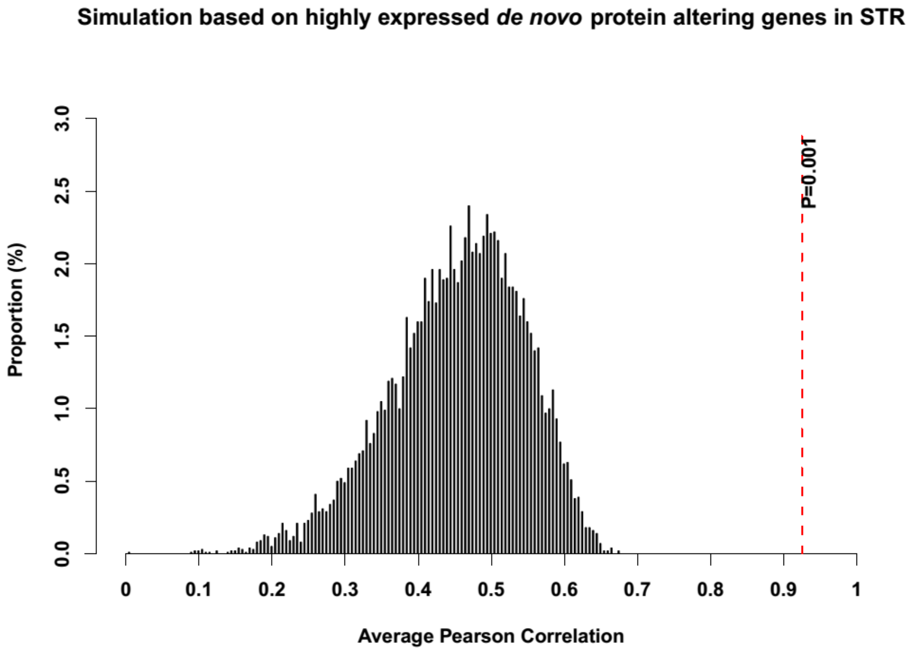
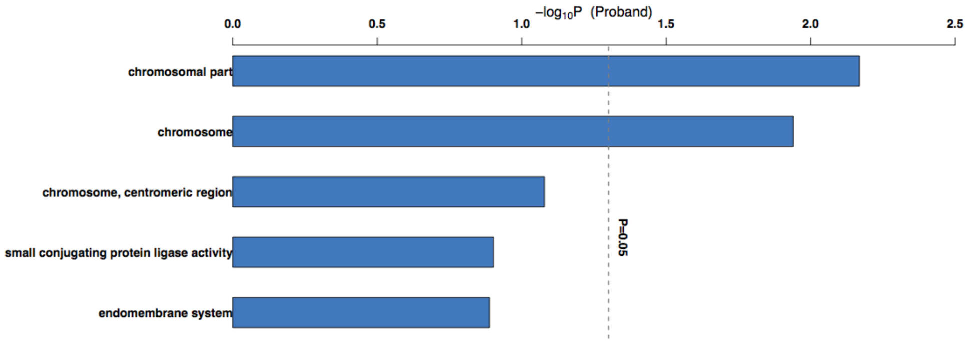
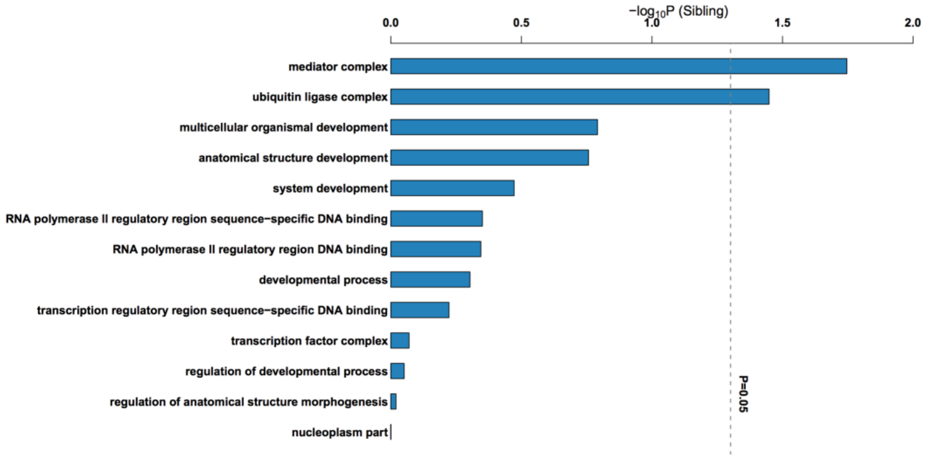


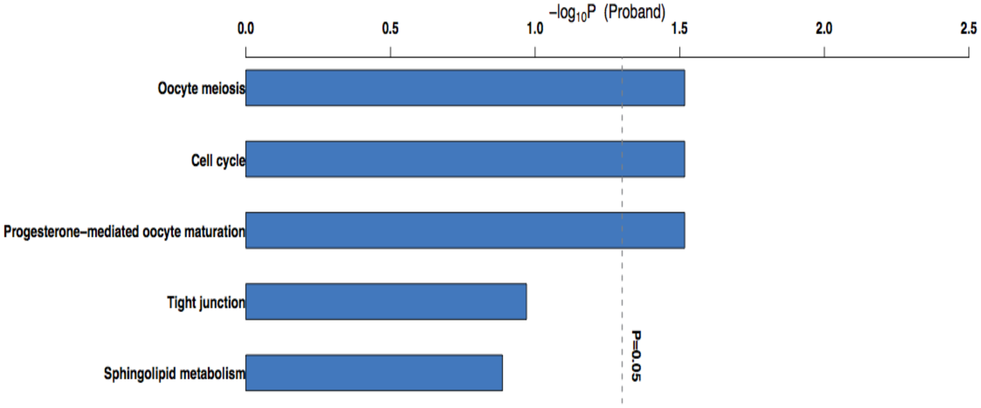
Fig. S9. The empirical distribution and P values for gene co-expression of the 12 PD new candidate genes. The empirical distribution was generated from 100,000 simulations, each contained randomly selected 12 genes. The gene co-expression for each network was calculated by average Pearson correlation. The red dashed line represents the co-expression of 12 PD new candidate genes. All the gene expressions were extracted from BRAINSPAN.



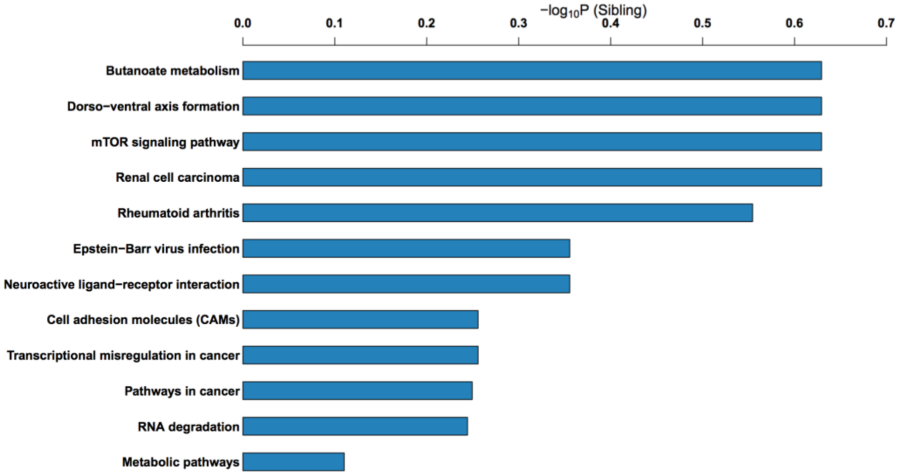
A. Gene Ontology enrichment for the probands



B. Gene Ontology enrichment for the siblings



C. KEGG pathway enrichment for the probands



D. KEGG pathway enrichment for the siblings

Fig. S10. Gene Ontology and KEGG pathway enrichment for the 12 new candidate genes in the probands and *de novo* altering genes in the siblings. Two GO terms (chromosome, $P_{corrected}$=6.78E-03; chromosomal part, $P_{corrected}$=1.15E-2) and three KEGG pathways (progesterone-mediated oocyte maturation $P_{corrected}$=0.03; cell cycle, $P_{corrected}$=0.03; oocyte meiosis, $P_{corrected}$=0.03) are significant in the probands. Two GO terms (mediator complex, $P_{corrected}$=1.79E-2; ubiquitin ligase complex, $P_{corrected}$=3.56E-2) and zero KEGG pathways are significant in the siblings. Chromosome: *TRIM24*, *NUP98*, *MAD1L1*, *PPP2CB*; chromosomal part: *TRIM24*, *NUP98*, *MAD1L1*, *PPP2CB*; progesterone-mediated oocyte maturation: *MAD1L1*, *PKMYT1*; cell cycle: *MAD1L1*, *PKMYT1*; oocyte meiosis: *PPP2CB*, *PKMYT1*; mediator complex: *MED12, MED23;* ubiquitin ligase complex: *MED12, MED23, FBXL15*.

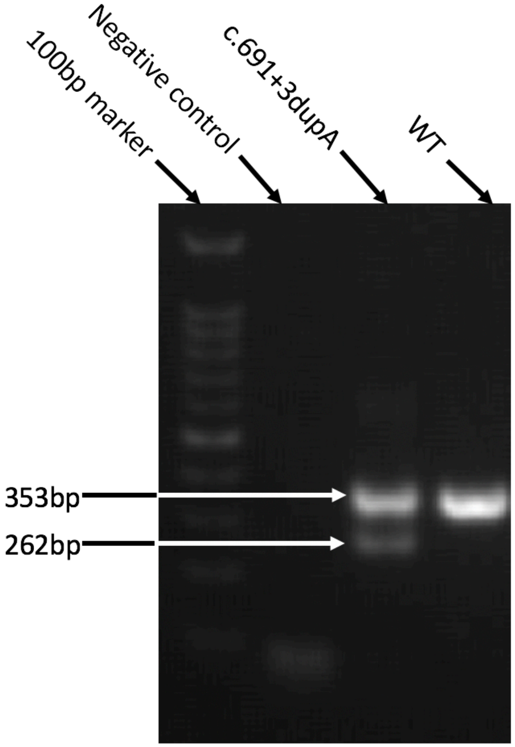


Fig. S11. Agarose gel electrophoresis of RT-PCR fragments produced by mRNAs extracted from patient and age-matched healthy control.

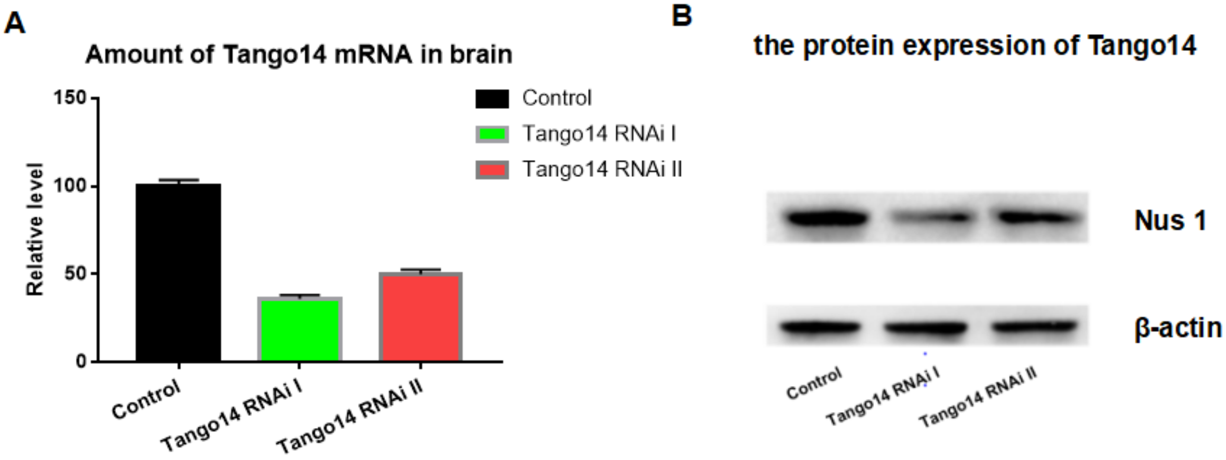


Fig. S12. The expression of Tango14 in the two Tango14 RNAi lines. A: mRNA expression level; B: protein expression level

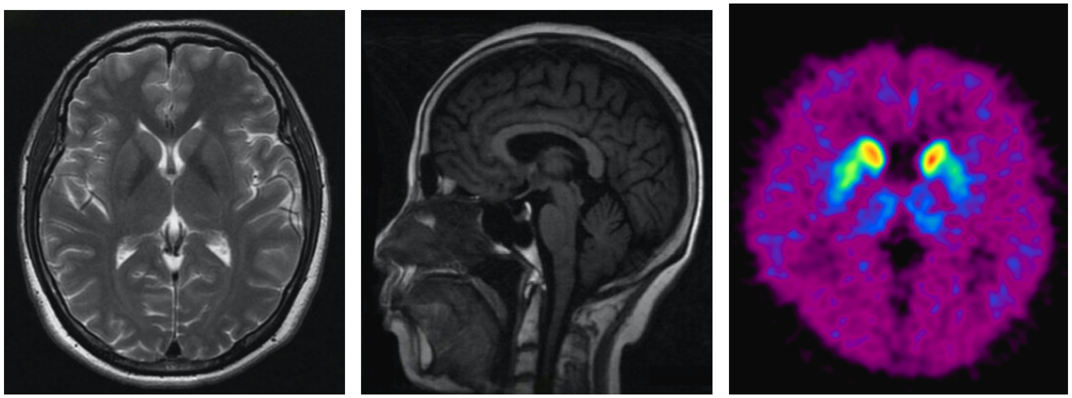


Fig. S13. Examples of brain MRI and PET examination of the EOPD probands. MRI (T2-weighted and T1-weighted) shows one patient with no evident lesions in the brain. PET results exhibit reduction of DAT binding (^{11}C-CFT) in posterior putamen nucleus in another patient.

	Mean ± Standard deviation
Read length (bp)	90
Number of individuals	137
Raw reads (Gb)	10.89±1.40
Mapped reads (Gb)	10.80±1.38
Mapped reads on target region (Gb)	3.95±0.49
Mapping rate (%)	98.39±0.31
Average sequencing depth (fold)	61.49±7.63
Proportion of target region covered ≥1X (%)	99.13±0.19
Proportion of target region covered ≥4X (%)	97.90±0.33
Proportion of target region covered ≥10X (%)	96.08±0.49
Proportion of target region covered ≥20X (%)	91.63±1.63

Table S1. The general information of exome sequencing data from 39 EOPD families

Fam_id	Pheno	Chr	Type	Position[*]	Ref	Alt	AAC	ExAC	gnomAD	Conservation[#]	Function	Gene
Quad1	Proband	chr12	SNV	6483742	C	T	p.G93S	-	-	Y	nonsynonymous	*SCNN1A*
Quad2	Proband	chr11	SNV	3697839	T	G	p.S1695S	-	-	Y	splicing[+]	*NUP98*
Quad4	Proband	chr19	SNV	31770010	G	A	p.T230M	-	-	Y	nonsynonymous	*TSHZ3*
Quad5	Proband	chr14	SNV	21870167	T	G	p.R1058R	-	-	Y	synonymous	*CHD8*
Quad8	Proband	chr12	SNV	52307354	C	T	p.P109S	-	-	N	nonsynonymous	*ACVRL1*
Quad10	Proband	chr17	SNV	79180926	A	G	p.L129P	-	-	Y	nonsynonymous	*CEP131*
Quad10	Proband	chr16	SNV	68404992	C	T	p.V365M	-	-	Y	splicing[+]	*SMPD3*
Quad11	Proband	chr5	SNV	56180628	C	T	p.Y1319Y	-	4.1E-6	Y	synonymous	*MAP3K1*
Quad13	Proband	chr5	SNV	150889595	A	G	p.Y4016H	-	-	N	nonsynonymous	*FAT2*
Quad14	Proband	chr11	SNV	20483714	G	T	NA	-	-	Y	splicing	*PRMT3*
Quad14	Proband	chr2	SNV	219562211	C	T	p.A929A	9.1E-5	1E-4	N	synonymous	*STK36*
Quad15	Proband	chr19	SNV	691872	T	C	p.I123V	8.4E-6	8.1E-6	Y	nonsynonymous	*PRSS57*
Quad16	Proband	chr7	SNV	2255841	T	A	p.R254W	-	-	Y	nonsynonymous	*MAD1L1*
Quad16	Proband	chr17	SNV	41158986	G	A	NA	8.2E-6	8.1E-6	Y	splicing	*IFI35*
Quad16	Proband	chr16	SNV	2225368	C	T	p.L485L	-	-	Y	synonymous	*TRAF7*
Quad18	Proband	chr16	SNV	3023235	G	A	p.L444F	-	-	N	nonsynonymous	*PKMYT1*
Quad18	Proband	chr2	SNV	101638833	G	A	p.H876Y	-	-	Y	nonsynonymous	*TBC1D8*
Quad18	Proband	chr12	SNV	64519788	T	C	p.F752F	-	-	Y	synonymous	*SRGAP1*
Quad18	Proband	chr14	SNV	65237617	G	T	p.T1928T	-	-	Y	synonymous	*SPTB*
Quad19	Proband	chr1	SNV	43228142	T	C	p.N157S	-	-	Y	nonsynonymous	*LEPRE1*
Quad19	Proband	chr16	SNV	4702048	G	A	p.V98M	-	8.1E-6	Y	nonsynonymous	*MGRN1*
Quad20	Proband	chr15	SNV	34537570	G	A	p.T611I	-	-	Y	nonsynonymous	*SLC12A6*
Trio1	Proband	chr8	SNV	30651593	C	T	p.G193D	-	-	Y	nonsynonymous	*PPP2CB*
Trio3	Proband	chr11	SNV	59190311	G	A	p.T39M	9.8E-5	9.8E-5	N	nonsynonymous	*OR5A2*
Trio3	Proband	chr7	SNV	117407153	A	T	p.D952E	-	-	Y	nonsynonymous	*CTTNBP2*
Trio4	Proband	chr19	SNV	54599121	G	A	p.A228V	-	-	N	nonsynonymous	*OSCAR*
Trio4	Proband	chr4	SNV	6302743	T	G	p.H407Q	8.2E-6	8.1E-6	Y	nonsynonymous	*WFS1*
Trio4	Proband	chr18	SNV	44560405	G	A	p.Q411X	-	-	N	stopgain	*TCEB3B*
Trio5	Proband	chr7	SNV	138266452	G	A	p.R910H	8.2E-6	4.1E-6	Y	nonsynonymous	*TRIM24*
Trio6	Proband	chr1	SNV	62588713	A	T	p.T1676S	-	-	Y	nonsynonymous	*INADL*
Trio7	Proband	chr13	SNV	97639686	G	A	p.R110C	7.4E-5	6.9E-5	Y	nonsynonymous	*OXGR1*
Trio7	Proband	chr9	SNV	33933540	T	C	p.T686A	-	-	N	nonsynonymous	*UBAP2*

Trio8	Proband	chr4	SNV	145041693	G	C	p.T29S	-	-	N	nonsynonymous	*GYPA*
Trio10	Proband	chr6	Indel	118015345	-	A	NA	-	-	Y	splicing	*NUS1*
Trio16	Proband	chr6	SNV	129636689	G	C	p.K1208N	-	-	Y	nonsynonymous	*LAMA2*
Trio17	Proband	chr2	SNV	20182229	T	C	p.H70R	-	-	Y	nonsynonymous	*WDR35*
Trio17	Proband	chr9	SNV	35558225	A	G	p.N1031S	-	4.1E-6	Y	nonsynonymous	*RUSC2*
Trio17	Proband	chr12	SNV	11338711	G	A	p.S278L	-	8.1E-6	N	nonsynonymous	*TAS2R42*
Trio18	Proband	chr2	SNV	55490814	C	T	p.A61T	-	-	N	nonsynonymous	*MTIF2*
Trio18	Proband	chr3	SNV	157131859	A	G	p.P239P	8.2E-6	4.1E-6	Y	synonymous	*VEPH1*

All the variants were not found in dbSNP137 and 1000 Genomes project. *GRCh37 (hg19) human reference genome. [#] Conservation scores were calculated by phastCons. splicing+: predicted splice site. Pheno: phenotype (probands or siblings). Ref: reference allele. Alt: alternative allele. AAC: amino acid change. § One trio was removed prior to further analysis due to its exceedance of Mendelian errors

Table S2. *De novo* mutations confirmed in the 38 probands[§].

Fam_id	Pheno	Chr	Type	Position*	Ref	Alt	AAC	ExaC	gnomAD	Conservation#	Function
Quad2	Sibling	chrX	SNV	70357214	G	A	p.R1910H	-	-	Y	nonsynonymous
Quad3	Sibling	chr6	SNV	46657771	A	G	p.K636E	-	-	Y	nonsynonymous
Quad3	Sibling	chr3	SNV	52547914	C	T	p.R1122X	8.4E-6	8.2E-6	N	stopgain
Quad3	Sibling	chr1	SNV	176852015	C	A	p.L1114L	-	-	Y	synonymous
Quad4	Sibling	chr3	SNV	170828652	C	T	p.R712Q	3.0E-5	3.1E-5	Y	nonsynonymous
Quad5	Sibling	chr10	SNV	104181111	G	T	p.R18S	-	-	Y	splicing+
Quad6	Sibling	chr16	SNV	20638526	A	G	p.I471T	-	-	Y	nonsynonymous
Quad7	Sibling	chr19	SNV	10205574	C	A	p.R208L	8.2e-6	-	N	nonsynonymous
Quad7	Sibling	chr8	SNV	113318282	T	C	p.Q2635Q	-	-	Y	synonymous
Quad9	Sibling	chr1	SNV	63789444	C	G	p.R239G	8.2e-6	-	Y	nonsynonymous
Quad13	Sibling	chr4	SNV	71346641	A	G	p.R60R	8.2E-6	4.1E-6	N	synonymous
Quad14	Sibling	chr2	SNV	96781599	G	A	p.T97I	-	-	Y	nonsynonymous
Quad14	Sibling	chr20	SNV	62705375	C	T	p.R162Q	1.7E-5	1.2E-5	Y	nonsynonymous
Quad14	Sibling	chr5	SNV	5235236	G	A	p.E654K	1.7E-5	8.1E-6	Y	nonsynonymous
Quad14	Sibling	chr17	SNV	7256378	C	T	p.F39F	-	4.1E-6	Y	synonymous
Quad15	Sibling	chr12	SNV	57677634	C	T	p.G368S	4.1E-5	2.4E-5	Y	nonsynonymous
Quad15	Sibling	chr17	SNV	74398741	G	A	p.L210L	1.6E-5	1.2E-5	Y	synonymous
Quad16	Sibling	chr6	SNV	131917740	C	T	p.R905Q	-	4.1E-6	Y	nonsynonymous
Quad19	Sibling	chr15	SNV	80847418	A	G	p.K368E	-	-	Y	nonsynonymous

Quad20	Sibling	chr2	Indel	168099403	ACCGTT	-	p.499_501del	-	-	Y	nonframeshift_deletion

All the variants were not found in dbSNP137 and 1000 Genomes project. *GRCh37 (hg19) human reference genome. # Conservation scores were calculated by phastCons. splicing+: predicted splice site. Pheno: phenotype (probands or siblings). Ref: reference allele. Alt: alternative allele. AAC: amino acid change.

Table S3. *De novo* mutations confirmed in the 20 siblings.

	Subjects (#dnMs)	dnMR	nsMR	dnMs per sample	nsMs per sample	P value
Probands	38(39)	1.67E-08	1.46E-08	1.03	0.89	0.50
Siblings	20(19)	155E-08	1.14E-08	0.95	0.70	-
Total	58(58)	1.63E-08	1.35E-08	1.00	0.83	0.52

*Nonsynonymous mutations include nonsense, missense mutations, and the mutations in splicing sites; dnMs, de novo mutations; dnMR, de novo mutation rate; nsMs, nonsynonymous mutations; P value, from Fisher's exact test.

Table S4. The *de novo* mutation rate between probands and siblings. There is no significant difference between them.

	Probands		Siblings	
The number of *de novo* mutations	0 (n=14)	≥1 (n=24)	0 (n=8)	≥1 (n=12)
Male subjects (%)	8 (57.14%)	14 (58.33%)	5 (62.50%)	7 (58.33%)
PD age at onset (average age/year)	29.07	31.71	-	-
Paternal childbearing age (average age/year)	27.71	28.5	28.75	29.08
Maternal childbearing age (average age/year)	25.29	25.21	25.29	25.21

Table S5A. Basic descriptive statistics.

Sample characteristics	P value (Wilcoxon Rank Sum test, one tail)

	P_0+P_1 vs. S_0+S_1	P_0 vs. P_1	S_0 vs. S_1	P_0+S_0 vs. P_1+S_1
PD age at onset	-	0.96	-	-
Paternal childbearing age	0.82	0.27	0.36	0.28
Maternal childbearing age	0.63	0.49	0.44	0.48

P_0: probands without de novo mutations; P_1: probands with *de novo* mutations;

S_0: siblings without de novo mutations; S_1: siblings with *de novo* mutations

Table S5B. The evaluation 1. PD age at onset, and 2. Parental childbearing age between the subjects with or without *de novo* mutations by Wilcoxon rank sum test.

	Probands		Siblings		Probands+Siblings	
Sample characteristics	Male	Female	Male	Female	Male	Female
De novo mutation ≥1	14	10	7	5	21	15
De novo mutation =0	8	6	5	3	13	9
Male vs. Female (P value)	0.60		0.74		0.63	

Table S5C. Evaluations of the gender in the subjects with or without *de novo* mutations by Fisher's exact test.

Table S5. The comparison of offspring gender, parental childbearing age, and probands' age at onset between 1. probands, siblings, and 2. probands with or without *de novo* mutations.

	Probands (*de novo* mutations)	Siblings (*de novo* mutations)	Probands and Siblings (Private inherited mutations)
Nonsynonymous (NS)	32	14	13,878
Synonymous (S)	7	5	8,612
NS:S	4.57	2.80	1.61
P value*	**5.35E-3**	0.20	-
Odds ratio	4.22	1.74	-
Loss of function (LoF)	6	2	392
Missense	27	12	13,576
LoF:missense	0.22	0.17	0.03
P value	**2.94E-4**	0.06	-
Odds ratio	7.69	5.77	-

Nonsynonymous mutations include nonsense, missense mutations, and the mutations in splicing sites; Loss of function mutations include frameshift, nonsense mutations, and the mutations in splicing sites. *The significant P values calculated by Fisher's exact test are in bold.

Table S6. The *de novo* mutations and private inherited mutations in probands and siblings.

Gene names	**Number of Amino Acid**	**RNSV Carriers** **(Case:Control)**	**P value**	**Expressed in brain STR/SNc**
NUP98	1817	39:30	0.72	Y/N
MAD1L1	718	28:29	0.50	Y/N
PPP2CB	309	15:9	0.54	Y/Y
PKMYT1	499	24:18	0.76	Y/N
TRIM24	393	4:5	0.74	Y/Y
CTTNBP2	639	99:69	0.23	Y/Y
NUS1	293	6:0	0.03	Y/Y
SMPD3	655	30:25	1	Y/Y
MGRN1	552	35:28	0.90	Y/Y
RUSC2	1516	31:26	1	Y/Y
CEP131	1083	71:54	0.58	Y/Y

IFI35	286	18:14	0.86	Y/Y

Table S7A. Replication of 12 candidate genes in 1,852 cases and 1,565 controls.

	Cases	**RNSV Carriers (Case)**	**Controls**	**RNSV Carriers (control)**	**P value**
Replication1	1,852	6	1,565	0	0.03
Replication2	3,237	20	2,858	2	3.2E-4
Combined	5,089	26	4,423	2	1.01E-5

Table S7B. Replication of *NUS1* on two case-control cohorts.

Table S7. Replication of candidate genes carrying *de novo* mutations. The P values were calculated by Fisher's exact test and the significant P values from were in bold. RNSV: Rare NonSynonmous Variant.

GEO ID	No. of PD mice	No. of Control mice	Description
GSE4788	15	8	Dysregulation of Gene Expression in the 1-Methyl-4-Phenyl-1,2,3,6-Tetrahydropyridine-Lesioned Mouse Substantia Nigra
GSE7707	6	6	Gene expression changes in multiple brain regions of a mouse MPTP model of Parkinson's disease
GSE20547	7	12	A53T-α-synuclein overexpression mouse model signaling and striatal synaptic plasticity
GSE60414	12	12	Potentiation of neurotoxicity in double mutant mice with Pink1 ablation and A53T-*SNCA* overexpression
GSE60413	24	23	Parkinson Phenotype in Aged *PINK1*-Deficient Mice Is Accompanied by Progressive Mitochondrial Dysfunction in Absence of Neurodegeneration
GSE52584	12	12	Gene and microRNA transcriptome analysis of Parkinson's related *LRRK2* mouse models

Table S8A. The gene expression datasets collected from GEO to calculate the differential expression between PD mice and control mice for the candidate genes in protein-protein interaction networks.

Gene names	**Adjusted P value**
CTTNBP2	3.01E-02

MAD1L1	1.47E-02
NUP98	4.87E-03
NUS1	8.84E-03
PKMYT1	5.92E-02
PPP2CB	5.64E-03
TRIM24	3.46E-03

Table S8B. The smallest adjusted P value of candidate genes in the protein-protein interaction network in six expression datasets.

Table S8. The differential expression of genes involved in the protein-protein interaction networks.

Chr	Position*	Ref	Alt	ExAC	gnomAD	Conservation#	RNSV Carriers (Case)	RNSV Carriers (Control)
chr6	117996897	C	T	-	-	Y	1	0
chr6	117996940	T	C	-	-	Y	1	0
chr6	117996941	C	G	1.3E-5	1.3E-5	Y	1	0
chr6	117997007	G	T	2.7E-4	2.7E-4	Y	1	0
chr6	117997032	C	G	-	2.5E-5	Y	1	0
chr6	117997090	G	T	3.3E-5	3.3E-5	Y	3	0
chr6	117997098	G	T	-	-	Y	1	0
chr6	117997104	G	T	-	3.5E-5	Y	1	0
chr6	117997184	C	G	1.1E-5	1.1E-5	Y	1	0
chr6	118014221	T	G	-	-	Y	2	0
chr6	118014264	C	A	-	4.1E-6	Y	1	0
chr6	118014276	G	C	8.1E-6	8.1E-6	Y	4	1
chr6	118015279	G	C	-	-	Y	2	0
chr6	118024773	A	G	-	8.1E-6	Y	1	0
chr6	118024794	T	A	-	-	Y	1	0
chr6	118024866	G	A	-	-	Y	1	0
chr6	118028241	G	A	-	-	Y	0	1
chr6	118028193	A	C	8.2E-6	8.2E-6	N	3	0

*GRCh37 (hg19) human reference genome. # Conservation scores were calculated by phastCons.

Table S9. Rare nonsynonymous variants in *NUS1* from case-control replications.
RNSV: Rare NonSynonmous Variant

Additional Data Files Content

Dataset S1. Homozygous inherited variants in probands. (Excel file)

Dataset S2. The predicted microRNAs targets on 12 candidate genes and PD known causative genes. (Excel file)

Dataset S3. PD related microRNAs reported in the previous literatures[17-24]. (Excel file)

Dataset S4. The patients with *NUS1* rare variants in our study. (Excel file)

Dataset S5. MIPs designed for the 12 new candidate genes by MIPgen. (Excel file)

Dataset S6. The probands of 39 EOPD families' characteristics and phenotypic variables (Patients cohorts). (Excel file)

Reference:

1. Yan W*, et al.* (2017) TMEM230 mutation analysis in Parkinson's disease in a Chinese population. *Neurobiol Aging* 49:219 e211-219 e213.
2. Kang JF*, et al.* (2016) RAB39B gene mutations are not linked to familial Parkinson's disease in China. *Sci Rep* 6:34502.
3. Guo JF*, et al.* (2012) VPS35 gene variants are not associated with Parkinson's disease in the mainland Chinese population. *Parkinsonism Relat Disord* 18(8):983-985.
4. Shi CH*, et al.* (2011) PLA2G6 gene mutation in autosomal recessive early-onset parkinsonism in a Chinese cohort. *Neurology* 77(1):75-81.
5. Guo JF*, et al.* (2010) Mutation analysis of Parkin, PINK1 and DJ-1 genes in Chinese patients with sporadic early onset parkinsonism. *J Neurol* 257(7):1170-1175.
6. Sun QY*, et al.* (2010) Glucocerebrosidase gene L444P mutation is a risk factor for Parkinson's disease in Chinese population. *Mov Disord* 25(8):1005-1011.
7. Hughes AJ, Daniel SE, Kilford L, & Lees AJ (1992) Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. *J Neurol Neurosurg Psychiatry* 55(3):181-184.
8. DePristo MA*, et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491-498.
9. Woolfe A, Mullikin JC, & Elnitski L (2010) Genomic features defining exonic variants that modulate splicing. *Genome Biol* 11(2):R20.
10. Hebsgaard SM*, et al.* (1996) Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. *Nucleic Acids Res* 24(17):3439-3452.
11. Dogan RI, Getoor L, Wilbur WJ, & Mount SM (2007) SplicePort--an interactive splice-site analysis tool. *Nucleic Acids Res* 35(Web Server issue):W285-291.
12. Desmet FO*, et al.* (2009) Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res* 37(9):e67.
13. Betel D, Koppal A, Agius P, Sander C, & Leslie C (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol* 11(8):R90.
14. Rossin EJ*, et al.* (2011) Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet* 7(1):e1001273.
15. Boyle EA, O'Roak BJ, Martin BK, Kumar A, & Shendure J (2014) MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. *Bioinformatics* 30(18):2670-2672.

16. Nuttle X, Itsara A, Shendure J, & Eichler EE (2014) Resolving genomic disorder-associated breakpoints within segmental DNA duplications using massively parallel sequencing. *Nat Protoc* 9(6):1496-1513.

17. Wang T*, et al.* (2016) De novo genic mutations among a Chinese autism spectrum disorder cohort. *Nat Commun* 7:13316.

18. Purcell S*, et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559-575.

19. Wu B & Pankow JS (2016) On Sample Size and Power Calculation for Variant Set-Based Association Tests. *Ann Hum Genet* 80(2):136-143.

20. Whitworth AJ*, et al.* (2005) Increased glutathione S-transferase activity rescues dopaminergic neuron loss in a Drosophila model of Parkinson's disease. *Proc Natl Acad Sci U S A* 102(22):8024-8029.

21. Yang Y*, et al.* (2006) Mitochondrial pathology and muscle and dopaminergic neuron degeneration caused by inactivation of Drosophila Pink1 is rescued by Parkin. *Proc Natl Acad Sci U S A* 103(28):10793-10798.

22. Huang X, Warren JT, Buchanan J, Gilbert LI, & Scott MP (2007) Drosophila Niemann-Pick type C-2 genes control sterol homeostasis and steroid biosynthesis: a model of human neurodegenerative disease. *Development* 134(20):3733-3742.