Integrated Sequencing and Array Comparative Genomic Hybridization in Familial Parkinson's Disease

Laurie A. Robak, MD, PhD,[1,*], Renqian Du, PhD, DDS[1,*], Bo Yuan, PhD[1], Shen Gu, PhD[1], Isabel Alfradique-Dunham, MD[2], Vismaya Kondapalli, BS[1], Evelyn Hinojosa, BS[1], Amanda Stillwell, BS[1], Emily Young, BA[1], Chaofan Zhang, BS[1], Xiaofei Song, PhD[1], Haowei Du, MS[1], Tomasz Gambin, PhD[1], Shalini N. Jhangiani, PhD[3], Zeynep Coban Akdemir, PhD[1,4], Donna M. Muzny, MS[3], Anusha Tejomurtula, MS[1], Owen A. Ross, PhD[5,6,7], Chad Shaw, PhD[1], Joseph Jankovic, MD[2,8], Weimin Bi, PhD[1,4], Jennifer E. Posey MD, PhD[1], James R. Lupski, MD, PhD[1,3,9,10,+ #], Joshua M. Shulman MD, PhD[1,2,11,12,+,#]

Supplementary Methods

Subjects

We studied 110 PD cases evaluated in the Baylor College of Medicine (BCM) PD Center and Movement Disorders Clinic in Houston, TX with a family history of PD (45% of cases reported an affected first degree relative; 55% reported a second or third degree relative). All diagnoses were made by movement disorder specialists. Subjects in our cohort were unrelated, except for 2 brothers (Subjects 21 and 22) known to have *PRKN*-related PD.[1] However, neither ES nor aCGH were previously performed in this sibling pair. Parental and other family member samples were not available for any subjects. All subjects provided informed consent and the study was approved by the BCM Institutional Review Board. As a positive control for aCGH, we also included a sample from a known subject with an *SNCA* triplication.[2–4] We also interrogated a Baylor Genetics diagnostic laboratory sample including 12,922 clinical referral samples for aCGH from peripheral blood.[5] This analysis of aggregate clinical genomic data was also approved by the BCM Institutional Review Board. Subject numbers throughout the text are consistent with clinical and demographic details provided in table e-1.

Gene Set Definition and Variant Criteria

We focused our analyses on genes and variants established to cause familial PD, including the autosomal dominant loci, *SNCA* (*PARK1*, MIM#168601),[6] *GBA* (MIM#168600),[7,8] *LRRK2* (MIM#607060),[9–11] *GCH1* (MIM#600225),[12] *DNAJC13* (MIM#616361),[13,14] and *VPS35* (MIM#614203),[15,16] as well as the autosomal recessive loci, *PRKN* (*PARK2,* MIM#600116),[17] *PINK1* (MIM#605909),[18] and *PARK7* (*DJ1*, MIM#606324),[19]based on the available literature in April 2015 when this study was initiated.[20,21] Gene names in this study conform to current guidelines from the HUGO Gene Nomenclature Committee (https://www.genenames.org). Deletions of 22q11.2 have been recognized as incompletely-penetrant PD risk factors[22] and were also considered for CNV analyses. From the ES data, all candidate variants were reviewed by clinical geneticists (LR, JP, JL) and movement disorder neurologists (JS, JJ) to establish a consensus on pathogenic alleles, integrating data from multiple available resources (PubMed, ClinVar,[23] Human Genome Mutation Database[24]). All pathogenic alleles included in this study are well-established, non-synonymous coding variants with moderate to high penetrance (OR>2) meeting stringent evidence for replication across studies or within the same study. All other variants discovered in these genes but not previously reported in PD were considered variants of unknown significance (VUS).

Exome sequencing

Genomic DNA was extracted from peripheral blood samples obtained from each participant. ES for the first 50 subjects was obtained from a prior study,[25] and ES for the remaining 60 subjects was performed at the BCM Human Genome Sequencing Center (HGSC).[26] Preparation and sequencing of genomic DNA was performed as previously described with two capture reagents.[27] In brief, the BCM HGSC-developed library (VCRome 2.1 and Core)[28] was used for exome capture[27] and the Illumina HiSeq 2000 system was used for sequencing. Sequencing data were processed through the HGSC-developed Mercury pipeline using the Atlas2 variant calling method with GRCh37/hg19[29,30] and annotated using the Cassandra annotation pipeline[31] based on ANNOVAR.[32] Samples achieved an average of 95% of targeted exome bases covered to a depth of 20X or greater. Sequencing yields averaged 8.64 Gb per sample. Pathogenic SNVs detected via ES were confirmed via Sanger sequencing using Takara HotStarTaq or Qiagen HotStarTaq.

High-density array-based comparative genomic hybridization (aCGH) and breakpoint mapping

Genome-wide aCGH was performed on 99 out of 110 subjects for which sufficient DNA remained after ES. We used exon-targeted Baylor Genetics (https://www.baylorgenetics.com/) V10 2x400K high-density clinical-grade oligonucleotide microarrays, which were previously used for clinical diagnostic purposes. We defined potential CNVs as those regions with three or more consecutive probes showing consistent direction of effect. To confirm these potential CNVs, we secondarily employed a custom 8x60K high-density array with interrogating oligonucleotide spacing as high as ~200 bp per probe through Agilent (CA, USA). To fine-map potential CNVs detected by clinical arrays using the Agilent protocol (https://www.agilent.com), gender-matched controls were applied in hybridization (HapMap individual NA15510 as female control and NA10851 as male control). Arrays were scanned by Agilent DNA microarray scanner at 3μm resolution, and then processed by Agilent Feature Extraction software (version 10). The extracted files from the scanned images were analyzed using Agilent Genomic Workbench (version 7.0.4.0). A $\log_2$ ratio was calculated based on the signal intensity of input sample against control sample, i.e. $\log_2 (1/2) = -1$ for deletions, $\log_2 (2/2) = 0$ for diploids, $\log_2 (3/2) = 0.58$ for duplications, and $\log_2 (4/2) = 1$ for triplications. Human genome build GRCh37/hg19 was used for array designs and sequence alignment. Primers for breakpoint mapping were designed based on aCGH results and the oligonucleotides marking the transition in copy number state (see primer sequences below). Regular and long-range PCR of breakpoint junctions was then performed using Takara LA Taq and Qiagen HotStarTaq using manufacturer's protocols. Purified PCR products were sequenced using Sanger dideoxynucleotide sequencing. For the Baylor Genetics clinical cohort lookup, data was available from 12,922 subjects profiled for CNVs

using Baylor arrays (v9 or v10).[5] CNVs were filtered according to the following criteria: 3+ consecutive probes with consistent direction of effect, $\log_2$ ratios for loss <=0.5, gain>0.35. Filtered CNVs were manually assessed to exclude artifacts.[33] For statistical comparison of CNV frequencies between the familial PD cohort and the Baylor Genetics control sample, we performed a 2-tailed Z-test to test the hypothesis that the odds ratio (OR) is not equal to 1 (log(OR) not equal to 0).

Digital Droplet PCR (ddPCR)

The BioRad QX200™ AutoDG™ Droplet Digital™ PCR System was used to perform ddPCR with manufacturer protocols. In brief, *Hind*III-HF restriction enzyme (NEB, MA, USA) was used to digest genomic DNA prior to droplet generation. A 20µL reaction volume was employed for droplet generation, containing 10uL of 2x Q200 ddPCR EvaGreen Supermix, 15ng digested genomic DNA, and 0.2µM primers (See below). Next, PCR was conducted in cycling conditions using Bio-Rad's C1000 Touch Thermal Cycler as follows: 5 minutes (mins) at 95ºC for enzyme activation, 40 cycles of 30 seconds (sec) at 95ºC and 1 min at 60ºC for denaturation and annealing/extension, 5 min at 4ºC and 5 min at 90ºC for signal stabilization. The droplet reading was then performed to obtain concentrations of positive droplets (number of positive droplets per uL of reaction) for each PCR reaction using Bio-Rad QuantaSoft TM Software.

Primer Sequences: The following primer sequences were used for breakpoint sequencing and ddPCR. *Adopted from Gu *et al*[34]; **Re-designed primer for exon 6 of *GBA*.

Breakpoint Sequencing

Primer	Sequence (5' to 3')
SNCA-TRP-R	TCCTGTTCGGTTTTAGTCTGATGAAGCA
SNCA-DUP-R	GCCTCCACTGCACTTGTTATGAGCCA
SNCA-TRP-F	GCCAGCATATGACAAAACCCT
SNCA-DUP-F	ATGGTGAGCTCTATGAGGGC
3-*SNCA*-F	TGCTTATTTCCTGTCCCGGT
3-*SNCA*-R	TCAACCCAATCAACTGCAGC
1-*GBA*-F	GTAGGAATCCTGGAGTTGGGTGACG
1-*GBA*-R	CACTATGCAGAGGGGAAGTGGAAGG
20-*PRKN*-F	TGCCCCATGCAGTTAATGTTTTTGA
20- *PRKN* -R	CGGCTCACAAGCTTACAGAGAGTCG
11- *PRKN* -1F	CAGCTCAGTATCTTGCATTTGCTGCTA
11- *PRKN* -1R	TTGCTGTGTGCCGTCAGAGTAATAAAA
11- *PRKN* -2F	CCATCTATTTCGTCTTTCCTGCCTCAT
11- *PRKN* -2R	TTCAACATGAGTTTCCTTCCCCTTT
21- *PRKN* -F	AGTGACTGGCTTCTATCTGGGTTCACA
21- *PRKN* -R	ACTGTGGACTTTTCTCTGAGCATACCC
GBA-E6-F	ACAAGCAGACCTACCCTACA
GBA-E6-2-F**	ACAAGCAGACCTACCCTACC
GBA-E6-R	CTATGCAGACACCCCTGATG
GBA-E7-F	TCAGCATGGCTAAATGGGAG
GBA-E7-R	TTGGCTCAAGACCAATGGAG
GBA-E8-F	TTGTGGTGAGTACTGTTGGC
GBA-E8-R	TTACAGTTCTGGGCAGTGAC
GBA-E9-F	AAAGAGCATGGTGTTGGGGA
GBA-E9-R	CAGACCCAGAAGCAGCTAAA
GBA-E10-F	ACTGTCGACAAAGTTACGCA
GBA-E10-R	TCTCCCACATGTGACCCTTA
GBA-E11-F	GTCCAGGTCGTTCTTCTGAC
GBA-E11-R	TGGGTGGGTGACTTCTTAGA
GBA-E12-F	GGGAAAGTGAGTCACCCAAA
GBA-E12-R	GGCTTCCTGGAGACAATCTC
RPPH1-F*	AATGGGCGGAGGAGAGTAGTCTGAAT
RPPH1-R*	CGAAGTGAGTTCAATGGCTGAGGTG
TERT-F*	GCACACCTTTGGTCACTCCAAATTC
TERT-R*	CCACATAGGAATAGTCCATCCCCAGA

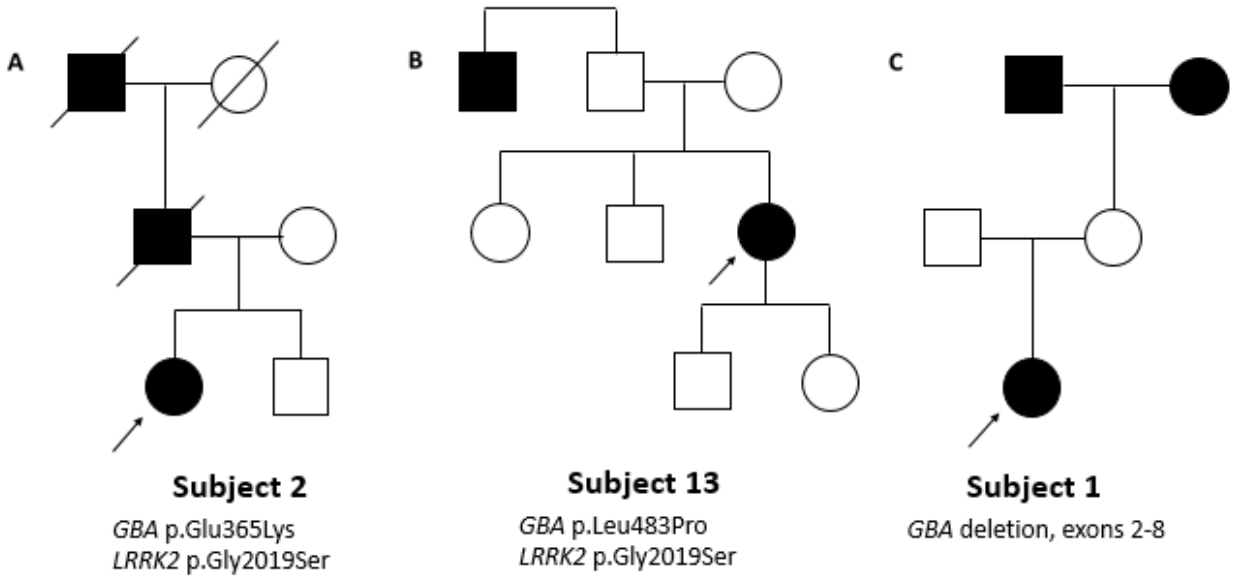
ddPCR

PRKN -E1-F	GAGGCGTGAGGAGAAACTAC
PRKN -E1-R	GGCTCTCCTGGGTTAAATCC
PRKN -E2-F	GAGGTCGATTCTGACACCAG
PRKN -E2-R	GTCCAGTCATTCCTCAGCTC
PRKN -E3-F	TCAGCAGAGCATTGTTCACA
PRKN -E3-4	TCAGTGTGCAGAATGACAGC
PRKN -E4-F	CGGGAAAACTCAGGGTACAG
PRKN -E4-R	CATGCTGACACTGCATTTCC
PRKN -E5-F	CCATCTTGCTGGGATGATGT
PRKN -E5-R	TGACCAGGTACTTACTGCAC
PRKN -E6-F	AGTCGGAACATCACTTGCAT
PRKN -E6-R	TCCCCAGGAAAGAGAGTTCA
PRKN -E7-F	AGCCCCGTCCTGGTTTTC
PRKN -E7-R	GAGTAGCCAAGTTGAGGGTC
PRKN -E8-F	AGGATTCTGGGAGAAGAGCA
PRKN -E8-R	AGCATGGTTTTCTTCCCCAT
PRKN -E9-F	CAGTATGGTGCAGAGGAGTG

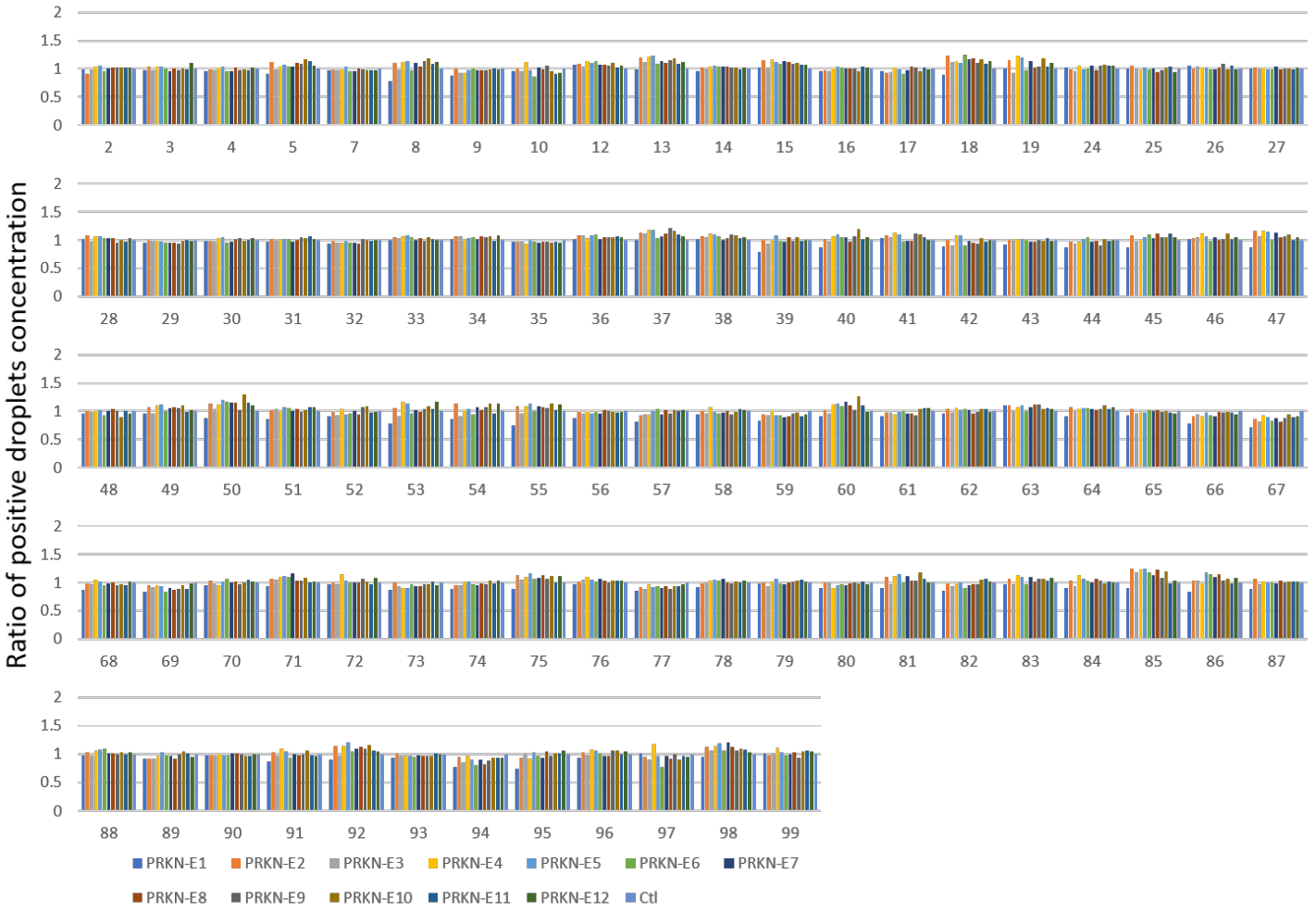
ddPCR (continued)

Primer	Sequence (5' to 3')
PRKN -E9-R	GTGACTTTCCTCTGGTCAGG
PRKN -E10-F	CCGGGAATGTAAAGAAGCGT
PRKN -E10-R	AGTTGTTCCTGAGGCTTCAA
PRKN -E11-F	AGGCCTACAGAGTCGATGAA
PRKN -E11-R	TTTCCACTGGTACATGGCAG
PRKN -E12-F	CCACTGGTTCGACGTGTAG
PRKN -E12-R	AGGTAGACACTGGGTATGCT
GBA-E1-F	GAGGCAAAACGAAATCCCAC
GBA-E1-R	GCATGAGTGACCGTCTCTTT
GBA-E2-F	GAAAACTCCATCCCCTCAGG
GBA-E2-R	CGGAATTACTTGCAGGGCTA
GBA-E3-F	TTGACTCACTCACCTGATGC
GBA-E3-R	AGTAGTTGAGGGGTGGAGAG
GBA-E4-F	CATACTCAGCTCCATCCGTC
GBA-E4-R	TCTGCAATGCCACATACTGT
GBA-E5-F	GTAGCAAATTTTGGGCAGGG
GBA-E5-R	ATTCCCTGTGGATGTCCTCA

<u>Supplementary Figure e-1</u>: **Selected study pedigrees**. (A) *Subject 2*, *GBA* p.Glu365Lys and *LRRK2* p.Gly2019Ser. (B) *Subject 13*, *GBA* p.Leu483Pro and *LRRK2* p.Gly2019Ser. (C) *Subject 1*, *GBA* deletion exons 2-8.



<u>Supplementary Figure e-2</u>: ***PRKN* ddPCR screening**. ddPCR was used to perform an exon-by-exon screen for copy number variation involving one or more exons of *PRKN*. Primer pairs for the 12 exons of *PRKN* and two housekeeping genes *RPPH1* and *TERT* (i.e. neutral copy number in human genome) were used to obtain positive droplet concentrations from PCR in 92 individuals from the PD cohort for whom sufficient DNA was available. The y-axis shows the exon-by-exon results indicated by 13 columns with different colors, showing comparable ddPCR results to that of the average value of *RPPH1* and *TERT* used as controls. A y-axis value of 0.5 is consistent with a deletion, 1 indicates copy neutral (no deletion, no duplication), and 1.5 a duplication.



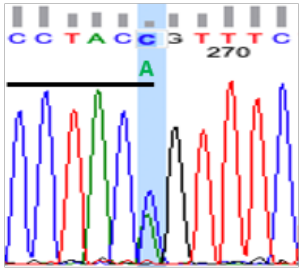
<u>Supplementary Figure e-3</u>: ***GBA* ddPCR screening**. ddPCR was used to perform an exon-by-exon screen for copy number variation involving one or more exons of *GBA* in 85 individuals from the PD cohort for whom sufficient DNA was available. Primer pairs for 6 of the 12 exons of *GBA* produce amplicons concurrently from *GBA* and its pseudogene, *GBAP1*. This results in a doubling of the apparent number of exon copies identified by ddPCR for these exons: four, instead of two copies (ratio = 2) indicate copy number neutral for these exons.

<u>Supplementary Figure e-4:</u> Amplification of exon 6 of *GBA* in subject 48. A) Re-designed exon 6 primers incorporating the SNV chr1:155208250A>C (rs1317644130) into the primer sequence and Sanger confirmation. B) Subsequent ddPCR reactions using 1) initial primers on subject 48 (blue); 2) a combination of the initial and re-designed primers on subject 48 (orange); 3) initial primers on HapMap individual NA10851 shown as a control (gray).

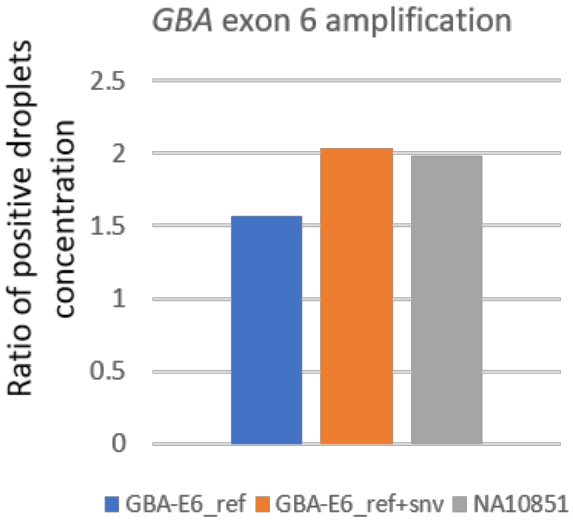
A

GBA-E6-F: ACAAGCAGACCTACCCTACA (initial primers)

GBA-E6-2F: ACAAGCAGACCTACCCTACC (re-designed primers)



B



Supplementary References:

1. Deng H, Le W-D, Hunter CB, et al. Heterogeneous phenotype in a family with compound heterozygous parkin gene mutations. Arch Neurol. 2006;63:273–277.
2. Spellman GG. Report of familial cases of parkinsonism. Evidence of a dominant trait in a patient's family. JAMA. 1962;179:372–374.
3. Muenter MD, Forno LS, Hornykiewicz O, et al. Hereditary form of parkinsonism--dementia. Ann Neurol. 1998;43:768–781.
4. Singleton AB, Farrer M, Johnson J, et al. -Synuclein Locus Triplication Causes Parkinson's Disease. Science (80-). 2003;302:841–841.
5. Gambin T, Yuan B, Bi W, et al. Identification of novel candidate disease genes from de novo exonic copy number variants. Genome Med. 2017;9:83.
6. Polymeropoulos MH, Lavedan C, Leroy E, et al. Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. Science. 1997;276:2045–2047.
7. Sidransky E, Nalls MA, Aasly JO, et al. Multicenter Analysis of Glucocerebrosidase Mutations in Parkinson's Disease. N Engl J Med. 2009;361:1651–1661.
8. Nalls MA, Duran R, Lopez G, et al. A Multicenter Study of Glucocerebrosidase Mutations in Dementia With Lewy Bodies. JAMA Neurol. 2013;70:727.
9. Funayama M, Hasegawa K, Kowa H, Saito M, Tsuji S, Obata F. A new locus for Parkinson's disease (PARK8) maps to chromosome 12p11.2-q13.1. Ann Neurol. 2002;51:296–301.
10. Paisán-Ruíz C, Jain S, Evans EW, et al. Cloning of the Gene Containing Mutations that Cause PARK8-Linked Parkinson's Disease. Neuron. 2004;44:595–600.
11. Zimprich A, Biskup S, Leitner P, et al. Mutations in LRRK2 Cause Autosomal-Dominant Parkinsonism with Pleomorphic Pathology. Neuron. 2004;44:601–607.
12. Mencacci NE, Isaias IU, Reich MM, et al. Parkinson's disease in GTP cyclohydrolase 1 mutation carriers. Brain. 2014;137:2480–2492.
13. Gustavsson EK, Trinh J, Guella I, et al. *DNAJC13* genetic variants in parkinsonism. Mov Disord. 2015;30:273–278.
14. Vilariño-Güell C, Rajput A, Milnerwood AJ, et al. DNAJC13 mutations in Parkinson disease. Hum Mol Genet. 2014;23:1794–1801.
15. Zimprich A, Benet-Pagès A, Struhal W, et al. A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset Parkinson disease. Am J Hum Genet. 2011;89:168–175.
16. Vilariño-Güell C, Wider C, Ross OA, et al. VPS35 Mutations in Parkinson Disease. Am J Hum Genet. 2011;89:162–167.
17. Kitada T, Asakawa S, Hattori N, et al. Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. Nature. 1998;392:605–608.
18. Valente EM, Abou-Sleiman PM, Caputo V, et al. Hereditary Early-Onset Parkinson's Disease Caused by Mutations in PINK1. Science (80-). 2004;304:1158–1160.
19. Bonifati V, Rizzu P, van Baren MJ, et al. Mutations in the DJ-1 Gene Associated with Autosomal Recessive Early-Onset Parkinsonism. Science (80-). 2003;299:256–259.
20. Brás J, Guerreiro R, Hardy J. SnapShot: Genetics of Parkinson's disease. Cell. Elsevier; 2015;160:570-570.e1.

21. Hernandez DG, Reed X, Singleton AB. Genetics in Parkinson disease: Mendelian versus non-Mendelian inheritance. J Neurochem. 2016;139 Suppl 1:59–74.
22. Butcher NJ, Kiehl T-R, Hazrati L-N, et al. Association Between Early-Onset Parkinson Disease and 22q11.2 Deletion Syndrome. JAMA Neurol. 2013;70:1359.
23. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42:D980–D985.
24. Stenson PD, Mort M, Ball E V, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet. 2017;136:665–677.
25. Farlow JL, Robak LA, Hetrick K, et al. Whole-Exome Sequencing in Familial Parkinson Disease. JAMA Neurol. 2016;73:68.
26. Posey JE, O'Donnell-Luria AH, Chong JX, et al. Insights into genetics, human biology and disease gleaned from family based genomic studies. Genet Med. 2019;21:798–812.
27. Du R, Dinckan N, Song X, et al. Identification of likely pathogenic and known variants in TSPEAR, LAMB3, BCOR, and WNT10A in four Turkish families with tooth agenesis. Hum Genet. 2018;137:689–703.
28. Bainbridge MN, Wang M, Wu Y, et al. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. Genome Biol. 2011;12:R68.
29. Shen Y, Wan Z, Coarfa C, et al. A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res. 2010;20:273–280.
30. Reid JG, Carroll A, Veeraraghavan N, et al. Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. BMC Bioinformatics. 2014;15:30.
31. Bainbridge MN, Wiszniewski W, Murdock DR, et al. Whole-Genome Sequencing for Optimized Patient Management. Sci Transl Med. 2011;3:87re3-87re3.
32. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164–e164.
33. Innes J, Reali L, Clayton-Smith J, et al. CNVs affecting cancer predisposing genes (CPGs) detected as incidental findings in routine germline diagnostic chromosomal microarray (CMA) testing. J Med Genet. 2018;55:89–96.
34. Gu S, Posey JE, Yuan B, et al. Mechanisms for the Generation of Two Quadruplications Associated with Split-Hand Malformation. Hum Mutat. 2016;37:160–164.