Science Translational Medicine
AAAS

stm.sciencemag.org/cgi/content/full/12/560/eaau3960/DC1

Supplementary Materials for

A patient-based model of RNA mis-splicing uncovers treatment targets in Parkinson's disease

Ibrahim Boussaad, Carolin D. Obermaier, Zoé Hanss, Dheeraj R. Bobbili, Silvia Bolognin, Enrico Glaab, Katarzyna Wołyńska, Nicole Weisschuh, Laura De Conti, Caroline May, Florian Giesert, Dajana Grossmann, Annika Lambert, Susanne Kirchen, Maria Biryukov, Lena F. Burbulla, Francois Massart, Jill Bohler, Gérald Cruciani, Benjamin Schmid, Annerose Kurz-Drexler, Patrick May, Stefano Duga, Christine Klein, Jens C. Schwamborn, Katrin Marcus, Dirk Woitalla, Daniela M. Vogt Weisenhorn, Wolfgang Wurst, Marco Baralle, Dimitri Krainc, Thomas Gasser, Bernd Wissinger, Rejko Krüger*

*Corresponding author. Email: rejko.krueger@uni.lu

Published 9 September 2020, *Sci. Transl. Med.* **12**, eaau3960 (2020)
DOI: 10.1126/scitranslmed.aau3960

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at stm.sciencemag.org/cgi/content/full/12/560/eaau3960/DC1)

Data file S3 (Microsoft Excel format). Raw data.

Materials and Methods

Fibroblast culture

Native fibroblasts were obtained from skin biopsies with the written informed consent of donors and cultured as previously described(*17*). Control fibroblast lines C1 and C5-C8 from healthy donors were obtained from the biobank of the Hertie-Institut für klinische Hirnforschung in Tübingen, Germany. Native fibroblasts were used for experiments and also for reprogramming and immortalization. For immortalization, human dermal fibroblasts were transduced with a lentiviral vector expressing SV40 (pLenti-III-SV40; Applied Biological Materials – abm Biotechnology Company) in the presence of polycation Polybrene (ThermoFisher Scientific). Subsequently, cells were incubated for 14 h before the medium was replaced with standard medium. Immortalized fibroblasts were characterized based on morphology and separated from non-immortalized fibroblasts by serial passaging at 1:10 dilutions.

iPSC culture

Control iPSC lines C2 and C3 were previously described and characterized as lines 2132 and 2135(*46*), respectively. Control iPSC line C4 was previously described and characterized as line C1-1(*16*). Fibroblasts of patients with PD were reprogrammed using an adapted protocol(*47*). Fibroblasts were transduced with a mix of retroviral vectors encoding *oct4*, *sox2*, *klf4* and *c-myc*. iPSC were cultured in hES medium [KnockOut DMEM with 20% KnockOut Serum Replacement, 1% MEM Non-Essential Amino Acids solution (100X), 1% GlutaMAX Supplement, 1% penicillin-streptomycin (10,000 U/mL) (all ThermoFisher Scientific), 20 µM β-mercaptoethanol (Carl Roth GmbH) and 10 ng/µl of recombinant human FGF basic (R&D Systems)] on a mouse embryonic fibroblast feeder layer. For differentiation iPSC were passaged

onto Matrigel (Corning Incorporated)-coated plates, and the medium was switched to Essential 8 medium (ThermoFisher Scientific). Midbrain specific dopaminergic (mDA) neurons were derived as previously described(*48*).

iPSC characterization

Silencing of recombinant reprogramming and relative expression of endogenous stem cell markers were analyzed by qPCR in comparison to fibroblasts obtained 4 days post-transduction or native fibroblasts, respectively. Expression was calculated using the amount of the housekeeping gene *HMBS* and the second derivative maximum method. For immunocytochemistry (ICC) of stem cell markers, iPSC were grown on Matrigel-coated coverslips under standard conditions. Cells were fixed in 4% formaldehyde (FA) in PBS for 20 min at room temperature and permeabilized with ice-cold 100% methanol for 5 min at -20°C. Samples were blocked using 10% FCS in PBS for 30 min at room temperature. Primary antibodies were incubated for 1-2 h at 37°C or at 4°C overnight, and secondary antibodies were added for 1 h at room temperature in the dark. Subsequently, DAPI (4',6-diamidino-2-phenylindole, dihydrochloride) (ThermoFisher Scientific) was added for 10 min at room temperature. Washed cover slips were mounted on glass slides using Mowiol/DABCO (Carl Roth GmbH). Karyoptype G-banding of iPSC was performed at the Cytogenetics Research Group at the Institute for Medical Genetics and Applied Genomics of the University Hospital Tübingen.

smNPC culture

Small molecule-derived neural precursor cells (smNPC) were differentiated from iPSC using a previously described protocol(*15*). Neuronal differentiation of smNPC was performed for 30 d following the described (*15*) protocol to generate midbrain-specific dopaminergic neurons.

Generation and analysis of midbrain organoids

The generation of midbrain organoids was performed as previously described(*27*). Briefly, colonies of 9,000 NESCs were plated on ultra-low-attachment 96-well round-bottomed plate (Corning) and cultured in N2B27 medium (DMEM-F12/Neurobasal 50:50 with 1:200 N2 supplement, 1:100 B27 supplement lacking vitamin A, 1% L-glutamine and 1% penicillin-streptomycin, Invitrogen) containing 3 μM CHIR-99021 (Axon Medchem), 0.75 μM purmorphamine (Enzo Life Science), and 150 μM ascorbic acid (Sigma). After eight days, 3D colonies were embedded in droplets of GelTrex (ThermoFisher). At day 10, differentiation was started with N2B27 medium supplemented with 10 ng/mL human brain-derived neurotrophic factor, 10 ng/mL human glial-derived neurotrophic factor, 500 μM dibutyryl cyclic AMP (Peprotech), 200 μM ascorbic acid (Sigma), and 1 ng/mL transforming growth factor β3 (Peprotech). 1 μM purmorphamine (Enzo Life Science) was added to the medium for the first six days of neuronal differentiation only. On day 14, the plates were placed on an orbital shaker (IKA), rotating at 80 rpm, and kept until day 35th. Organoids from the isogenic pair were generated three times.

Organoids were fixed in 4% paraformaldehyde overnight at 4°C and washed three times with PBS for 1h. They were then embedded in 3% low-melting point agarose in PBS. 50 μm sections were cut using a vibratome (Leica VT1000s). Sections were permeabilized with 0.5 % Triton X-100 in PBS and blocked for 90min in 5 % normal goat serum, 2 % BSA, 0.1 % Triton X-100. Sections were incubated on a shaker for 72 h at 4 °C with primary antibodies in the blocking

buffer at the following dilutions: rabbit anti-TH (1:1000, Santa Cruz Biotechnology), anti-chicken Tuj1 (1:1000, Millipore). After incubation with the primary antibodies, sections were washed three times with 0.05% Tween-20 in PBS and incubated for 2h at RT with the secondary antibodies (Invitrogen).

Organoid sections were acquired with an Operetta High-Content Imaging System (Perkin Elmer). 3D images of midbrain organoids were analyzed in Matlab (Version 2017b, Mathworks). The in-house developed image analysis algorithms automate the segmentation of nuclei and neurons, with structure-specific feature extraction. The image preprocessing for the segmentation of nuclei was computed by convolving the raw Hoechst channel with a Gaussian filter. For segmentation of neurons, a median filter was applied to the raw Tuj1 channel. For segmentation of dopaminergic neurons, a median filter was applied to TH raw channel. Skeleton of the resulting TH mask was used to identify nodes and links as total number of branch/end-points and total number of linking elements, respectively. The expression of TH was expressed as positive pixel of the marker, normalized by the pixel count of Hoechst.

Targeted re-sequencing of the PARK7 gene

Targeted re-sequencing of the entire PARK7 gene was performed using a combination of next-generation sequencing (NGS) of overlapping long distance (LD) PCR fragments and Sanger sequencing of two gap closure amplicons for high GC-content segments. Individual LD PCR reactions were performed with 20 pmoles of each primer, 4*150 µM dNTPs, 300 ng genomic DNA and 5U of LongAmp Taq DNA Polymerase (New England Biolabs) in a volume of 50 µl for 36 thermal cycles. Equimolar amounts of the LD PCR products as estimated by side-by-side agarose gel electrophoresis were pooled and fragmented by sonification using a Hielscher UP100H instrument equipped with a microtip (Hielscher Ultrasonics GmbH,). Fragmented DNA

was size-selected by double SPRI applying Agencourt AMPure XP beads (Beckman Coulter GmbH) and then used for NGS library preparation using the GS FLX Titanium Rapid Library Preparation Kit (Roche Diagnostics GmbH). The MID adaptors RL13 or RL14 were used during library preparation for indexing individual libraries. NGS libraries were quantified by qPCR using the KAPA 454 Library Quantification Kit (KAPA Biosystems), and 4E6 molecules were used for emulsion PCR and subsequent purification employing the GS Junior Titanium em-PCR Kit_Lib-L as recommended by the manufacturer (Roche Diangostics GmbH). NGS libraries were pooled, and a total of 5E5 beads were loaded on a GS Junior NGS instrument for 200 sequencing cycles using the GS Junior Titanium Sequencing Kit (Roche Diagnostics GmbH). Data analysis was performed using the instrument's Reference Mapper software, and ambiguities were resolved by manual inspection. High-confidence variant calls were assessed for population frequency using the 1000 Genomes V3 dataset. Rare variants with overall population frequencies below 0.01 were assessed for potential impact on splicing by applying the Human Splicing Finder (http://www.umd.be/HSF/).

Viral transduction

cDNAs of wt *PARK7* and Δex3 *PARK7* were cloned into the lentiviral vector pLL3.7. A construct of the human U1 promotor followed by the sequence of either wt U1 snRNA or G>C U1 snRNA were cloned into the vector pENTR-U6-mPGK-eGFP lacking the U6 promotor. The cloned constructs were recombined with the vector pCDH-DEST-EF1-Puro to create the lentiviral vector plasmids pCDH-U1-hU1(wt)-PGK-eGFP-EF1a-Puro and pCDH-U1-hU1(mut)-PGK-eGFP-EF1a-Puro. Cloning of U1 constructs and packaging of all 4 vectors were performed by SIRION BIOTECH (Martinsried, Germany). Cells were transduced overnight with an M.O.I.

of 10, and stably transduced cells were enriched by flow cytometric sorting of GFP-expressing cells to a purity of >85% using the FACS AriaIII cell sorter (Becton Dickinson and Company).

Minigene assay

Minigene assays are used to analyze splicing of an exon of interest by cloning it into a vector between two given exons with subsequent analysis of the cDNA by RT-PCR. To assess the effect of U1 on c.192G>C *PARK7* splicing, U1 was cotransfected together with the minigene constructs into HEK293 cells. HEK293 cells were transfected at 80 to 90% confluence using the Lipofectamine® 2000 Transfection Reagent (ThermoFisher Scientific) according to the manufacturer's instructions. Cells were transfected with 4 µg minigene plasmid and 4 µg U1 snRNA. RNA isolation and cDNA synthesis were performed as previously described(*49*). Subcloning of RT-PCR products into the pCR2.1 vector (Invitrogen GmbH) was performed according to the manufacturer´s instructions

Western blot

Cell pellets were lysed in PBS/ 1% Triton X-100 supplemented with 1 tablet/50 ml cOmplete, EDTA-free Protease Inhibitor Cocktail (Roche Applied Science) for 20 min on ice and centrifuged for 20 min at 4^0C at 14,000 rpm. The supernatant was collected, and the protein concentration was determined by Bradford assay. Samples were denatured by incubating at 96°C for 5 min in Laemmli buffer. Equal amounts of total protein (15-80 µg) were run on SDS-PAGE gels and subsequently blotted onto a nitrocellulose membrane (ThermoFisher Scientific) for 7 min at 20 V using the iBlot device (Invitrogen GmbH). Primary antibodies used included rabbit anti-DJ-1 (D29E5 XP, Cell Signaling Technology), mouse anti-DJ-1 (mAB E2.1, Invitrogen), rabbit anti-DJ-1 (mAB EP2816y, Abcam), mouse anti-β-actin (A5441, Sigma Aldrich Chemie),

mouse anti-GAPDH (MAB374, Merck KGaA), rabbit anti-LC3B (2775S, Cell Signaling Technology) and mouse anti-mono- and poly-ubiquitinylated conjugates (mAB FK2, Enzo Life Sciences BVBA). Secondary antibodies used included Amersham ECL mouse IgG, HRP-linked whole Ab NA931 and Amersham ECL Rabbit IgG, HRP-linked whole Ab NA934 (GE Healthcare). Proteins were detected after incubation with a 1:1 Amersham ECL Western Blotting Detection Reagent (GE Healthcare) mixture. X-ray films were exposed for 5 s to blots and developed using an X-ray developer (Fujifilm). To increase the sensitivity of detection, membranes were incubated with Amersham ECL Prime Western Blotting Detection Reagent (GE Healthcare) and measured using the ODYSSEY FC Imaging System (LI-COR) and exposure times were set to 30 s to 2 min. Analysis of protein amounts by densitometry was performed using ImageJ (Wayne Rasband, NIH) software.

Compound treatments

To study protein degradation, smNPC were treated with the indicated concentrations of MG132 (SigmaAldrich) or bafilomycin A1 (ENZO Life Sciences) for 16 h overnight prior to lysis. For rescue experiments of full-length *PARK7* mRNA and protein with PB (4-Phenylbutyric Acid, Sigma) and RECTAS (2-chloro-N-(furan-2-ylmethyl)-7H-purin-6-amine, Enamine) in vitro, differentiated neurons and immortalized fibroblasts were cultured under standard conditions. A 1 M PB stock solution was prepared with ethanol, and a 25 mM stock solution of RECTAS was prepared in DMSO. Medium was freshly supplemented with PB and RECTAS as indicated and changed every other day. To avoid loss of efficiency, aliquots of RECTAS were used only once.

Midbrain-specific organoids were treated from day 10 onwards until the end of the experiment at day 35. Medium was freshly supplemented with 1 mM PB and indicated concentrations of RECTAS prior every feeding.

RNA isolation, RT-PCR and qPCR

Total RNA was isolated from cells using the High Pure RNA isolation kit (Roche). cDNA was reverse transcribed using the Transcriptor High Fidelity cDNA Synthesis Kit (Roche). Amplification of both full-length *PARK7* cDNA and Δex3 *PARK7* cDNA by PCR was performed using DJ-1 fwd primer (acgaattcgaatggcttccaaaagagctctggt) and DJ-1 rev primer (ggctccacttgttcttaaagactaggcggccgct). SYBR green qPCR was performed using LightCycler 480 SYBR Green I Master (Roche) with each biological sample run in triplicate. Quantification of *PARK7* full-length or Δex3 *PARK7* cDNA by multiplex qPCR was performed using the LightCycler 480 Probes Master kit (Roche) and hydrolysis probes detecting β-actin and each of the *PARK7* variants in the same reaction. (For detailed information about primers and hydrolysis probes see table S2 and S3.)

Sequencing

Sequencing was performed using the BigDye Terminator v3.1 kit (Applied Biosystems) following the manufacturer's instructions on the ABI 3100 Genetic Analyzer (Applied Biosystems/Ambio) or using the service of Eurofins Genomics.

Subcellular protein and RNA fractionation

Cells from one 100-mm 80% confluent dish were washed twice with cold PBS and harvested. The pellet of cells was resuspended in five volumes of buffer N [15 mM Tris-HCl pH 7.5, 60 mM KCl, 15 mM NaCl, 5 mM $MgCl_2$, 1 mM $CaCl_2$, 1 mM DTT, 0.25 M sucrose, Complete Protease Inhibitor Cocktail (Roche Applied Science)]. Cytoplasmic membrane lysis was obtained by adding an equal amount of buffer N plus 0.4% NP-40. Following 5 min of incubation, nuclei were pelleted and the cytoplasmic fraction recovered. Nuclei were then

washed in 1 ml of solution 1 (10 mM HEPES pH 7.9, 10 mM KCl, 1.5 mM $MgCl_2$, 0.1 mM EGTA, 0.1 mM DTT) and again pelleted and lysed using one volume of solution 2 (10 mM HEPES pH 7.9, 10 mM KCl, 1.5 mM $MgCl_2$, 0.1 mM EGTA, 0.1 mM DTT, 5% glycerol, 0.4 M NaCl). Then, 2 µl and 10 µl were taken from the nuclear and cytoplasmic fractions, respectively, and the quality of fractionation was assessed by Western blot using anti-p84 (Abcam; ab487) and anti-α-tubulin (Merck Millipore; CP06) antibodies. Nuclear and cytoplasmic RNA was extracted using Eurogold Trifast reagent (EuroClone; EMR507200) according to the manufacturer's instructions.

Northern blot assay

Nuclear and cytoplasmic RNA samples were loaded on a 1% formaldehyde agarose gel, transferred to Hybond N+ nylon membranes (Amersham Biosciences # RPN119B) and probed with an internally 32P-labeled *PARK7* exon 4 fragment or *GAPDH* sequence following prehybridization in ULTRAhyb Ultrasensitive Hybridization Buffer (Ambion #AM8670). Pre-hybridization and hybridization were carried out at 54°C. Visualization of transcripts was carried out with a Cyclone Plus Storage Phosphor Scanner and the included OptiQuant Software (Perkin Elmer).

Mass spectrometry

In-gel digestion and peptide concentration determination followed by mass spectrometric analysis were performed according to the protocol published by Plum *et al.*, 2013, with slight modifications(*50*). In total 350 ng of peptides were used for nanoLC -MS/MS analysis with a gradient from 5-40% acetonitrile within 98 min. Mass spectrometric data were analyzed with MaxQuant software (version 1.5.3.30). MS/MS spectra were searched against the

Uniprot/SwissProt human proteome database (UniProtKB/Swiss-Prot UniProt release 2016.05;551,193; downloaded 2015-08-29) using the search engine Andromeda, including 262 common contaminants and concatenated with the reversed versions of all sequences. The precursor and fragment ion mass tolerance were set to 5 ppm and 20 ppm, respectively. The enzyme specificity was set to trypsin, and two missed cleavages were allowed. The minimum peptide length was set to 7 amino acids. Cysteine carbamidomethylation (C) was set as fixed, and methionine oxidation (M) as well as phosphorylation (STY) were set as variable modifications. A maximum of 6 modifications per peptide was set. For both peptide spectrum matches (PSMs) and the protein amount the false discovery rate (FDR) was set to 1%. For the calculation of the protein abundance, label-free quantification (LFQ) was performed with an LFQ minimum ratio count of two. LFQ normalized intensities were used for further data analysis.

Mitochondrial membrane potential (MMP)

Assessment of MMP by fluorescence microscopy was performed on fibroblasts that had been cultured under standard conditions. Cells were stained for 20 min in PBS containing 100 nM Tetramethylrhodamine, methyl ester (TMRM) (ThermoFisher Scientific) and 2 µg/ml Hoechst 33342 Trihydrochloride, Trihydrate (ThermoFisher Scientific) at 37°C and 5% CO_2 prior to imaging. Pictures were analyzed using ImageJ software by applying the “despeckle” function and the “Lookup Tables–Fire” condition, followed by the “analyze particles” option. To measure MMP by flow cytometry, trypsinized cells were stained for 20 min in medium containing 50 nM tetramethylrhodamine, ethyl ester (TMRE) at 37°C and 5% CO_2. Subsequently, the cells were washed and resuspended in PBS/ 50 nM TMRE, and flow cytometric measurements were

performed using the BD LSRFortessa (Becton Dickinson and Company). Analysis of flow cytometry was performed using Flowjo software (Flowjo LLC).

Prokaryotic expression of DJ-1

cDNA of wt *PARK7* and Δex3 *PARK7* were cloned into the pcDNA 3.1/V5-His B vector (Invitrogen). Plasmids were transformed into T7 competent BL21 (DE3) *E. coli* (C2527, NEB). Bacterial cultures were induced with 0.6 mM IPTG overnight at 18°C while shaking at 200 rpm. RNA was extracted by beating the pellet in Trizol (Invitrogen) with glass beads (Sigma). After purification, cDNA was reverse transcribed and amplified with DJ-1 primers as already described. PCR products were then run by high-resolution capillary electrophoresis using the QIAxcel Advanced System and analyzed with QIAxcel® ScreenGel software (Qiagen). For protein extraction, pellets were resuspended in lysis buffer containing 20 mM HEPES (Sigma), 5 mM EDTA (Sigma), 1 M DTT (Carl Roth GmbH), 1 mg/ml Lysozyme (Sigma) and 1 tablet/50 ml cOmplete, EDTA-free Protease Inhibitor Cocktail (Roche Applied Science), freeze/thawed 3 times, incubated for 30 min on ice and centrifuged for 40 min at 15,000 rpm. The supernatant was used to perform Western blotting with 10 µg of total protein.

In vitro transcription and translation of DJ-1

The same pcDNA 3.1-DJ-1 plasmids as in the prokaryotic expression experiments were used for in vitro translation of DJ-1. The plasmids were linearized by XhoI digest and used for in vitro transcription and capping of mRNA following the manufacturer's instructions of the mMESSAGE mMACHINE T7 ULTRA Transcription Kit (ThermoFisher). 1 µg of in vitro transcribed mRNA was used per in vitro translation reaction using the Retic Lysate IVT Kit (ThermoFisher). The translation was performed according to manufacturer's instructions for

reactions with a final potassium concentration of 125 mM. 5 µl of each reaction was used in SDS-page and subsequent Western blot analysis.

RNA secondary structure analysis

Secondary structure predictions for native wt *PARK7* mRNA and Δex3 *PARK7* mRNA were generated using the RNAfold software (the "avoid isolated base pairs" option was activated, and default parameters were used for all other settings). The predicted centroid secondary structures, for example the structures with minimal base-pair distance to all structures in the thermodynamic ensemble prediction, were visualized using the dot-bracket notations of the structures as input for the PseudoViewer software (see Fig. 5D; paired bases are highlighted in blue, unpaired bases in yellow). To evaluate the stability of the predicted structures, the total minimum free energy estimates were compared, and the relative number of unpaired bases was determined.

Gene correction in human fibroblasts

The targeting vector for homologous recombination contains a 2.4 kb 5´-homology arm including the corrected exon 3, a loxP and FRT-flanked neomycin resistance cassette and a 1.9 kb 3´-homology arm. Two different sgRNA sequences, 5'-AGACGTTTGTAATCCATACA-3' and 5'-ACATCACGGCTACACTGTAC-3', were cloned in between two BbsI sites as complementary oligonucleotides into a plasmid containing a U6 promoter (pBS-U6) and the sgRNA backbone V1.0 based on Jinek *et al.*(*51*). Patient-derived dermal fibroblasts were cultured on Corning Matrigel-coated plates in DMEM (GIBCO, Invitrogen) supplemented with 10% FBS (Sigma-Aldrich), penicillin (100 U/ml) and streptomycin (100 µg/ml). Nucleofection was performed with the 4D-Nucleofector X Unit using the P2 Primary Cell 4D-Nucleofector X Kit L and Program C DS-150 according to the manufacturer's protocol (Lonza, Basel,

Switzerland). A total of 0.5 µg of each sgRNA plasmid, 0.25 µg of the targeting vector and 0.25 µg of pCAG-Cas9-D10A expressing spCas9 nickase based on Cong *et al.*(*52*) were used. After nucleofection, cells were incubated in RPMI (GIBCO) for an additional 10 min at 37°C; after 2 days of recovery, cells were selected with G418 (50 µg/ml, GIBCO) for 8 days and subsequently seeded as single cells in a 96-well format. Gene-corrected clones were identified via PCR using the following primer pairs: 5'-GGAGACGGTCATCCCTGTAG-3' and 5'-TCCAGACTGCCTTGGGAAAA-3' for correct 5' integration, and 5'-TCGCCTTCTTGACGAGTTCT-3' and 5'-GGTCCAGCAATCCCACTACT-3' for correct 3' integration. Correct clones were confirmed by Sanger sequencing and Western blot analysis.

Gene editing of iPSC

Insertion of the the c.192G>C mutation into the control line C4 via CRISPR/Cas9 technology was performed by Applied StemCell, Inc., Milpitas, CA, USA.

Burden analysis

Discovery cohort (PPMI)

Data

The Parkinson's Progression Markers Initiative (PPMI) study is an effort to identify the biomarkers of PD progression. We used the whole exome sequencing (WES) data available as part of this project. Detailed information about this initiative and the data can be found on the project website (http://www.ppmi-info.org/). Briefly, the variants were called following GATK(*53*) best practices by the authors of the original study. The data was obtained in the form of a Variant Call Format file (VCF).

Pre-processing

Sample QC

Samples with >3 standard deviation (SD) from the QC metrics (number of alternate alleles, number of heterozygotes, Ti/Tv ratio, number of singletons and call rate) that were calculated by using PLINK/SEQ i-stats (https://atgu.mgh.harvard.edu/plinkseq/) were excluded from the analysis. For population stratification we selected the variants that were common between our dataset and hapmap version 3.30(*54*), present in autosomal chromosomes, not in linkage equilibrium, call rate > 80%, allele frequency > 5 % and Hardy-Weinberg equilibrium P-value < 0.001 and used PLINK(*55*) multi-dimensional scaling (mds)(*56*) to identify outliers. Each sample that was > 3 SD of the first and the second principal components was considered as ethnicity outlier and excluded from further analyses. By using the same set of variants as described above, relatedness check was performed up to second degree applying PLINK(*55*) and KING(*57*) algorithms. From the identified related sample pairs one sample was chosen randomly to be included in the final analyses.

Variant QC

Multi-allellic variants were decomposed by using variant-tests(*58*) and left normalized by bcftools(*59*). The authors of the PPMI study used the variant quality score recalibration (VQSR) method as recommended by GATK best practices(*53*) to filter out low quality variants. Additionally, we used GATK hard filtering to select only high quality SNVs. Variant genotypes with a read depth (DP) < 10 and genotype quality (GQ) < 20(*60*) were converted to missing by using bcftools(*59*) and only variants with a call rate of > 0.9 were kept for further analyses.

Variant annotation and filtering

As the current study is focused on U1 splice site variants we restricted our further analyses to the 5' consensus splice site positions, +3 to -6 from the exon/intron boundary. The exon-intron intervals were obtained from the UCSC table browser based on hg19 reference genome. Variants were annotated by using ANNOVAR(*61*) version 2016December05 using RefSeq gene annotations and the dbNSFP v3.0(*45*) prediction scores. Only rare variants(*62*), as defined by variants with a minor allele frequency of < 5% in the European population of 1000 genomes(*63*), ExAC (NFE (non-finnish Europeans), release 0.3)(*64*), and the Exome variant server (http://evs.gs.washington.edu/EVS) were selected. In order to prioritize the 5' splice variants based on their deleteriousness, we used three different scores. The first score is generated by using the MaxEntScan method(*43*) which is based on the maximum entropy principle. The other two scores were ensemble scores (*dbscSNV_ADA* and *dbscSNV_RF*) generated from multiple splice site prediction tools(*44*) which are available as part of dbNSFP database(*45*).

Generation of MaxEntScan score

To prioritize variants using MaxEntScan method, for each SNV that lies in the consensus splice site region a wild type 9 mer (WT) was extracted from the reference genome (hg19). Then, the variant was introduced within the WT sequence by using the python module pyfaidx(*65*), hence creating a mutated consensus splice site (MUT) sequence for each variant. In the next steps, the scores were calculated for both WT and MUT sequences by using the scripts provided in the MaxEntScan website (http://genes.mit.edu/burgelab/maxent/Xmaxentscan_ scoreseq.html). The relative percentage change (*maxentscan_change*) was calculated by using (*44*, *66*):

$$maxentscan\_change = \left(\frac{wild\_score - mut\_score}{wild\_score}\right) * 100$$

Benchmarking of MaxEntScan score

We were interested in the highly deleterious splice variants and, in line with our hypothesis, one recent study(*67*) has shown that, 21 variants out of 30 variants tested within BRCA1 genes were predicted by MaxEntScan method and were later confirmed by the functional validation. Out of the 21 variants that were predicted to be deleterious 18 of them had a *wild_score* > 5 and a *maxentscan_change* > 70. In order to benchmark our methods and determine reliable cut-offs, we used two datasets: 1) The professional version of Human Gene Mutation Database (HGMD)(*68*) version February 2017, and 2) gnomAD (*64*), which is variant data from 123,136 exome sequences and 15,496 whole-genome sequences from individuals which were sequenced as part of various disease-specific and population genetic studies.

We only selected the variants annotated as high confidence and pathogenic ("DM" flag) in HGMD (HGMDpatho) variants. VCF files were generated for HGMD and gnomAD datasets for only those variants that were present within the U1 consensus splice regions annotated in a similar way as we did for the discovery cohort. Density plots based on various scores were generated for HGMDpatho variants and gnomAD splice variants (fig. S7A, C).

Determination of cut-offs for *wild_score* and *maxentscan_change*

In the current study, we were interested in variants having a high likelihood causing splice changes. For the HGMDpatho variants a clear separation was observed in the distribution of majority of variants at a wild_score of 5 (fig. S7A) and at a maxentscan_change of 70% (fig. S7B). Whereas, a reversed distribution could be seen for the gnomAD variants (fig. S7C, D). HGMDpatho variants showed *dbscnv_RF* and *dbscnv_ADA* scores of > 0.9 (fig. S7E, F). Based on the above inferences and the results based on a previous study(*67*), we choose the following cut-offs for further processing: Deleterious splice site variants (DEL.splicing) were defined as SNVs with the following criteria: *wild_score* > *5* and *maxentscan_change* > *70* and

dbscSNV_ADA score > 0.9 and *dbscSNV_RF score > 0.9*. If the ensemble scores were not available for any particular variant only MaxEntScan method (*wild_score > 5* and *maxentscan_change > 70*) was used.

Splice site burden tests

The *wild_score* generated from the wild type 9mer by MaxEntScan is used to identify a true splice site. The higher the *wild_score* the higher the probability of being a true splice site(*44*). We separated the variants into different classes: 1) All the deleterious splicing variants (DEL.splicing), 2) DEL.splicing variants in coding regions (DEL.exonic.splicing), 3) DEL.splicing variants within intronic regions (DEL.intronic.splicing), 4) DEL.exonic.splicing variants present in the genes that are expressed in brain (DEL.exonic.brain.splicing), 5) DEL.exonic.splicing variants present in the genes that are not expressed in brain (DEL.exonic.nonbrain.splicing), and 6) rare synonymous variants as a negative class. We used a previously published list of brain expressed genes(*69*) to test if there is an increased burden in brain expressed genes (n= 14,177) compared to the non-brain expressed genes (n=6,428). Our hypothesis was that cases carry a higher number of DEL.splicing variants compared to the controls. For each variant class a VCF file was generated and the variant counts per sample was calculated by using bcftools(*59*) stats command.

We performed burden testing by constructing the generalised linear regression models(*70–73*) using R version 3.4.1 while correcting for various confounding factors for each sample such as: 1) Sex 2) total number of variants remaining after final QC, 3) TiTv ratio of other variants relative to the dbSNP version 138(*74*), 4) TiTv ratio of variants present in dbSNP version 138, 5) heterozygous variants to homozygous variants ratio, 6) first ten principal components derived from the multi-dimensional scaling.

Replication cohort (PDGSC)

Data

We used the WES data available as part of the ongoing Parkinson's Disease Genome Sequencing Consortium (PDGSC) project. The PDGSC dataset is an effort to integrate PD WES data generated from multiple studies across different sequencing centres. The variant calling was performed by the consortium using GATK best practices version 3.4. Similar to the discovery cohort, we obtained the data in the form of VCF file. Since the PPMI samples are also part of the PDGSC cohort, all samples overlapping between PPMI and PDGSC were excluded in beforehand from the PDGSC dataset. PPMI samples within PDGSC were identified based on their sample ids as well as using relatedness test (see above).

Sample QC

Sample QC was performed by the PDGSC consortium. Briefly, samples were excluded based on the following parameters: 1) < 15x mean coverage 2) discordance between genetic and reported sex, 3) < 85% call rate, 4) outliers for various parameters such as variant counts (all, non-reference genotypes, hets, singletons, mean minor allele rates), TiTv ratio, mean quality scores for non-reference variants and mean depth for non-reference variants, 5) heterozygosity outliers (-0.1 < F <0.1) , 6) ancestry outliers > 6 SD from means of CEU and TSI for PC1 and PC2 , 7) extract probands randomly from pairs related at > 12.5% and 8) exclude samples < 18 years of age or with missing age data.

Variant QC

Similar to the discovery cohort a VQSR filtering method was employed by the authors of original study. In addition, we used the same filtering procedure as described above for the

discovery cohort with one difference in the threshold for call rate. As the data was generated at multiple centres by employing different sequencing protocols we might lose true positive variants if we would filter too stringently leading to loss of statistical power ultimately. Hence, we used a less stringent, although a standard threshold(*75*) of call rate > 0.8 for a variant to be included in the analysis.

Variant detection and annotation

Variants were annotated and splice variants were scored using the same procedure as for the discovery cohort.

Burden testing

We employed the same procedure for burden testing by adjusting for all the covariates that were described above for discovery cohort. In order to further adjust for study wide differences, we used the total number of sites that were fully called within each sample as an additional covariate along with the other covariates. This approach allowed us to account for any exome-wide biases arising due to different sequencing protocols that were employed at different sequencing centres and other confounding factors arising from technical differences(*70*). The same can also be noted from the fact that there is no statistically significant difference between the number of synonymous variants (neutral variants) between cases and controls (Fig. 7).

Multiple testing adjustment

The P-values from burden analysis of both discovery and replication cohorts were adjusted for multiple testing by the function “p.adjust” (R version 3.4.1) using the false discovery rate (FDR) method for discovery and replication cohort separately.

Literature mining

Biomedical literature contains wealth of information about functional relations between various concepts - for example proteins, chemicals, small molecules, phenotypes, diseases and more. Relation types depend on the concepts involved, and are meant to express how the concepts are connected to each other. For example, which proteins participate in a regulatory event or how certain mutation affects a pathological process.

We have built a pipeline for processing publicly available biomedical text, abstracts, full text articles, conference proceedings, books and electronic health records, starting from searching the web and downloading raw files, to extraction and storing of concepts (entities) and semantic relations between them (events) into a knowledge base.

Our text mining analysis consists of the following steps:

- Concept identification. This process is known as "named entity recognition" (NER), and aims at labelling words and phrases as proteins, chemicals, diseases, etc. We use Reflect tool(*76*) for this purpose.
- Finding of lexical patterns which may indicate an event. For example, phrase like “activation of” or its synonyms ”activate(s) or activated (by)” would indicate an event of positive regulation if at least one of its arguments is a protein. We have compiled a dictionary of such patterns based on a corpus of manually annotated biomedical articles(*77*).
- Morphological and syntactic analysis of the text. The goal of this stage is to determine syntactic dependencies between words and phrases in the sentence. In particular, types of syntactic dependencies between entities and event patterns are important. We use Genia tagger(*78*) and Stanford dependency parser(*79*) at this stage of the pipeline.

- Semantic interpretation during which syntactic dependencies between the concepts and patterns are mapped onto functional relations between them. We have developed a set of rules which combine syntactic dependencies, concept and event pattern types, and which guide the mapping. Each relation is expressed as a subject-predicate-object triplet. For example, in "Phenylbutyrate up-regulates the DJ-1 protein ...", the subject is "phenylbutyrate", the object is "DJ-1", and the predicate is "up-regulates". This triplet describes a positive regulation of a protein DJ-1, caused by a chemical phenylbutyrate".

The interactions discovered by our text mining system are stored in a database in the so called triple store or RDF format. The database can be accessed automatically by a software using an API or searched by a human using a web interface. While searching one can specify interaction type and / or terms of interest. For this particular work we searched the database for the interactions of type "Regulation" / "Positive regulation" with "chemical" as a cause, and "protein" as an object. By the time of this specific search (May, 2016) we have extracted about 2000 compounds involved in regulation of protein expression. We filtered the list by selecting compounds which have been mentioned in at least 10 different articles and sorted them by the number of different proteins whose expression they increased. This step reduced the list of compound candidates to 97. We then manually inspected the list of all the sentences describing protein (up)regulation by the selected compounds, giving preference to interactions involving DJ-1 (*PARK7*), neurodegenerative diseases, and in particular the Parkinson's disease.

Additional authors note

PDGSC

PDGSC (Parkinson Disease Genetic Sequencing Consortium) is a collaborative group of investigators working in the area of PD genetics through the analysis of high content sequencing. The PDGSC has been supported by The Michael J. Fox Foundation for Parkinson's Research, the National Institute on Aging Intramural Research Program, and the National Institute of Neurological Disorders and Stroke. A full list of the participants support is provided below.

The members of the consortium are: Marco Abreu (Indiana University School of Medicine, USA), Gary W. Beecham (John P. Hussman Institute for Human Genomics, University of Miami, USA), Sara Bandres-Ciga (National Institute on Aging, USA), Cornelis Blauwendraat (National Institute on Aging and National Institute of Neurological Disorders and Stroke, USA), Jose Bras (University College London, UK), Alexis Brice (Brain and Spine Institute (ICM), France), Kathrin Brockmann (University of Tübingen/DZNE, Germany), Zeynep Akdemir (Baylor College of Medicine, USA), Patrick F Chinnery (University of Cambridge, UK), Jean-Christophe Corvol (Brain and Spine Institute (ICM), France), Fabrice Danjou (Brain and Spine Institute (ICM), France), Aaron Day-Williams (MRL, Merck & Co., Inc., Boston, MA, USA), John D Eicher (MRL, Merck & Co., Inc., Boston, MA, USA), Karol Estrada (Biogen, USA), Daniel M Evans, Faraz Faghri (National Institute on Aging, USA, University of Illinois at Urbana-Champaign, USA), Samuel Evetts (University of Oxford, UK), Ilaria Guella and Matthew J Farrer (Centre for Applied Neurogenetics, University of British Columbia, Canada),

Tatiana Foroud (Indiana University School of Medicine, USA), Steve Finkbeiner (Gladstone Institutes/UCSF, USA), Thomas Gasser (University of Tübingen/DZNE, Germany), J Raphael Gibbs (National Institute on Aging, USA), John Hardy (University College London, UK), MTM Hu (University of Oxford, UK), Joseph Jankovic (Baylor College of Medicine, USA), Hallgeir Jonvik (University College London, UK), Demis A Kia (University College London, UK), Christine Klein (Institute of Neurogenetics, University of Luebeck, Germany), Rejko Krüger (Luxembourg Centre for Systems Biomedicine, Luxembourg), Dongbing Lai (Indiana University School of Medicine, USA), Suzanne Lesage (Brain and Spine Institute (ICM), France), Christina M. Lill (Institute of Neurogenetics, University of Luebeck, Germany), Steven J. Lubbe (Ken and Ruth Davee Department of Neurology, Northwestern University, Chicago, IL, USA), Timothy Lynch (Dublin Neurological Institute, Mater Misericordiae University Hospital, Ireland), Kari Majamaa (University of Oulu, Finland), Eden R. Martin (John P. Hussman Institute for Human Genomics, University of Miami, USA), Patrick May (Luxembourg Centre for Systems Biomedicine, Luxembourg), Brit Mollenhauer (University Medical Center Goettingen, Department of Neurology, Germany), David Murphy (University College London, UK), Huw R Morris (University College London, UK), Mike A Nalls (National Institute on Aging, USA, Data Tecnica International, USA), Khanh-Dung Nguyen (Biogen, USA), Karen Nuytemans (John P. Hussman Institute for Human Genomics, University of Miami, USA), Lasse Pihlstrom (Oslo University Hospital, Norway), Alan Pittman (University College London, UK), Lea R'Bibo (University College London, UK), Laurie Robak (Baylor College of Medicine, USA), Owen A. Ross (Mayo Clinic Jacksonville, USA), Cynthia Sandor (University of Oxford, UK), Barbara Schormair (Institute of Neurogenomics, Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), Munich, Germany), William K. Scott

(John P. Hussman Institute for Human Genomics, University of Miami, USA), Manu Sharma (Centre for Genetic Epidemiology, Institute for Clinical Epidemiology,University of Tubingen, Germany), Joshua M Shulman (Baylor College of Medicine, USA), Ari Siitonen (University of Oulu, Finland), Javier Simón-Sánchez (University of Tübingen/DZNE, Germany), Andrew B Singleton (National Institute on Aging, USA), David J Stone (MRL, Merck & Co., Inc., West Point, PA, USA), Konrad Szewczyk-Krolikowski (University of Oxford, UK), Manuela MX Tan (University College London, UK), Paul Tomlinson (University of Oxford, UK), Mathias Toft (University of Oslo, Norway), Richard Wade-Martins (University of Oxford, UK), Claudia Trenkwalder (Dept. Neurosurgery, University Medical Center, Goettingen, Germany), Caleb Webber (University of Oxford, UK), Wei Wei (University of Cambridge, UK), Jeffery M. Vance (John P. Hussman Institute for Human Genomics, University of Miami, USA), Nigel M Williams (Cardiff University, UK), Juliane Winkelmann (Institute of Neurogenomics, Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), Munich, Germany), Zbigniew K. Wszolek (Mayo Clinic Jacksonville, USA), Pauli Ylikotila (University of Turku, Finland), Alexander Zimprich (Department of Neurology, Austria).

For correspondence regarding PDGSC please contact: Andrew B Singleton, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, USA; singleta@mail.nih.gov

The work of the PDGSC was supported by: The Intramural Research Program of the National Institute on Aging, National Institutes of Health, part of the Department of Health and Human

Services (ZO1 AG000949), The Extramural Research Program of the National Institutes of Health (R01NS096740, R01NS037167, R01NS078086, P50NS071674, P50NS072187), The Michael J. Fox Foundation for Parkinson's Research, The Department of Defense (USAMRAA, the Mitochondrial Disease (W81XWH-17-1-0249) and the Parkinson's Research Program managed through CDMRP), the National Institute of Neurological Disorders and Stroke, the Canadian Consortium for Neurodegeneration in Aging, the Canada Excellence Research Chairs program, the Canadian Institutes of Health Research, the Hermann and Lilly Schilling Foundation, the German Research Foundation (FOR2488), the German Federal Ministry of Education and Research (BMBF) under the funding code 031A430A and the EU Joint Programme -Neurodegenerative Diseases Research (JPND) project under the aegis of JPND (www.jpnd.eu) through Germany, BMBF, funding code 01ED1406, Sigrid Juselius Foundation, the Medical Research Council UK (MC_UP_1501/2 & 13044), Parkinson's UK (grants 8047, J-0804, G-1502), the Medical Research Council UK (G0700943, G1100643) the Wellcome Trust (101876/Z/13/Z), the Fonds National de Recherche (FNR; NCER-PD) Luxembourg.

Disclosure statement

Dr Mike A. Nalls' participation is supported by a consulting contract between Data Tecnica International LLC and the National Institute on Aging, NIH, Bethesda, MD, USA. Dr Nalls also consults for Genoom Health, Illumina Inc, The Michael J. Fox Foundation for Parkinson's Research and University of California Healthcare among others.

The PDGSC data set did also include controls from the Alzheimer's Disease Sequencing Project (ADSP).

The Alzheimer's Disease Sequencing Project (ADSP) is comprised of two Alzheimer's Disease (AD) genetics consortia and three National Human Genome Research Institute (NHGRI) funded Large Scale Sequencing and Analysis Centers (LSAC). The two AD genetics consortia are the Alzheimer's Disease Genetics Consortium (ADGC) funded by NIA (U01 AG032984), and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) funded by NIA (R01 AG033193), the National Heart, Lung, and Blood Institute (NHLBI), other National Institute of Health (NIH) institutes and other foreign governmental and non-governmental organizations. The Discovery Phase analysis of sequence data is supported through UF1AG047133 (to Drs. Schellenberg, Farrer, Pericak-Vance, Mayeux, and Haines); U01AG049505 to Dr. Seshadri; U01AG049506 to Dr. Boerwinkle; U01AG049507 to Dr. Wijsman; and U01AG049508 to Dr. Goate and the Discovery Extension Phase analysis is supported through U01AG052411 to Dr. Goate, U01AG052410 to Dr. Pericak-Vance and U01 AG052409 to Drs. Seshadri and Fornage. Data generation and harmonization in the Follow-up Phases is supported by U54AG052427 (to Drs. Schellenberg and Wang).

The ADGC cohorts include: Adult Changes in Thought (ACT), the Alzheimer's Disease Centers (ADC), the Chicago Health and Aging Project (CHAP), the Memory and Aging Project (MAP), Mayo Clinic (MAYO), Mayo Parkinson's Disease controls, University of Miami, the Multi-Institutional Research in Alzheimer's Genetic Epidemiology Study (MIRAGE), the National Cell Repository for Alzheimer's Disease (NCRAD), the National Institute on Aging Late Onset Alzheimer's Disease Family Study (NIA-LOAD), the Religious Orders Study (ROS), the Texas Alzheimer's Research and Care Consortium (TARC), Vanderbilt University/Case Western Reserve University (VAN/CWRU), the Washington Heights-Inwood Columbia Aging Project

(WHICAP) and the Washington University Sequencing Project (WUSP), the Columbia University Hispanic- Estudio Familiar de Influencia Genetica de Alzheimer (EFIGA), the University of Toronto (UT), and Genetic Differences (GD).

The CHARGE cohorts are supported in part by National Heart, Lung, and Blood Institute (NHLBI) infrastructure grant HL105756 (Psaty), RC2HL102419 (Boerwinkle) and the neurology working group is supported by the National Institute on Aging (NIA) R01 grant AG033193. The CHARGE cohorts participating in the ADSP include the following: Austrian Stroke Prevention Study (ASPS), ASPS-Family study, and the Prospective Dementia Registry-Austria (ASPS/PRODEM-Aus), the Atherosclerosis Risk in Communities (ARIC) Study, the Cardiovascular Health Study (CHS), the Erasmus Rucphen Family Study (ERF), the Framingham Heart Study (FHS), and the Rotterdam Study (RS). ASPS is funded by the Austrian Science Fond (FWF) grant number P20545-P05 and P13180 and the Medical University of Graz. The ASPS-Fam is funded by the Austrian Science Fund (FWF) project I904),the EU Joint Programme - Neurodegenerative Disease Research (JPND) in frame of the BRIDGET project (Austria, Ministry of Science) and the Medical University of Graz and the Steiermärkische Krankenanstalten Gesellschaft. PRODEM-Austria is supported by the Austrian Research Promotion agency (FFG) (Project No. 827462) and by the Austrian National Bank (Anniversary Fund, project 15435. ARIC research is carried out as a collaborative study supported by NHLBI contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). Neurocognitive data in ARIC is collected by U01 2U01HL096812, 2U01HL096814, 2U01HL096899, 2U01HL096902, 2U01HL096917

from the NIH (NHLBI, NINDS, NIA and NIDCD), and with previous brain MRI examinations funded by R01-HL70825 from the NHLBI. CHS research was supported by contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the NHLBI with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629, R01AG15928, and R01AG20098 from the NIA. FHS research is supported by NHLBI contracts N01-HC-25195 and HHSN268201500001I. This study was also supported by additional grants from the NIA (R01s AG054076, AG049607 and AG033040 and NINDS (R01 NS017950). The ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F4-2007-201413 by the European Commission under the programme "Quality of Life and Management of the Living Resources" of 5th Framework Programme (no. QLG2-CT-2002-01254). High-throughput analysis of the ERF data was supported by a joint grant from the Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043). The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, the Netherlands Organization for 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. Genetic data sets are also supported by the Netherlands Organization of Scientific Research NWO Investments (175.010.2005.011, 911-03-012), the Genetic

Laboratory of the Department of Internal Medicine, Erasmus MC, the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), and the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA), project 050-060-810. All studies are grateful to their participants, faculty and staff. The content of these manuscripts is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the U.S. Department of Health and Human Services.

The four LSACs are: the Human Genome Sequencing Center at the Baylor College of Medicine (U54 HG003273), the Broad Institute Genome Center (U54HG003067), The American Genome Center at the Uniformed Services University of the Health Sciences (U01AG057659), and the Washington University Genome Institute (U54HG003079).

Biological samples and associated phenotypic data used in primary data analyses were stored at Study Investigators institutions, and at the National Cell Repository for Alzheimer's Disease (NCRAD, U24AG021886) at Indiana University funded by NIA. Associated Phenotypic Data used in primary and secondary data analyses were provided by Study Investigators, the NIA funded Alzheimer's Disease Centers (ADCs), and the National Alzheimer's Coordinating Center (NACC, U01AG016976) and the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA, and at the Database for Genotypes and Phenotypes (dbGaP) funded by NIH. This research was supported in part by the Intramural Research Program of the National Institutes of health,

National Library of Medicine. Contributors to the Genetic Analysis Data included Study Investigators on projects that were individually funded by NIA, and other NIH institutes, and by private U.S. organizations, or foreign governmental or nongovernmental organizations.

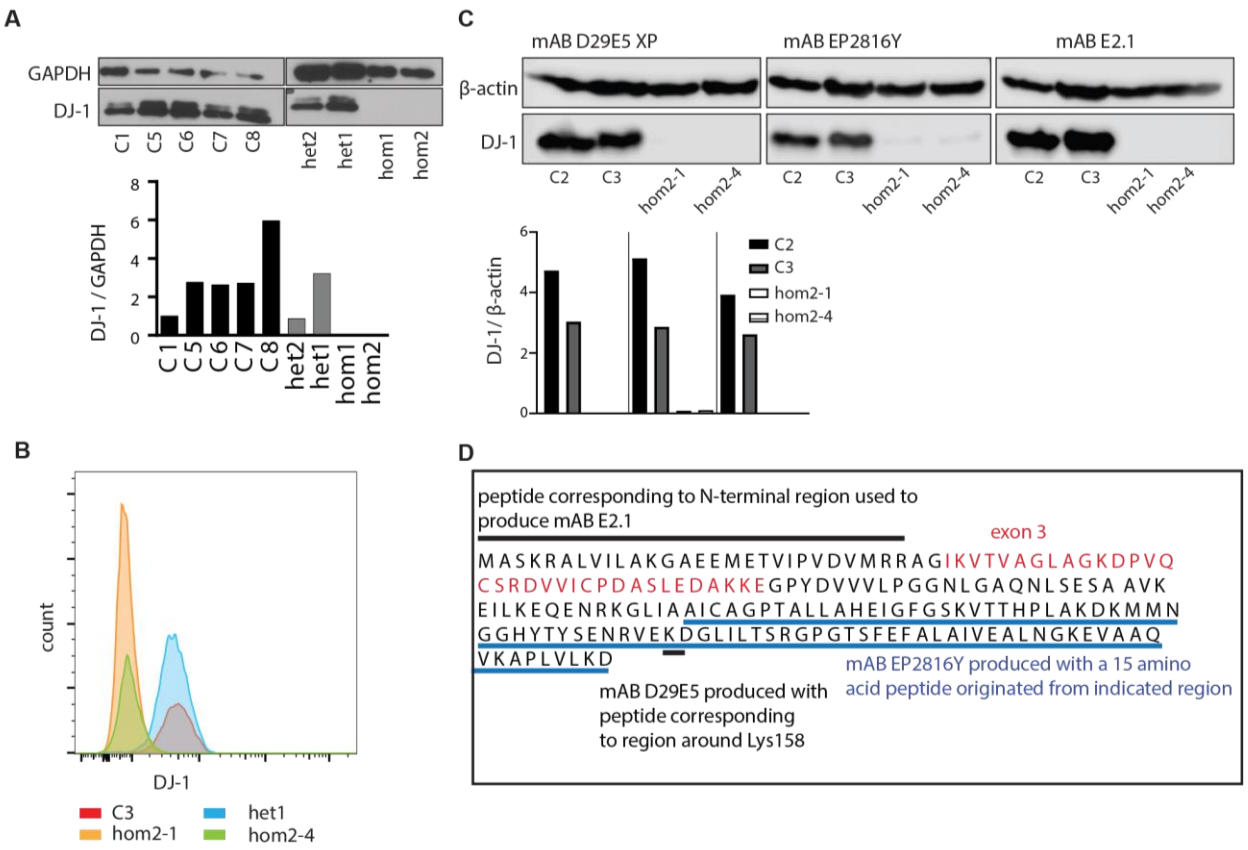


Fig. S1. Loss of DJ-1 protein in homozygous c.192G>C mutation carriers. (**A**) Comparison of DJ-1 amounts in fibroblasts by Western blot. (top) GAPDH and DJ-1 immunoblot of lysates from healthy control fibroblasts (left panels) and from fibroblasts of heterozygous and homozygous c.192G>C mutation carriers (right panels). (bottom) Quantification of the Western blot. (**B**) Flowcytometric analysis of DJ-1 content in smNPC derived from two independent iPSC of the index patient, one iPSC clone of individual het1 and control iPSC C3. (**C**) Immunoblots of the indicated smNPC lines comparing control lines with patient-derived lines using three different anti-DJ-1 monoclonal antibodies (mAB) from Cell Signaling (mAB D29E5 XP, left panels), Abcam (mAB EP2816Y, middle panels) and Invitrogen (mAB E2.1, right panels). (**D**) Regions corresponding to the peptides that were used to generate the three mAB are highlighted in the DJ-1 amino acid sequence, with exon 3 highlighted in red letters.

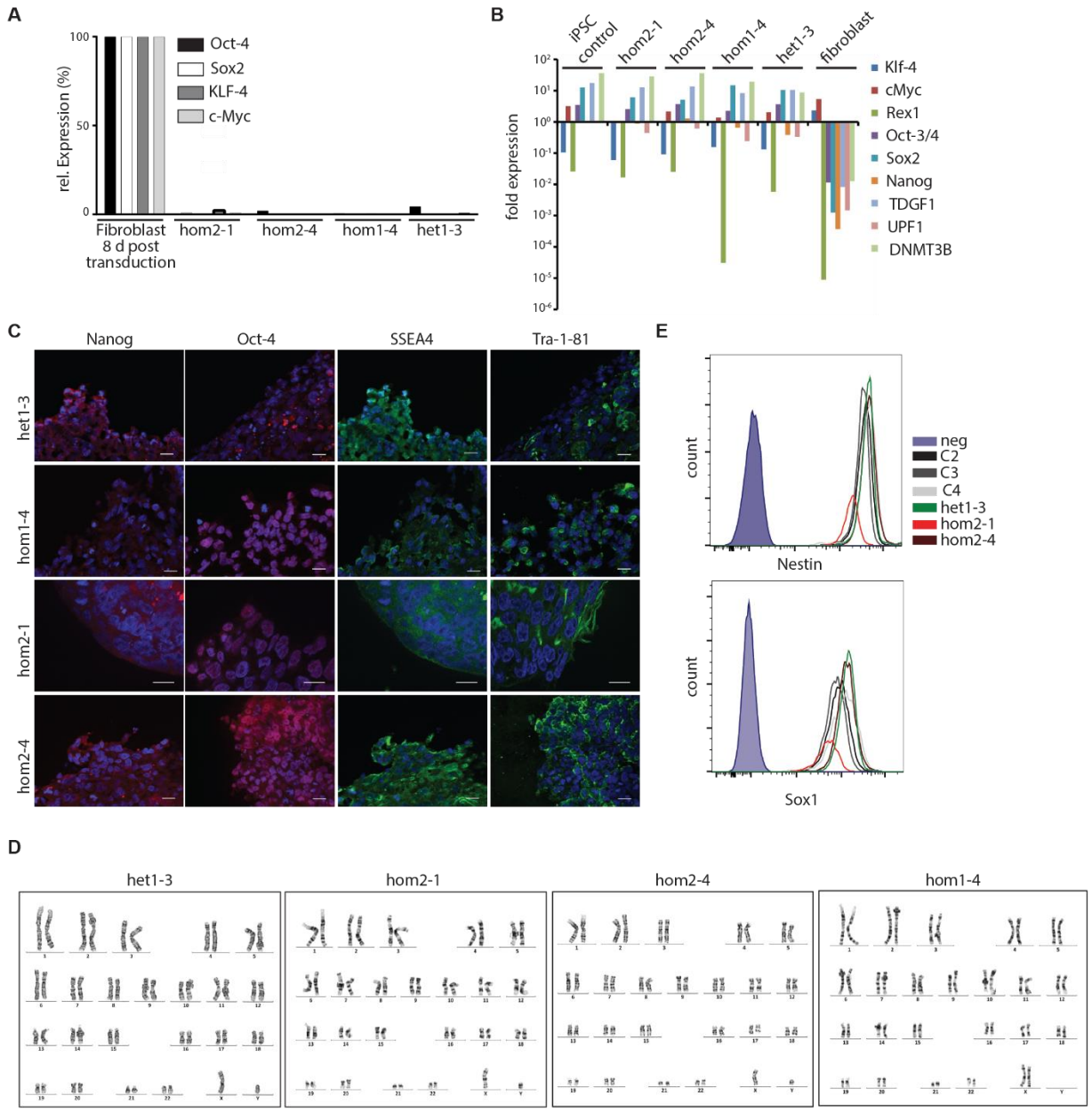


Fig. S2. Characterization of iPSC and smNPC. (**A**) Silencing of retroviral vectors in established iPSC clones was confirmed by SYBR green qPCR analysis with vector-specific primers. Expression was calculated using the housekeeping gene *HMBS* and compared to the amounts in fibroblasts that had been transduced 8 d prior to RNA extraction. (**B**) Expression of endogenous pluripotency markers in established iPSC clones in comparison to a fibroblast line were detected by SYBR green qPCR and calculated using the housekeeping gene HMBS. (**C**) Immunocytochemistry of established iPSC clones stained for nuclear DNA with DAPI and Nanog (red, first column), Oct-4 (red, 2nd column), SSEA4 (green, 3rd column) and TRA-1-81 (green, last column). (**D**) G-banding of established iPSC clones. (**E**) Expression of the neural precursor markers Nestin (left) and Sox1 (right) detected by flow cytometry in the indicated smNPC lines differentiated from iPSC.

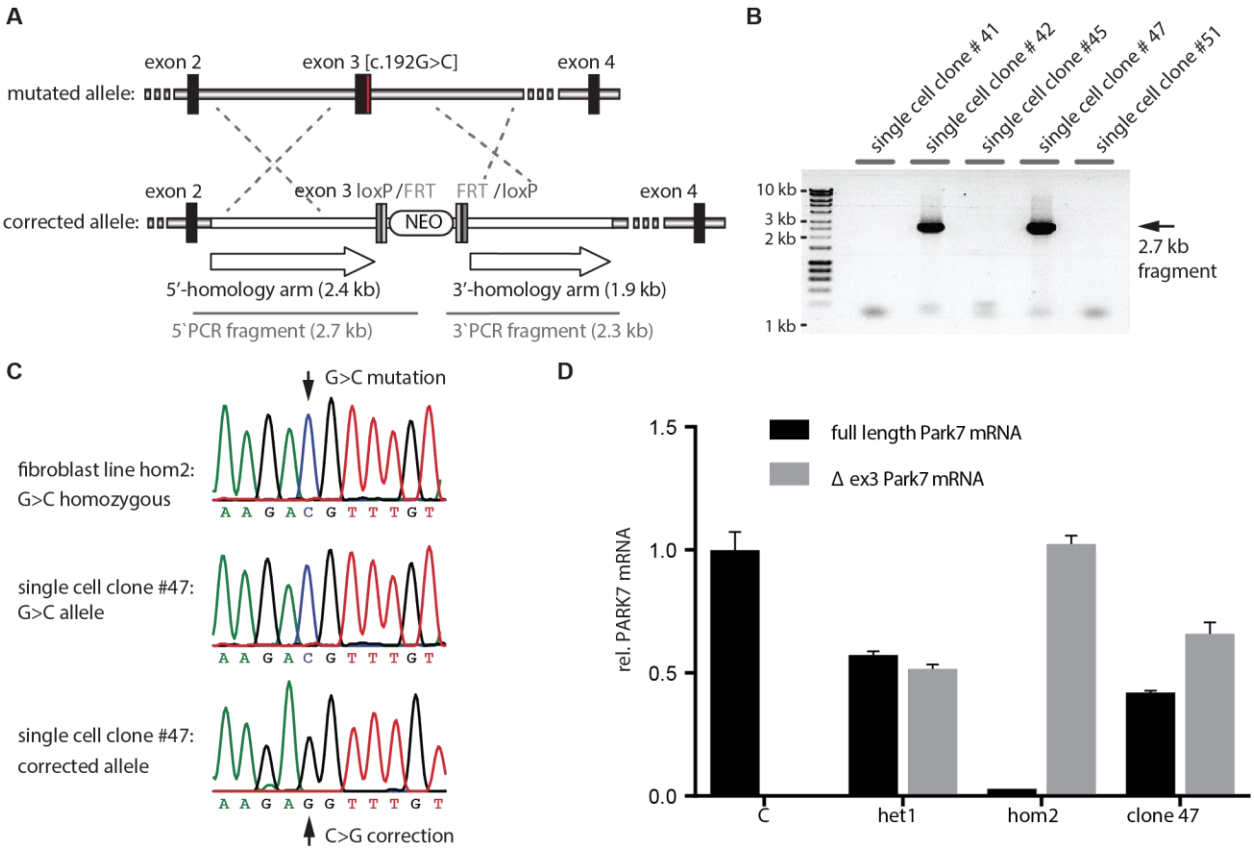


Fig. S3. Generation of gene-corrected fibroblasts. (**A**) Schematic of the targeted genomic region (upper graph) containing Park7 exons 2-4, with the mutation c.192G>C at the end of exon 3 highlighted in red and the homology construct (lower graph) used for gene editing containing a wt exon 3 and a floxed selection cassette encoding a neomycin resistance gene. The 5' and 3' homology arms are indicated by arrows, and the products of the PCRs that were used to screen for positively edited clones are indicated. For primer sequences, see methods. (**B**) Representative agarose gel of PCR of single-cell clones after gene editing. Lanes 3 and 5 show the 2 positively edited clones 42 and 47. Clone 47 was used for further studies. (**C**) Sanger sequencing result of the parental line hom2 and the 2 alleles of the edited clone 47, showing that the mutation was corrected in one allele in clone 47. (**D**) Expression amounts of full-length *PARK7* mRNA (black bars) and Δex3 *PARK7*1 mRNA (gray bars) in the indicated fibroblast lines and the gene-corrected clone 47. Expression amounts were detected by TaqMan qPCR and normalized to overall *PARK7* mRNA expression detected by a TaqMan probe binding the exon 6-7 junction. Values show mean + s.e.m. of technical replicates.

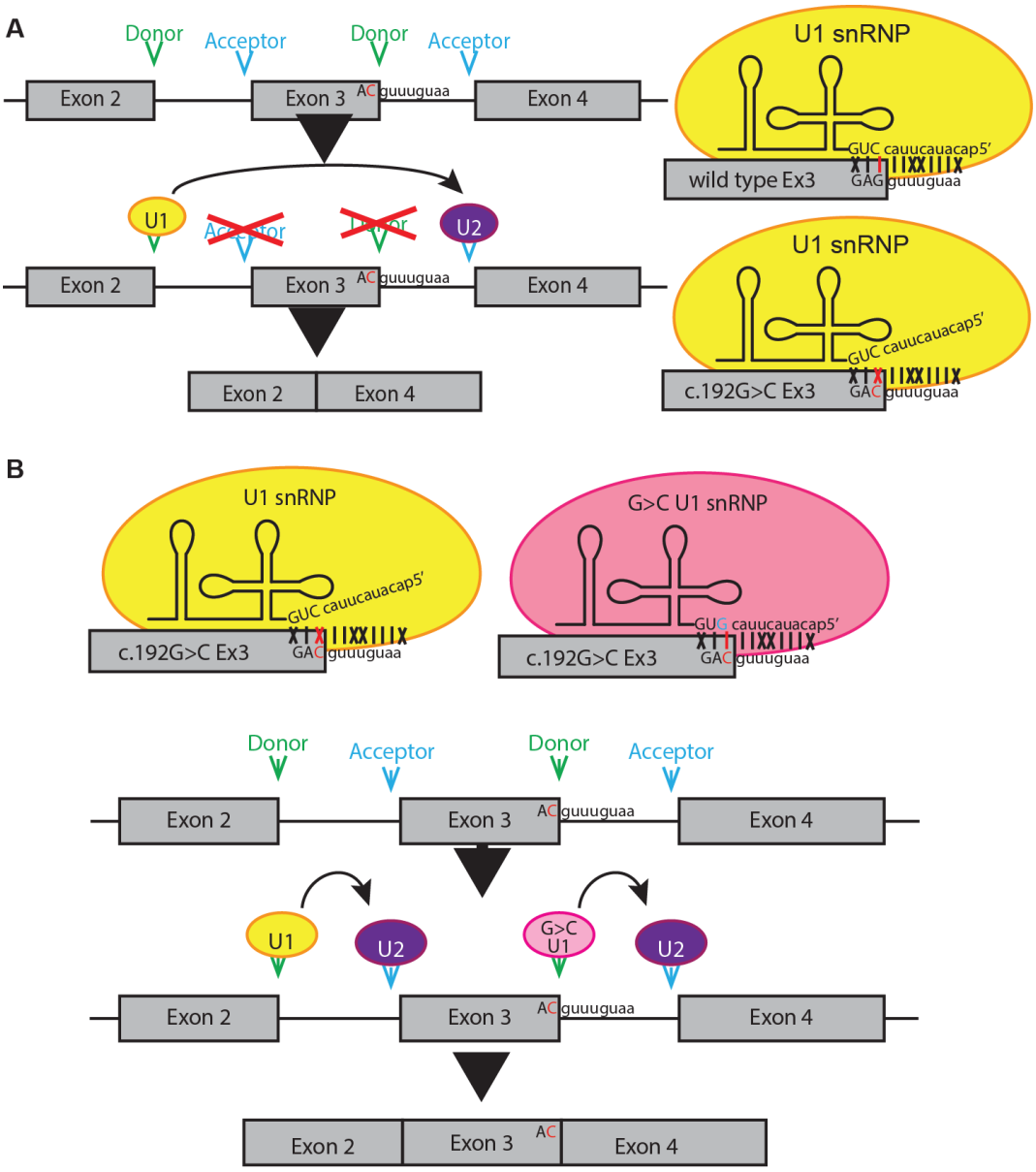


Fig. S4. Schematic of genetic intervention to rescue aberrant splicing of c.192G>C PARK7. (**A**) (left) Schematic image of predicted mis-splicing of pre-mRNA carrying the c.192G>C mutation. The mutation, last base of exon 3 (red), is located in the splice site donor. The mutation changes the donor site, which abolishes recognition of the donor site by U1 snRNP and leads to exon skipping. (right, top) Schematic image of U1 snRNP binding to wt *PARK7* exon 3. (right, bottom) U1 snRNA not binding to c.192G>C mutant *PARK7* exon 3. (**B**) Binding of C>G U1 snRNP to c.192G>C *PARK7* exon 3. (top) A C>G base exchange (indicated in cyan) was introduced in the wt U1 snRNA, restoring binding to c.192G>C *PARK7* exon 3. (bottom) Schematic of splicing of c.192G>C *PARK7* pre-mRNA in the presence of G>C U1 snRNP. Expression of recombinant G>C U1 snRNA leads to the formation of G>C U1 snRNP (pink) that will recognize the mutated splice site donor. Consequently, together with endogenous wt U1 snRNP (yellow) and U2 snRNP (purple), G>C U1 snRNP will include exon 3 in spliced c.192G>C *PARK7* mRNA. This mRNA will presumably be translated to E64D DJ-1 protein.

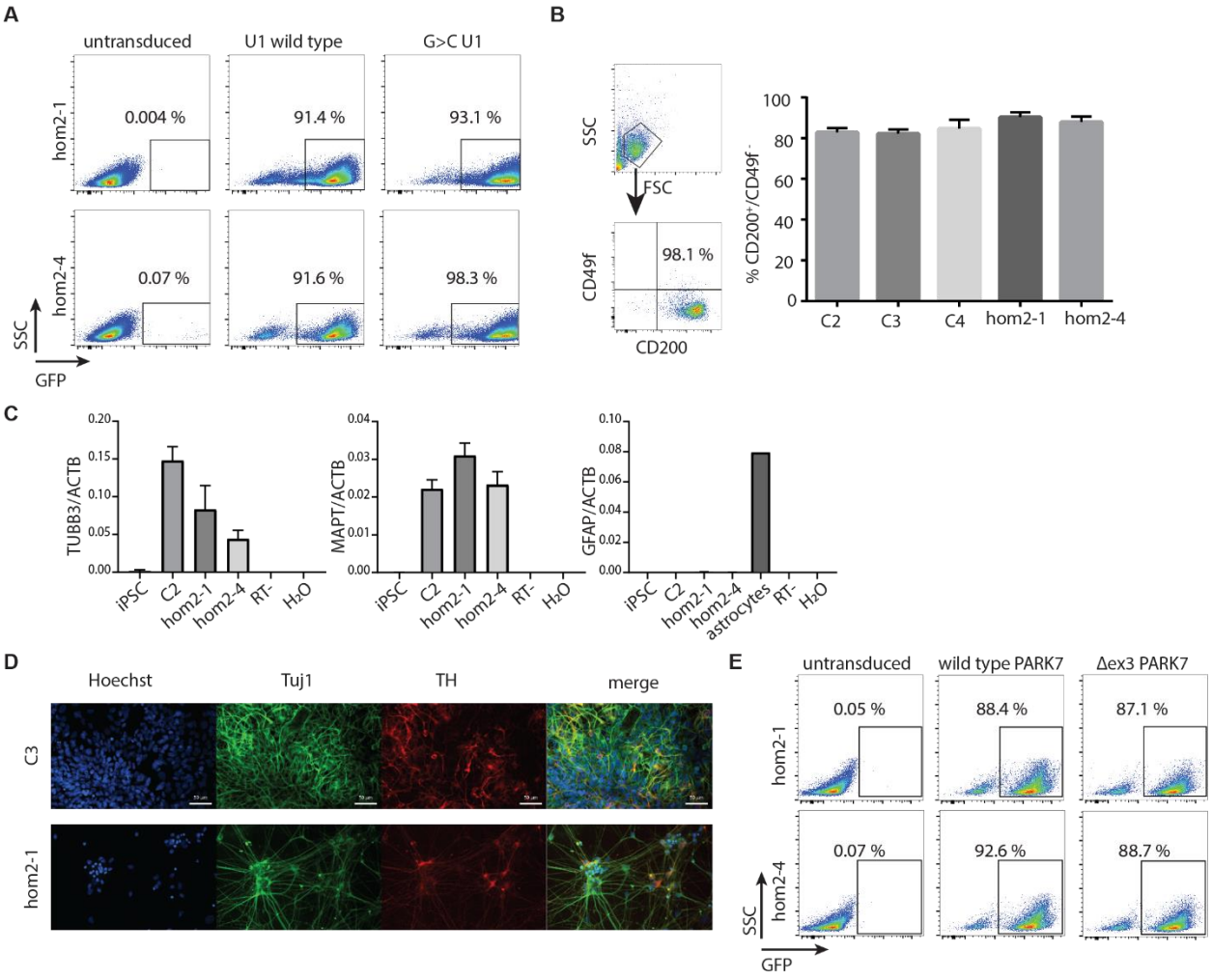


Fig. S5. Transduction and differentiation of smNPC. (**A**) Transduction efficiency with lentiviral vectors encoding GFP and either U1 wt snRNA or G>C U1 snRNA was detected by flow cytometry after enrichment by FACS sorting. (**B**) Neuronal differentiation of smNPC. (left) Representative plots of the flow cytometric analysis of neurons differentiated in vitro (DIV) for 30 d. Cells were stained with antibodies against CD200 and CD49f, and gates were set to detect percentages of $CD200^{+}/CD49f^{-}$ neurons. (right) Average percentage of neurons after in vitro differentiation from indicated smNPC lines. Bars show mean + s.e.m., n = 6. (**C**) qPCR analysis of RNA from in vitro differentiated neurons from experiment in figure 6E. Expression amountsof neuronal markers TUBB3 and MAPT and of astrocyte marker GFAP were analysed. RNA from iPSC and from in vitro differentiated astrocytes were used as negative and positive control, respectively. Bars show mean + s.e.m., n = 5 - 6. (**D**) Representative immunocytochemistry image of mDA neurons differentiated in vitro from smNPC. (**E**) Transduction efficiency with lentiviral vectors encoding GFP and either wt or Δex3 *PARK7* was detected by flow cytometry after enrichment by FACS sorting.

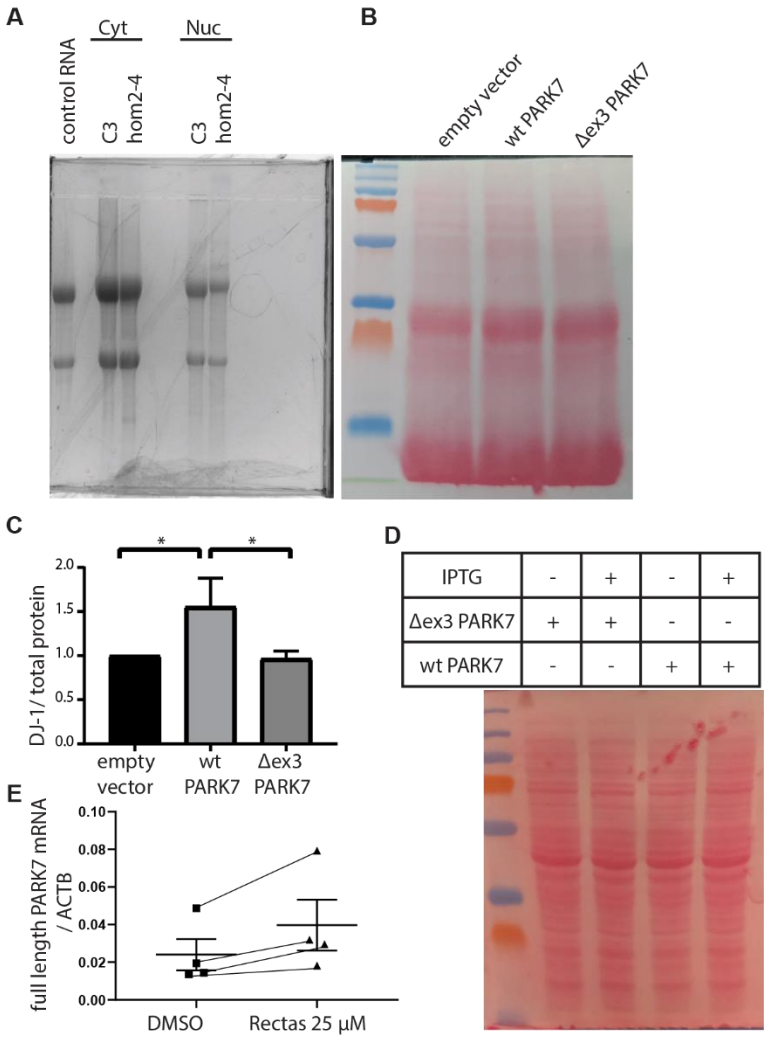


Fig. S6. Loading controls from Fig. 5 and RECTAS single treatment. (**A**) Formaldehyde gel of RNA from the cytosolic and nuclear fractions of the indicated smNPC that served as loading control for the Northern blot shown in Fig. 5E. RNA was visualized by ethidium bromide. (**B**) Ponceau red staining of the nitrocellulose membrane used in Fig. 5F was used to confirm equal loading of protein amounts of all samples. (**C**) Quantification of DJ-1 Western blots of in vitro translation normalized to total protein loaded. Endogeneous DJ-1 signal of the reticulocyte lysate (empty vector) was set to one, n = 4. Bars represent means +s.e.m. Kruskal-Wallis test followed by Dunn's multiple comparison test. *$p<0.05$ (**D**) Ponceau red staining of the nitrocellulose membrane used in Fig. 5G was used to confirm equal loading of protein amounts of all samples. (**E**) Single treatment of neurons differentiated in vitro from the index patient-derived smNPC lines hom2-1 and hom2-4 with 25 µM Rectas. Full-length *PARK7* mRNA was detected by duplex One-step RT-qPCR using TaqMan probes detecting full-length *PARK7* and *ACTB* mRNA and normalized to expression of control C2. Lines connect samples from the same experiment.

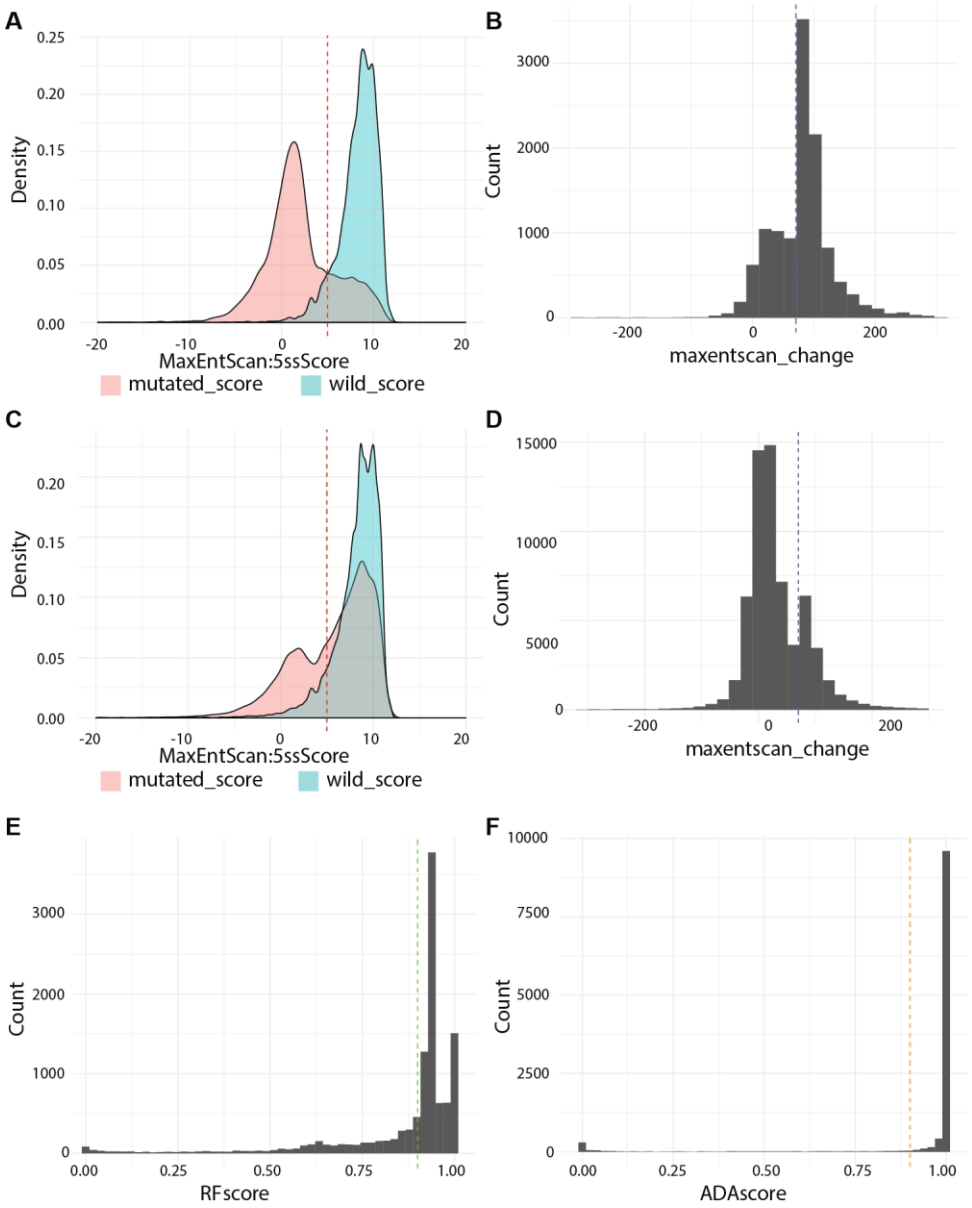


Fig. S7. Determination of cutoffs for wild_score and maxentscan_change. Dashed lines in each plot indicate the cut-offs that were used to define a variant as deleterious (DEL.splicing). (**A**) Distribution of *wild_score* and *mutated_score* of HGMDpatho variants, (**B**) distribution of *maxentscan_change* of HGMDpatho variants, (**C**) distribution of *wild_score*, and *mutated_score* of gnomAD variants, (**D**) distribution of *maxentscan_change* of gnomAD variants, (**E**) distribution of *dbscnv_RF score* of HGMDpatho variants, and (**F**) distribution of *dbscnv_ADA score* of HGMDpatho variants. mutated_score = *maxentscan score* of mutated 9mers and *wild_score* = *maxentscan score* of all wild type 9mers.

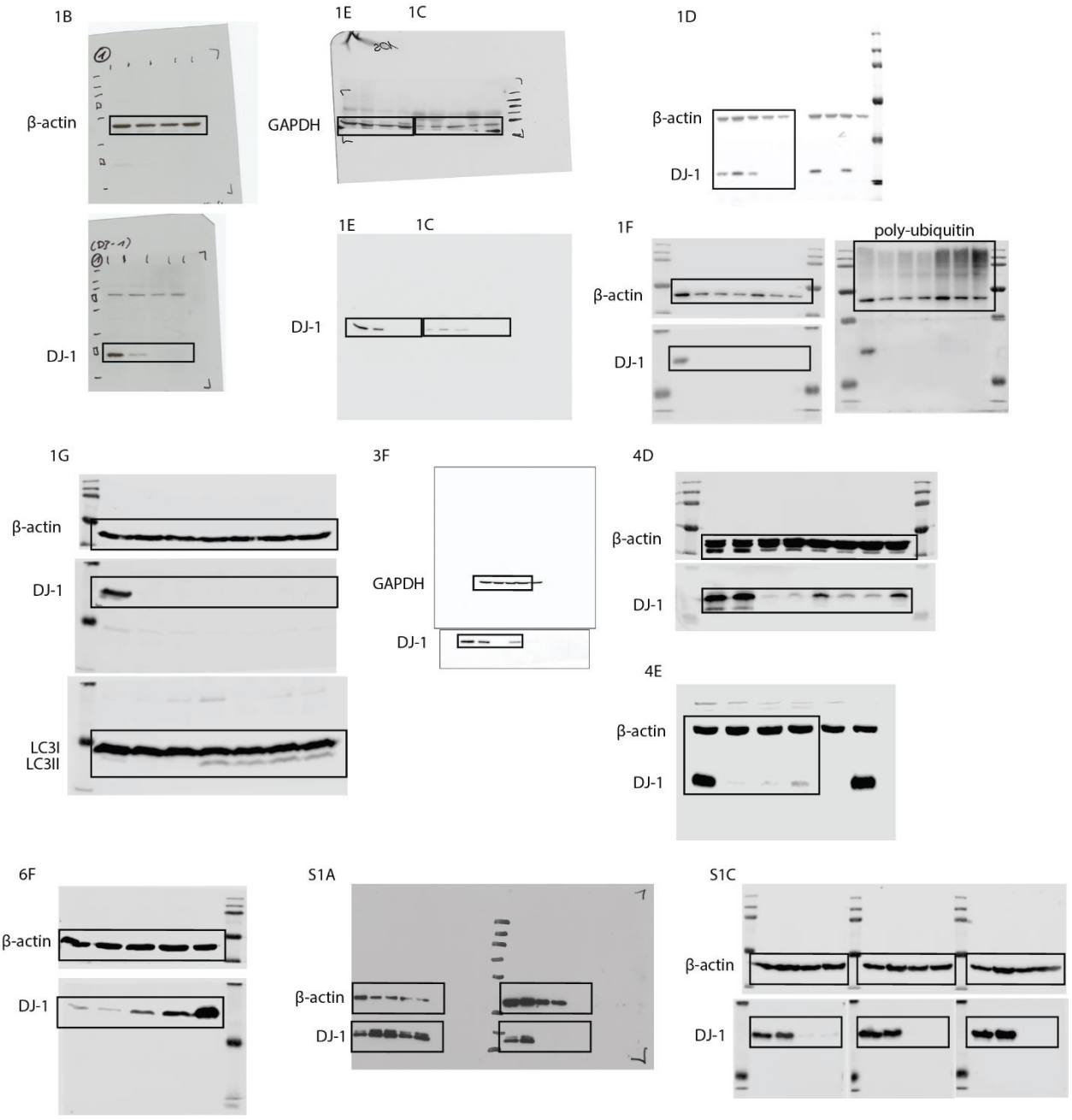


Fig. S8. Western blots. Images of whole Western blots of all representative Western blots from which only segments are shown in the figures. Corresponding figures are indicated for each blot.

Table S1. Features of midbrain organoids extracted from image analysis.

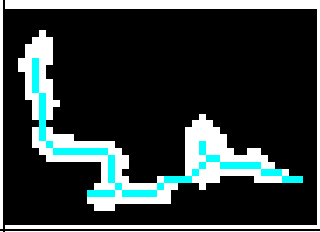
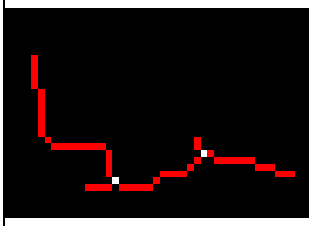
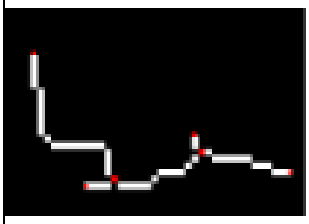
Feature	***Description.***
TH Pixels	Count of TH positive pixels
Skel TH	Count of TH skeleton pixels within an image (blue line) 
TH Links	Total number of links in the TH skeleton (red lines) 
TH Nodes	Total number of branches and end-points in the TH skeleton (red points) 
TH %	Number of nuclei positive for TH
TH by Hoechst	Ratio of TH positive pixels and Hoechst positive pixels
Hoechst Pixels	Count of nuclear mask pixels
Tuj1 Pixels	Count of neuronal mask pixels
TH Fragmentation	Surface to Volume ratio of TH Mask
Tuj1 by Hoechst	Ratio of Tuj1 positive pixels and Hoechst positive pixels

Table S2. List of primers used for determination of gene expression by CYBR green qPCR.

Label	Target	Purpose	Sequence	Company
endo_Klf4 fwd	endogenous *klf4*	stem cell marker expression	5‘- acagtctgttatgcactgtggtttca -3‘	Metabion international AG (Steinkirchen, Germany)
endo_Klf4 rev			5‘- catttgttctgcttaaggcatacttgg -3‘	
endo_cMyc fwd	endogenous *cmyc*	stem cell marker expression	5‘- ccagcagcgactctgagga -3‘	
endo_cMyc rev			5‘- gagcctgcctcttttccacag -3‘	
REX1 fwd	*rex1*	stem cell marker expression	5‘- gcacactaggcaaacccacc -3‘	
REX1 rev			5‘- catttgtttcagctcagcgatg -3‘	
endo_OCT4 fwd	endogenous *oct4*	stem cell marker expression	5‘- ggaaggaattgggaacacaaagg -3‘	
endo_OCT4 rev			5‘- aacttcaccttccctccaacca -3‘	
endo_SOX2 fwd	endogenous *sox2*	stem cell marker expression	5‘- tggcgaaccatctctgtggt -3‘	
endo_SOX2 rev			5‘- ccaacggtgtcaacctgcat -3‘	
NANOG fwd	*nanog*	stem cell marker expression	5‘- cctgtatttgtgggcctg -3‘	
NANOG rev			5‘- gacagtctccgtgtgaggcat -3‘	
TDGF1 fwd	*TDGF1*	stem cell marker expression	5‘- ctgctgcctgaatgggggaacctgc -3‘	
TDGF1 rev			5‘- gccacgaggtgctcatccatcacaagg -3‘	
UPF1 fwd	*UPF1*	stem cell marker expression	5‘- ccgtcgctgaacaccgccctgctg -3‘	
UPF1 rev			5‘- cgcgctgcccagaatgaagcccac -3‘	
DNMT3B fwd	*DNMT3B*	stem cell marker expression	5‘- gctcacagggcccgatactt -3‘	
DNMT3B rev			5‘- gcagtcctgcagctcgagttta -3‘	
OCT4 viral fwd	viral *oct4*	silencing of reprogramming factors	5‘- ggctctcccatgcattcaaac -3‘	
OCT4 viral rev			5‘- catggcctgcccggttatta -3‘	
SOX2 viral fwd	viral *sox2*	silencing of	5‘- gcacactgcccctctcacac -3‘	

SOX2 viral rev		reprogramming factors	5‘- caccagaccaactggtaatggtagc -3‘	
KLF4 viral fwd	viral *klf4*	silencing of reprogramming factors	5‘- cctcgccttacacatgaagagaca -3‘	
KLF4 viral rev			5‘- caccagaccaactggtaatggtagc -3‘	
c-MYC viral fwd	viral *cmyc*	silencing of reprogramming factors	5‘- gctacggaactcttgtgcgtga -3‘	
c-MYC viral rev			5‘- caccagaccaactggtaatggtagc -3‘	
HMBS fwd	*HMBS*	housekeeping level	5‘- atgccctggagaagaatgaagt -3‘	
HMBS rev			5‘- ttgggtgaaagacaacagcatc -3‘	
RTPCR DJ1 fwd	*PARK7*	detection of total *PARK7* mRNA	5‘- ggttctaccaggaggtaatctgg -3‘	
RTPCR DJ1 rev			5’- atttcatgagccaacagagc -3’	
RTPCR Ex3 fwd	*PARK7 with exon 3*	detection only of full length *PARK7* mRNA	5‘- cgagctgggattaaggtcac -3‘	
RTPCR Ex3 rev			5‘- atcttcaaggctggcatcag -3‘	
RT_ACTb_fwd	*ACTB*	housekeeping level	5‘- ctggaacggtgaaggtgaca -3‘	Eurogentec (Liège, Belgium)
RT_ACTb_rev			5‘- aagggacttcctgtaacaatgca -3‘	

Table S3. Hydrolysis probes and primers. List of hydrolysis probes and corresponding primers used to determine the expression of full-length *PARK7* and Δex3 *PARK7* mRNA in duplex qPCR. Custom-made primers and probes are listed in the table in the following order: forward primer, reverse primer, and hydrolysis probe. All other targets were detected with commercial primers/probe kits from ThermoFisher Scientific.

Label	Target	Purpose	Sequence/ Assay ID	Company
EX3-4_FAM	full length *PARK7* mRNA	detects only correctly spliced mRNA	5‘- ggagacggtcatccctgtagat -3‘ 5‘- ctacactgtactgggtcttttcca -3‘ 5‘- acggtgaccttaatccca -3‘	ThermoFisher Scientific
EX2-4_FAM	Δex3 *PARK7* mRNA	detects only mis-spliced mRNA	5‘- ggagacggtcatccctgtagat -3‘ 5‘- tcctggtagaaccaccacatca -3‘ 5‘- tatggtcccccagctcgc -3‘	
ACTB-Vic	*β-actin mRNA*	housekeeping level	Hs01060665_g1	
EX6-7_FAM	*overall PARK7 mRNA*	detects overall levels of *PARK7* mRNA	Hs00994896_g1	
EX2-3_FAM	full length *PARK7* mRNA	detects only correctly spliced mRNA	Hs00994893_g1	

Data file S1. High-confidence sequence variants in the proband as determined by resequencing of the PARK7 gene.

Data file S2. List of brain-expressed genes harboring variants uniquely identified in cases from both cohorts.

Data file S3. Raw data.