Identification of genetic risk factors for Parkinson's disease

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften

der Mathematisch-Naturwissenschaftlichen Fakultät
und
der Medizinischen Fakultät
der Eberhard-Karls-Universität Tübingen

vorgelegt

von

Anastasia Illarionova
aus Podolsk, Russland

2023

Tag der mündlichen Prüfung: 15.03.2023

Dekan der Math.-Nat. Fakultät: Prof. Dr. Thilo Stehle

Dekan der Medizinischen Fakultät: Prof. Dr. Bernd Pichler

1. Berichterstatter: Prof. Dr. Peter Heutink

2. Berichterstatter: Prof. Dr. Thomas Gasser

Prüfungskommission: Prof. Dr. Peter Heutink

Prof. Dr. Thomas Gasser

Prof. Dr. Stefan Bonn

Prof. Dr. Kay Nieselt

Erklärung / Declaration:

Ich erkläre, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:

“Identification of genetic risk factors for Parkinson’s disease”

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

I hereby declare that I have produced the work entitled “Identification of genetic risk factors for Parkinson’s disease”, submitted for the award of a doctorate, on my own (without external help), have used only the sources and aids indicated and have marked passages included from other works, whether verbatim or in content, as such. I swear upon oath that these statements are true and that I have not concealed anything. I am aware that making a false declaration under oath is punishable by a term of imprisonment of up to three years or by a fine.

Tübingen, den 15.03.2023 .....

Table of Contents

Abstract

Parkinson's disease (PD) is a common progressive neurodegenerative disorder with a complex and heterogeneous genetic landscape. Approximately 90% of all PD cases are driven by the cumulative effect of several common low-risk genetic variants. Over the last years, genetic studies of familial and sporadic PD cases identified a range of high and low-risk variants, representing approximately 40% of estimated heritability. However, the role of structural variants (SV) in the PD missing heritability remains understudied. Therefore, we investigated SVs in the human cohort enriched for the PD phenotype to expand our knowledge about the putative PD genetic risk factors. We leveraged the matching omics datasets obtained from 95 iPSC lines differentiated into the dopaminergic neuronal-like state to run the SV calling and to directly assess their impact on the gene and transcript expression. We demonstrated a conceptual approach for the genome-wide SV annotation and pathogenicity assessment, addressing the challenges of functional SV effect prediction based on the known properties of genome regions and available multi-omics data. Using this approach, we prioritized a group of non-coding SVs absent in the healthy controls with a strong association with the differential expression of genes whose dysregulation can trigger the development of PD or PD-related phenotype. Discovered variation impacts molecular mechanisms involved in the regulation of signaling processes, oxidative stress response, and neuronal DNA reparation. Additional analysis on the larger PD patient and control cohort has to be conducted for variant-expression association validation and exploration of the allele effect size and penetrance of the prioritized hits. The dataset is publicly available to facilitate the further discovery of SV PD risk association as well as to study sequence signatures and neurological disease-specific SV hot spots.

1 Introduction

1.1 Parkinson's disease - missing heritability and molecular pathways

Parkinson's disease (PD) is a complex progressive neurodegenerative disorder that manifests both motor and non-motor impairments. In Europe, prevalence and incidence rates for PD are estimated at approximately 108–257/100 000 and 11–19/100 000 per year, respectively [1]. In Germany, PD prevalence is estimated to be 217/100,000 [2]. Globally, PD is expected to reach a prevalence of 12.4 million cases by 2040 [3].

PD is characterized by a selective loss of dopaminergic neurons in the substantia nigra pars compacta, accumulation of α-synuclein aggregates, and, in some cases, Lewy Bodies formation. The clinical hallmark of PD is motor symptoms, including resting tremor, rigidity, bradykinesia, and gait alterations. Several clinical challenges accompany the PD complexity, including difficulties to make a definitive diagnosis during the early disease stages, personalized treatment to ameliorate motor and non-motor symptoms, and lack of technology to slow down the neurodegenerative process [4]. PD diagnosis and diagnostic differentiation from atypical parkinsonian disorders are challenging in routine clinical practice. The diagnostic accuracy was assessed to be only 80.3%, and 10% of cases with alternative pathologies were diagnosed with PD [5,6].

Both environmental and genetic factors were shown to contribute to the disease via a system of complex interactions. For example, potential associations were found between PD and coffee intake, smoking, and exposure to pesticides [4,7]. The genetics of PD is represented by a complex interplay of highly pathogenic rare causal variants and more frequent variants with a small risk effect size. Even though the variant penetrance in PD is very likely to be affected by different genetic and non-genetic factors, the term ‘monogenic’ is being actively used for PD as a convenient simplification to describe familial cases with highly penetrant mutations [8]. It was shown that approximately 10% of all PD cases exhibit a clear Mendelian inheritance pattern through a dominant or recessive mode with a high risk of PD recurrence within a family [9].

Up to 20 genes have been discovered to cause monogenic forms of PD, and the gene list is constantly updated [8,10]. The first proven genetic factor was discovered in 1997 during a family study which occurred to be a missense variant in SNCA (A53T) [11]. SNCA pathogenic mutations can be classified into three main classes: point missense mutations, repeat expansions in the promoter region, and loci multiplications [10]. Missense mutations in SNCA are highly penetrant and associated with early onset autosomal dominant (AD) PD and a good response to L-DOPA treatment [12]. The main pathogenic effect of the SNCA mutations occurs through a change in the affinity of SNCA to lysosomal transmembrane receptors, thus inhibiting protein autophagy-dependent clearance and triggering the formation of protein oligomers [13–15]. In addition, a growing amount of evidence suggests that mutated SNCA prevents the dopamine vesicle release leading to the neurotransmitter cytoplasmic accumulation and subsequent metabolic dysfunction in dopaminergic neurons due to oxidative stress [16–18].

The most frequent genetic cause of familial AD PD is pathogenic variation in LRRK2 [19,20]. LRRK2 mutations can be both low and high penetrant and typically associated with late-onset AD PD features, presumably including motor symptoms [21]. Known LRRK2 mutations lead to hyperactivation of the enzyme catalytic domain triggering an increased level of autophosphorylation and LRRK2 target protein phosphorylation and affecting microtubule elaboration [22,23]. The complexity of LRRK2 and its participation in many crosstalk molecular pathways make this kinase one of the main targets for developing a PD cure [24].

Another gene implicated in the AD form of PD is VPS35 [25]. The only known missense variant in PD (D620N) is present in different populations with an overall prevalence of 0.115% [26,27]. The mutation impairs the endosomal trafficking pathway causing disturbance of endolysosome maturation and membrane receptor recycling and affecting autophagy and mitophagy processes [28,29].

Mutations in GBA, a gene associated with the lysosomal storage disorder Gaucher disease, are among the highest genetic risk factors for PD. Approximately 3-20% of PD patients in different populations harbor GBA pathogenic mutations [30–32]. It was recently discovered that certain variants increase the risk of PD development in GBA carriers at SNCA and CTSB loci [33]. GBA mutations are acting through both

loss- and gain-of-function mechanisms affecting autophagic-lysosomal pathways, triggering aggregation of SNCA protein and inflammatory response [34].

Pathogenic variants in other PD genes, such as DJ-1, PRKN (parkin), PINK1, and FBXO7 were associated with the early-onset autosomal recessive (AR) forms of PD affecting presumably mitochondrial and mitophagy functions [35–38]. Mutations in PRKN are the most common genetic factors in AR early-onset PD forms, followed by pathogenic variation in PINK1 [39]. The main PD causal small variation types described for PRKN and PINK1 include missense and frameshift mutations [40]. Mutations in these genes disturb the PINK1/parkin mitophagy signaling pathway affecting mitochondrial homeostasis [41,42].

Other known genes with PD causal variants impairing lysosomal and mitochondrial functions and synaptic transmission process are ATP13A2 [43], CHCHD2 [44], DNAJC6 [45], PLA2G6 [46], VPS13C [47], and SYNJ1 [48].

However, most PD cases cannot be explained by a highly penetrant pathogenic variation. PD is believed to follow the “common disease common variant” (CDCV) hypothesis that suggests that the genetic component of PD is the cumulative result of several common low-risk variants [49]. Genome-wide association study (GWAS) is a powerful tool to explore the complex genetics of PD, allowing us to investigate most of the common human genetic variation hypothesis-freely. Recent meta-GWA studies identified 92 independent genome-wide significant association signals in European and Asian populations, representing 20-36% of estimated heritability [50,51]. Several GWAS signals are localized close to known PD genes such as SNCA, LRRK2, GBA, and VPS13C, implying that more frequent variants in these genes can increase the risk for PD. Notably, a recent GWAS discovered several loci that influence PD onset [52].

It is essential not only to identify the genetic risk loci but also to investigate the biological effect of the observed variation to determine the actual causal mutation and perturbed molecular pathways [53]. Expression quantitative loci analysis (eQTL) allowed researchers to associate the discovered GWAS SNPs with gene expression and gene tissue-specific expression, indicating possible molecular interactions between PG GWAS loci and pointing out the most relevant tissues and cell types[50]. Collected omics data and results of functional studies are

accumulated on the PD GWAS Locus Browser platform developed by the International Parkinson's Disease Genomics Consortium (IPDGC) to facilitate further novel gene and variant prioritization: Locus Browser https://github.com/ipdgc/PD-Wiki-Loci [53]. However, despite the valuable insights into PD genetics provided by recent studies, the role of structural variants is understudied due to the technical and biological challenges. Nevertheless, SVs are believed to contribute to the missing heritability of PD substantially [54].

1.2 Structural variants as risk factors for neurological disorders

Structural variants (SVs) are arbitrarily defined as chromosomal genomic rearrangements greater than 50 bps [55]. SVs are classified based on their nature into classes that include DNA unbalanced gains - duplicates (DUPs) and insertions (INSs), or losses such as deletions (DELs), and balanced rearrangements that occur without dosage alterations such as inversions (INVs) and translocations (TRAs). Unbalanced genomic rearrangements are usually referred to as copy number variants (CNVs) [55].

The source of SVs roots in mutational processes occurring during DNA recombination, replication, and repair mechanisms. A common mechanism of SV formation is non-allelic homologous recombination (NAHR) [56,57]. The recombination occurs between non-homologous loci (ectopic recombination) due to the high similarity of the sequences, which are most often located in the repetitive regions, including transposable elements or segmental duplications. The ectopic recombination usually happens on the same chromosome or between two homologous chromosomes; more rarely, it occurs between two non-homologous chromosomes giving rise to DELs and DUPs with a size of several kbps to Mbps [58–60]. It was discovered that most NAHR events happen due to recombination errors in SINEs [61]. Shorter SVs are often generated during the DNA replication and transposable elements activity dominated by Alu-Alu-mediated events in primates [62–64]. The investigation of genome-wide distribution patterns of large

repeats revealed that they quite often overlap SVs associated with different pathogenic phenotypes [65,66].

SVs play a profound role in human genome evolution and genetic adaptation, shaping population diversity [67–70]. The adaptive and detrimental effects of SVs can be explained through the alteration of gene expression. The direct impact can manifest through the gene dosage change or coding region alteration [71,72]. SVs can affect regulatory elements and locally change the 3D DNA structure impacting the organization of topologically associating domains (TAD) in the genome and leading to the dysfunction or rewiring of gene-enhancer interactions [73–75]. Several SVs, presumably CNVs, were already associated with neurological and neurodevelopmental disorders. It was estimated that approximately 15% of neurodevelopmental disorders are caused by large CNVs [76]. For example, DUP or recombination-derived DEL in PMP22 gene leads to Charcot–Marie–Tooth disease type 1A or hereditary neuropathy with liability to pressure palsies respectively [77–79], and SNCA locus multiplication causes the familial type of PD and parkinsonism [80,81]. Structural variants in PRKN account for 43.2% of AR familial PD cases associated with PRKN mutations. This group includes exonic deletions, duplications, and triplications [40,82]. Exonic duplications and deletions in PINK1 and exonic deletions in DJ1 were discovered to be causative for PD and PD-related phenotypes[40]. Several dosage-altering events were associated with autism spectrum disorder, schizophrenia, intellectual disability, epilepsy, and other neurodevelopmental diseases [83–86]. Variation in transcript structure also plays a significant role in human disease manifestation [87,88]. SVs affecting splicing regulation and leading to transcript structure disruption were associated with several monogenic diseases, including familial PD [89–91].

Apart from the simple canonical SVs, more complex genetic rearrangements present in the human genome possess clinical relevance [55]. Several complex SVs were associated with rare genetic disorders [92]. For example, a complex event that involves duplication, triplication, and inversion (DUP-triplication-INV-DUP) affects MECP2 and PLP1 loci leading to the development of MECP2 duplication syndrome, Lubs syndrome, or Pelizaeus-Merzbacher disease [93,94]. Investigation of signaling pathways enriched for the genes affected by validated pathogenic CNVs

causal for brain diseases revealed proteasomal and vesicular functioning pathway clusters to be the most significant [95].

Repeat expansions were proven to be causal pathogenic factors in a number of neurological diseases [96]. For instance, intronic extensions of ATTCC in ATXN10 cause progressive spinocerebellar ataxia [97]. Another expansion pattern of the exact repeat in ATXN10 was associated with early-onset levodopa-responsive parkinsonism [98]. Non-coding CAG expansions in HTT and SCA1 were proven causal in Huntington's disease and spinocerebellar ataxia type 1 [99]. Retrotransposition of SINE elements within an intron of TAF1 was associated with early stages of X-linked dystonia-parkinsonism [100].

The largest SV reference dataset was compiled and highly curated for several populations, including African/African-American, Latino, East Asian, European, and others. The dataset includes 433,371 SVs and is part of the Genome Aggregation Database (gnomAD) [101]. Other large-scale human genetic studies, such as the 1000 Genomes Project, Genome of the Netherlands Project, and Genotype-Tissue Expression Project (GTEx) revealed 68,818[55], 67,357[102] and 23,602[103] SVs, respectively. The accumulating short- and LR WGS datasets allow us to expand the genome-wide knowledge about structural variation and to improve the estimation of sequence-specific mutation rates and genome intolerance to non-coding SVs.

1.3 Discovery and genotyping of SVs

1.3.1 Chromosomal microarray analysis

Systematic and comprehensive SV discovery and genotyping are essential for investigating inherited human diseases. Microarrays are often used for CNV genotyping as a standard approach in clinical diagnosis as well as in genetic studies. In addition, the technology allows to identify loss of heterozygosity, mosaicism, and uniparental disomy events [104,105]. The microarray-based methods are represented by array comparative genomic hybridization technique [106] (array-CGH) and SNP microarrays [107]. The array-CGH experiment design is based on the comparison of labeled test and reference samples hybridized with long

oligonucleotides or bacterial artificial chromosome (BAC) clones on the same chip. The signal ratio between the test and the reference samples is then used as a proxy to estimate the genome material gain or loss in the test sample [108][109]. SNP microarray platforms are also based on the hybridization technique and are used for single-nucleotide and copy number variation detection between DNA sequences [110,111]. The advantage of the SNP microarray platform is the possibility of performing the allele-specific CNV calling [107]. Today, microarray chips available on the market are equipped with hundreds of thousands to millions of probes allowing for an accurate determination of the CNV breakpoints and detection of CNVs with up to 0.5kbp and allele frequency up to 0.5% [112,113]. Due to the CNV's high genotyping throughput and low-cost microarray analysis is actively used in the clinics for the diagnosis of inherited disorders such as autism [114], intellectual disability [115], neurodevelopmental disorders[116], and other rare genetic diseases[116,117].

The limitations of chromosomal microarray analysis are associated with the hybridization process and resolution, which is influenced by the coverage and density of the chosen microarray chips. The availability of many commercial array platforms loaded with different probe content and density made the method sensitive to the choice of analysis algorithm [118–120]. Due to the SV complexity and tendency to locate in the repetitive genomic regions and segmental duplications, accurate SV characterization remains difficult [121,122]. The microarray analysis ability to discriminate the signal coming from the duplicated regions drops significantly if the copy ratio does not match the expected diploid ratio [123]. The whole genome CNV screen is limited to variation with a size larger than the region between two adjacent probes. The microarray-based assays do not detect balanced rearrangements such as TRAs and INVs, and their detection ability is skewed towards the DEls [124].

1.3.2 Next generation sequencing and assembly-based approaches

The rise of the high-throughput sequencing era with the development of next generation sequencing technology (NGS) significantly accelerated SV discovery and annotation [55,125,126]. A previous study showed that SV calls from NGS are at least as sensitive as those from microarray genotyping [127]. Whole-genome sequencing (WGS) allows the entire genome hypothesis-free screen to detect balanced and unbalanced DNA rearrangements. NGS can also be performed on a specific region of the genome. For example, whole-exome sequencing (WES) and targeted genome sequencing involve the pre-selection of specific DNA sequences, enriching the DNA fragment library for the coding portion of the genome or specific target loci [128]. NGS data is becoming more widely used in clinical medicine to detect hereditary forms of different diseases, including neurological disorders [129–131].

Modern alignment algorithms start with building an index database from the reference genome and then query it for the read subsequences (seeds) to determine read global position in the genome. The next step includes a pairwise alignment between the read and each of the genome corresponding regions which can be implemented through different algorithms and heuristics such as Needleman-Wunsch [132] and Smith-Waterman [133–139] algorithms, Hamming distance approach [140,141], Dynamic programming [142], Non-Dynamic programming heuristic [143–146], and methods combinations.

The choice of the aligner and the alignment's accuracy is important for detecting SV signatures. The sequencing-based SV detection methods are based on the identification of abnormally oriented or split reads from the test samples after the reads mapping to the reference genome. The obtained alignment signatures and patterns are then used to determine the position, size, and type of SVs. This goal is tackled from several directions, which are used in a complementary manner by most of the current SV calling pipelines. The first approach is based on a so-called read pair concept, where aberrations in the distances between the mapped paired-end reads or in the read orientation are used as a landmark for the presence of structural variation [147–149]. For instance, a smaller insert size indicates the presence of a DEL. However, small SVs can be missed if the length is within the insert size

standard variability. Thus, it is important to combine several approaches. In the split-read approach, the events are detected where a part of the read was not mapped to the reference genome or mapped to another region (soft-clipped reads) [150,151]. The number of copies of certain genomic features can be assessed via a read-depth approach [152,153]. This algorithm records coverage changes within the sliding window, which screens through the whole genome or a target sequence. SV signature clustering and the solvent of the classification problem are performed during the final and the most crucial step of SV calling. This task is often approached with machine learning methods using alignment signatures, local nucleotide content, and alignment quality (MAPQ) as the input features [154–156].

Although a range of small SVs can be successfully resolved with the mentioned methods directly from the read alignments, larger SVs complicate the discordant read mate clustering process. Assembly-based methods suggest first assembling overlapping reads into longer fragments using the overlap layout consensus (OLC) and de Bruijn graph methods [157,158]. Whole genome assembly and local assembly can be utilized for SV calling. However, the whole genome assembly and alignment remain long and computationally intensive. In contrast, local assembly approaches decrease computational requirements and are often applied to detect SVs in a whole genome or a targeted fashion [159–161]. To increase the method power and sensitivity, the local assembly algorithms extract aberrantly mapped or unmapped reads to enrich the sequence subset in reads that support an SV. The variant-supporting reads are then assembled into longer contigs which are subsequently mapped to the reference genome [161]. Modern short-read (SR) SV calling pipelines include several approaches and tools to output reliable consensus results, which could be used for population genetics studies and clinical diagnostics [126].

NGS-based SV discovery comes with its limitations. For example, coverage-based SV detection approaches are complicated by GC content bias [162]. Currently, the typical read length for an NGS study ranges from 100 to 250 bps. This fact challenges the read alignment in regions with SVs, leads to the generation of high numbers of false positive results, and causes limitations to the accurate SV detection of larger sizes [163,164]. SVs within the repetitive regions or segmental duplications are usually underestimated within NGS datasets [126]. According to a recent

estimation, more than half of SVs are located in repetitive human genome sequences [68]. De novo assembly methods come in handy to increase the efficiency of SV calling. However, global assembly-based workflows remain computationally intensive, and local assembly-based methods often miss SVs near centromeres and within simple repeats [161]. Long reads can account for these challenges allowing more sophisticated SV calling within the problematic genome regions and reducing alignment ambiguity.

1.3.3 Long-read sequencing technologies

The third-generation sequencing technologies produce reads with an average length of more than a thousand bps. The Pacbio platform performs real-time sequencing of single DNA/RNA molecules through uninterrupted template-directed synthesis using fluorescently labeled nucleotides: single-molecule real-time sequencing (SMRT) [165]. Another actively used 3GS technique, the Oxford nanopore technology (ONT), is based on the measurement of system impedance changes induced by a single-stranded DNA/RNA molecule passing through a nanopore. The converted impedance changes are then converted to the nucleotide sequences during basecalling [166,167]. The development of the HiFi sequencing method allowed Pacbio to produce long-read (LR) sequencing datasets with an accuracy of more than 99.5% [168]. The ONT is less costly; however, it also yields less accurate sequencing results. Recent advancements in the nanopore system design and basecalling methods increase the sequencing accuracy reaching more than 90% [169]. For SV calling purposes, however, accuracy is dominated by sufficient coverage and read length [170].

Multiple studies reported a significant improvement in LR-based SV calling precision and recall (especially for SVs located in ‘dark DNA’), facilitating large-scale SV studies in various populations worldwide [68,163,171–173] (Table I). Approximately 25% of unresolved rare genetic disease cases can be recovered when longer reads are applied [174]. Interestingly, a recent population-scale study discovered that even though the majority of SVs were missed in the SR-based

analysis, more than 60% of LR-based SV calls can be genotyped and validated with NGS sequencing [175].

A recent hybrid SR and LR-based study presented a more comprehensive small variant and SV benchmarks for hard-to-access regions, including clinically-relevant genes [176]. A family-based study showed that the usage of LR datasets increased the recall of de novo small and structural variants by 20% [177]. 3GS technologies allow for a precise genotyping of short tandem repeat expansion (STR) expansions overcoming the challenges of low sequence complexity, large repeat size, and high GC content [178–181]. LR data were proven to be indispensable for the sophisticated resolvement and validation of complex SVs associated with Mendelian disorders [92,182].

The usage of one reference genome for all populations leads to the mapping bias towards the reference allele, subsequent incorrect SV genotyping, and inaccurate SV population frequency estimation [68,183]. LR technologies enable resolving this issue via an efficient construction of pan-genomes, graph-based structures which incorporate reference and alternative alleles, enabling more accurate read alignment [184,185]. However, large complex rearrangements remain to be resolved only by the whole genome assembly approach being completely inaccessible with standard benchmarking pipelines, including the sequencing-based methods [176]. The de novo genome assembly complexity is successfully simplified with ONT and SMRT datasets [186–188]. The LR sequencing datasets are often coupled with the SR and chromatin interactions information data (Hi-C) to improve the basecalling accuracy and produce more continuous error-free whole genome assemblies [188,189]. The Telomere-to-Telomere (T2T) Consortium took advantage of LR sequencing to resolve the complete human genome by releasing telomere-to-telomere gap-free haploid human assembly [190]. The Human Pan Genome Project leverages hybrid and pairwise assembly alignment algorithms using minmap2 and Winnowmap to capture genetic variation in different populations [191]. The project aims to create a global map for genetic diversity and facilitate the further annotation of the complete human genome and its application in evolution and disease research [191].

Table I. LR and hybrid-based population-scale structural variation studies.

Reference study	Sample size	3GS technology	Method used for SV discovery	Represented populations
Huddleston et al., 2017 [175]	2 (CHM13 and CHM1 cell lines)	ONT	Long read/whole genome assembly alignments	-
Chaisson et al., 2019 [171]	3 (HG00733, HG00514, NA19240)	ONT, SMRT	Long read /Haplotype-resolved whole genome assembly alignments	Native American, Han Chinese, African
Audano et al., 2019)[68]	15 (2 from Huddleston et al. (2017))	ONT, SMRT	Long read/whole genome assembly alignments	African, South Asian, Han Chinese, Vietnamese, European, Native American
Yan et al., 2021 [192]	15 (from Audano et al. (2019))	SMRT	Long read alignment	African, South Asian, Han Chinese, Vietnamese, European, Native American
Quan et al., 2021 [70]	25	ONT	Long read alignment	Asian (Chinese Tibetan and Han)
Ebert et al., 2021 [193]	35	ONT	Haplotype-resolved whole genome assembly alignment	South Asian, East Asian, Admixed American, African, European,
Beyter et al., 2021 [183]	3,622	ONT	Long read alignment	European (Icelandic)
Sano et al., 2022 [194]	15	ONT	Long read alignment	Asian (Japanese)

1.4 Structural variant annotation and clinical prioritization

While treatments are available for alleviating PD symptoms in some patients, no cure is currently available for the disease. Systematic screening of disease-causing variants and their effect on gene expression is required to bridge the gap between molecular and clinical phenotypes and to facilitate the identification of therapeutic targets for precision medicine. The genetic variant interpretation remains challenging given that not only pathogenicity estimation should be assessed, but also the true causal variant has to be prioritized among putatively pathogenic variation [126,195]. Estimation of the functional effects of SVs is even more complex due to their large size, which very often leads to the impact of several molecular targets. In addition, one has to take into account the mechanistic type of the SVs, since the deleteriousness effects do not result only from the DNA sequence alteration [196].

The standard procedure for the variant interpretation pipeline was designed and tested by a collaborative effort of the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP), who published their guidelines first for small variants [195] and later expanded them with CNVs in collaboration with Clinical Genome Resource (ClinGen) project[197] providing a joint consensus recommendation in the form of the semi-quantitative system [198]. The proposed categorization rules classify variants into five different classes: “Pathogenic”, “Likely pathogenic”, “Benign”, “Likely benign”, and “Variant of Uncertain Significance (VUS)” based on the different characteristics such as population allele frequencies, functional annotation, and predicted degrees of pathogenicity. A growing number of SVs datasets allow to obtain variant counts and frequencies across a variety of populations and clinical phenotypes (PS4 and PM2 ACMG/AMP criteria for strong and moderate evidence of pathogenicity), however additional evidence levels such as a consensus verdict from several *in silico* methods and functional variants annotation have to be systematically assessed (PP3 and PS3 ACMG/AMP criteria for supporting and strong evidence of pathogenicity).

A number of *in silico* approaches based on a scoring system or machine learning algorithms were developed to incorporate ACMG/AMP guideline criteria and automatically predict the pathogenicity and phenotype association [196,199–202]. Even though most of the algorithms exhibit good recall and specificity during the

benchmarking, the training datasets and positive controls are represented mainly by known highly detrimental coding SVs and “easy-to-detect-and-associate” regulatory variants leaving the method's ability to detect and prioritize cryptic causal non-coding SVs usually unassessed.

SV annotation benefits from the accumulated datasets of genome annotations, including coding and non-coding genes [203–206], epigenetic and regulatory activity [207,208], 3D genome organization [206,209], and conserved/constrained genome regions [210–212]. Functional variant annotation is becoming less challenging with the advancement of omics assay techniques and a growing number of publicly available omics datasets. For example, RNA sequencing allows for accurate measurement of genome-wide gene expression, alternative splicing events, and allele-specific expression [213–215]. Chromatin 3D organization and accessibility can be explored with Hi-C and ATAC sequencing assays [216,217]. Epigenetic markers can be used as proxies to assess gene and regulatory element regions' activity via methylation, acetylation, and other histone modification profiling [218–220]. Proteomics datasets obtained through mass spectrometry or single-molecule protein sequencing are indispensable to investigating the variant effect on the protein isoform molecular organization, function, activity, and interaction networks [221,222]. The development of single cell multiomics technologies [223] shed light on the specific cell types which are affected the most by the DNA variation leading to a better understanding of molecular mechanisms triggering the pathogenic processes and providing refined information about the targets for precision medicine.

It is essential to perform a semi-automated or manual assessment of individual candidate SVs using such tools as UCSC genome browser, Integrated Genome Viewer (IGV) [224], samplot [225], Ribbon [226], or svviz [227]. First, visualization of the read alignments helps to discard false positive calls and artifacts which were missed by the automatic quality control system. Second, the complement of SV loci with the gene annotation, functional regulatory elements associations, trait-associated regions, and omics data analysis results has occurred to be a powerful tool to trace the variant effect on molecular pathways and prioritize candidates for the genetic disease risk variation.

Before incorporation into the clinical diagnostic panel, causal SVs should be validated using standard approaches, including Sanger sequencing and PCR-based validation [228]. The findings have to be replicated in other unrelated and nonoverlapping cohorts of the minimal size, which depend on the power of the analysis methods [228,229].

2 Aim of my study

The aim of this study is to expand our knowledge about the role of the most understudied type of DNA variation in the genetics of PD via a systematic discovery and annotation of SVs in the cohort enriched for the familial and idiopathic PD cases. For this purpose, we used a long-read sequencing-based SV detection method, which already laid down an important foundation for SV investigation covering the full frequency spectrum and genomic regions so far inaccessible to other technologies.

3 Materials and methods

iPSC lines and donor background

The induced pluripotent stem cell (iPSC) lines were obtained from the Parkinson's Progression Marker Initiative (PPMI; https://www.ppmi-info.org/) and differentiated to the dopaminergic like state in the scope of The Foundational Data Initiative for Parkinson's Disease project (FOUNDIN-PD; (https://www.foundinpd.org/). 95 samples were included in the current analysis. The cell line collection included healthy controls (n=9), PD cases without mutations in known PD mendelian and high-risk genes (n=36), and PD-affected and unaffected individuals harboring pathogenic mutations, including LRRK2+ G2019S (n=25) or R1441G (n=1), GBA1+ N370S (n=19), T369M (n=1) or E326K (n=1), SNCA+ A53T (n=4). One iPSC line carries both LRRK2 G2019S and GBA1 p.N370S, and another iPSC line carries both LRRK2 G2019S and GBA1 T369M.

FOUNDIN-PD omics datasets

The molecular readouts protocol and data analysis for 95 iPSCs on day0, da25 and day65 are described in FOUNDIN-PD resource paper (Bressan et al., 2022, Unpublished manuscript)

Bulk RNA sequencing data generated on day 65 was used. A non-redundant genome annotation combined from GENCODE 29 and LNCipedia5.2 50 (https://github.com/FOUNDINPD/annotation-RNA) was used for the read counting. The analysis pipeline can be found here: https://github.com/FOUNDINPD/bulk_RNASeq.

Single cell RNA sequencing data generated on day 65 was used. The pipelines used in this study are available at https://github.com/FOUNDINPD/FOUNDIN_scRNA.

Bulk ATAC-seq dataset obtained from day 65 iPSC lines was used. The full analysis pipeline is stored here: https://github.com/FOUNDINPD/ATACseq_bulk

Methylation profiling dataset obtained from day 65 iPSC lines was used. The quality controlled data normalized using quantile normalization was included in the current exploratory analysis. The methylation profiling analysis pipeline is stored here: https://github.com/FOUNDINPD/METH

Nanopore sequencing

DNA has been isolated from the iPS FOUNDIN cell lines at day 0 by QIAamp DNA Mini kit according to the standard operating procedure provided by QIAamp. The extracted DNA was detected by NanoDrop™ 2000 spectrophotometer (Thermo Fisher Scientific, USA) for DNA purity (OD260/280 ranging from 1.8 to 2.0 and OD 260/230 is between 2.0 and 2.2); then, Qubit 3.0 Fluorometer (Life Technologies) was used to quantify DNA accurately. The Short Read XS Eliminator Kit (Circulomics) was used to size-select long DNA fragments. The sequencing adapters from the SQK-LSK109 kit were attached to the DNA ends. Finally, Qubit 3.0 Fluorometer (Life Technologies) was used to quantify the size of library fragments. The Nanopore PromethION 24 sequencer (Oxford Nanopore Technologies, UK) was used, one flow cell per sample. All samples were sequenced with 1D R9.4.1 nanopores. Each genome was sequenced to a minimum of 28X coverage depth.

RNA was successively purified and concentrated using the Qiagen RNeasy micro kit according to the standard operating procedure provided by Qiagen. RNA concentration was determined by Qubit HS RNA assay and RNA integrity was determined by analyzing 1ul on a RNA Tape on a 4200 TapeStation. Libraries were prepared using the ONT cDNA-PCR sequencing kit SQK-PCS109. This kit is suitable for low total RNA input amounts and generates high cDNA data output.

Processing of raw ONT DNA and RNA sequencing datasets

Fast5 files containing nanopore ionic current underwent basecalling analysis. Basecalling was performed with Guppy (v 4.4.1), a 3GS data processing toolkit that includes a neural network-based basecaller. Guppy was used to filter low quality

reads with a score less than 7. RNA sequencing datasets underwent the primer and adapter trimming, and read reorientation with Pychopper (v2.5.0). Quality control (QC) reports were obtained with NanoPlot (v1.32.1).

LR aligner benchmarking analysis

Benchmarking analysis was run against the Genome in a bottle (GIAB) SV truth set for HG002 consisting of 7281 sequence-resolved insertions and 5464 deletions within benchmark regions covering 2.5 Mbp [230].

HG002 ONT sequencing dataset was downloaded from https://github.com/human-pangenomics/HG002_Data_Freeze_v1.0. The raw FASTQ files were filtered for the minimal read quality equal to 7 and obtained reads were submitted to the SV calling pipeline. SV calling pipeline includes LRA or Winnowmap as aligner and CuteSV as LR SV caller.

The workflow script can be found here:
https://github.com/illarionovaanastasia/pipeline-structural-variation

Reads were aligned to the human assembly build GRCh38/hg38 build. The pipeline discovered simple SVs including INS, DEL, DUP and INV classes. Obtained SV calls were filtered for SVs located within GIAB HG002 benchmark regions. The GIAB HG002 SV truth set and benchmark regions were downloaded from ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/. Positional annotation of filtered SV calls and The GIAB HG002 SVs was performed with Variant Effect predictor (VEP). The GIAB HG002 SVs were clustered with the filtered SV calls using Jasmine [231] to detect common SVs and calculate sensitivity and specificity of the two tested versions of SV calling pipeline. Pipeline sensitivity and specificity were calculated using custom R (v 4.4.1) script with default R packages. Benchmarking results, SV counts and SV length distribution were visualized with R package ggplot2 (v 3.3.6).

SV calling on FOUNDIN PD dataset

The benchmarked SV calling pipeline included Winnowmap aligner and CuteSV variant caller. Reads were aligned to the human assembly build GRCh38/hg38 build. The pipeline was run sample-wise, then SV calls missing genotypes were removed. Obtained SV calls were clustered with Jasmine [231] to obtain the cohort-level SV callset.

Differential gene expression analysis

A list of SV-gene pairs was compiled to test for the variant-gene expression association. SV-gene pairs are defined based on SV annotation: intragenic SVs and SV within 5 kb gene flanking regions were included in the analysis. SV list was filtered down further to include only the SV-gene pairs which had the SV present in at least 3 and absent in at least 3 of the samples. For each of the SV-gene pairs, the samples were split into two groups: SV carriers (genotype 0/1 or 1/1) and non-carriers (genotype 0/0). Bulk RNA sequencing count results generated on day 65 were used to discover perturbed molecular pathways on the transcriptome level. Raw counts were analyzed in R (v 4.4.1) with DESeq2 which performs quantitative analysis of comparative RNA-seq data using shrinkage estimators for dispersion and fold change (v 1.30.1) [232]. PD status, iPSC line differentiation batch, donor sex, and age were included as confounding factors in the design matrix. FDR-adjusted p-value 5 % was used as a threshold for a statistically significant signal, logFC was calculated as log2 of the SV carrier to non-carrier normalized gene counts ratio.

Expression outlier analysis

The context-dependent outlier detection analysis was performed with R package OUTRIDER (v 1.8.0) [233] to reveal gene expression outliers based on the identification of significantly deviated gene counts in comparison to the count expectations in the given RNA-seq dataset. In brief, OUTRIDER fits a negative binomial model to gene read counts performing correction for variations in sequencing depth and known co-variations across the cohort. The OUTRIDER

autoencoder module was used to control for confounders. Bulk RNA sequencing count results generated on day 65 were used.

Differential transcript usage analysis

Reference-guided transcript annotation and quantification was performed with bambu (v 0.3.0) based on GENCODE (v.29) expanded with LNCipedia (v 5.2). Novel and annotated transcripts from annotated genes were used for the differential transcript usage analysis conducted in R with DRIMSeq (v 1.18.0).

PD GWAS hits and LD calculation

PD GWAS sentinel SNPs were obtained from Meta 5 PD GWAS [50]. LD regions for Meta5 SNPs were calculated using 1000 Genomes project European population (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.wgs.phase3_shapeit2_mvncall_integrated_v5c.20130502.sites.vcf.gz).

Plink 1.9 (https://www.cog-genomics.org/plink/) was used to obtain proxies for sentinel SNPs with R2 > 0.5

Segment liftover (https://github.com/baudisgroup/segment-liftover, 'First public version') was used to convert most left and most right sentinel SNPs proxies coordinated from hg19 to hg38 build. The chain file was downloaded from UCSC liftover http://hgdownload.cse.ucsc.edu/goldenPath/hg38/liftOver/

Bedtools (v2.26.0) were used to extract FOUNDIN-PD SVs located within obtain LD block coordinates.

Statistical tests and data visualization

Fisher exact test, Mann-Whitney U test and Chi Square test were run in R. Gene ontology enrichment analysis was performed in gProfiler (https://biit.cs.ut.ee/gprofiler/, Ensembl 106). Plots were visualized with ggplot2 (v 3.3.6) in R. Figures were arranged and annotated in BioRender (BioRender.com, DZNE license).

Data availability

- FOUNDIN iPSC lines are available upon request at https://www.ppmi-info.org/access-data-specimens/request-cell-lines/.
- Molecular assay protocols and generated data can be accessed at https://www.ppmi-info.org
- Data analysis code is available at https://github.com/FOUNDINPD and https://github.com/illarionovaanastasia/
- FOUNDIN data is available in the FOUNDIN-PD data browser located at https://www.foundinpd.org

4 SV calling: long read aligner benchmarking

4.1 SV calling summary statistics

Reads derived from highly repetitive loci or regions adjacent to unresolved gaps in the reference genome are prone to incorrect mapping which subsequently leads to an increase of false positive and false negative variant calling rate [234]. To investigate the effect of the read mapping pattern on SV calling of different SV lengths and in different genome regions, we conducted a benchmarking analysis comparing the performance of the two state-of-art long read aligners: lra [235] and Winnowmap [236]. CuteSV was used for the identification of SV signatures and SV filtering[237]. Benchmarking analysis was run against the Genome in a bottle SV truth set for HG002 consisting of 7281 sequence-resolved insertions and 5464 deletions within benchmark regions covering 2.5 Mbp [230]. ONT sequencing data from HG002 was mapped to GRCh37 with lra or Winnowmap. We first compared summary result metrics from the alignment step (Table1). An overall number of alignments reached 4,258,585 and 6,053,699 for lra and Winnowmap, respectively. The number of alignments passed MAPQ 10 reached 4,201,945 and 4,890,338 for lra and Winnowmap, respectively. Next, we checked the number of high-quality alignments within the repetitive regions. The number of mapped reads (MAPQ > 10) reached 3,957,088 and 4,357,175 for lra and Winnowmap, respectively.

Table 1. Comparison of mapping performance between lra and Winnowmap using HG002 ONT DNA sequencing dataset.

Aligner	Mapped reads	Unmapped reads (%)	Mapped reads with MAPQ > 10	Mapped reads with MAPQ > 10 within repeats
lra	4,258,585	33.02	4,201,945	3,957,088
Winnowmap	6,053,699	4.78	4,890,338	4,357,175

Alignments passed MAPQ10 were used for a subsequent SV calling with CuteSV (Figure 1, A). Obtained SV calls were filtered using the following metrics: SV type, minimum and maximum SV length, and a read support threshold defined automatically for each locus (see Materials and Methods section). The filtered SV set comprises 66,375 variants after lra mapping and 54,638 variants after Winnowmap mapping. Variant callset after lra has an imbalance of SV type frequency: a significant enrichment in DEL is observed in comparison to the SV calls after Winnowmap (Two-sided Fisher's exact test, odds ratio 1.4, p-value < 0.001) (Table 2). SV comparison against the truth set as well as between two aligners was performed using *truvari* without a sequence similarity comparison with the same SV type and with a genomic coordinate variance smaller than 1 kbp [238]. The common SV set encompasses 43,372 SVs (Figure 1, B).

Benchmarking results for SV discovery are depicted in Figure 1, C-D. The following measurements of method accuracy were used:

- Specificity, defined as the number of true positives (TP) divided by the sum of true positives and false positives (FP).
- Recall (sensitivity), as the number of true positives divided by the sum of true positives and false negatives (FN).
- F1 score, calculated as the number of true positives divided by the sum of true positives and the half sum of false positives and false negatives.

In general, CuteSV achieved a higher accuracy performance with Winnowmap: F1 score = 0.93 after Winnowmap mapping, F1 score = 0.89 after lra mapping (Figure 1, C). No significant difference between TP callset counts and total affected genome length was observed (Table 2).

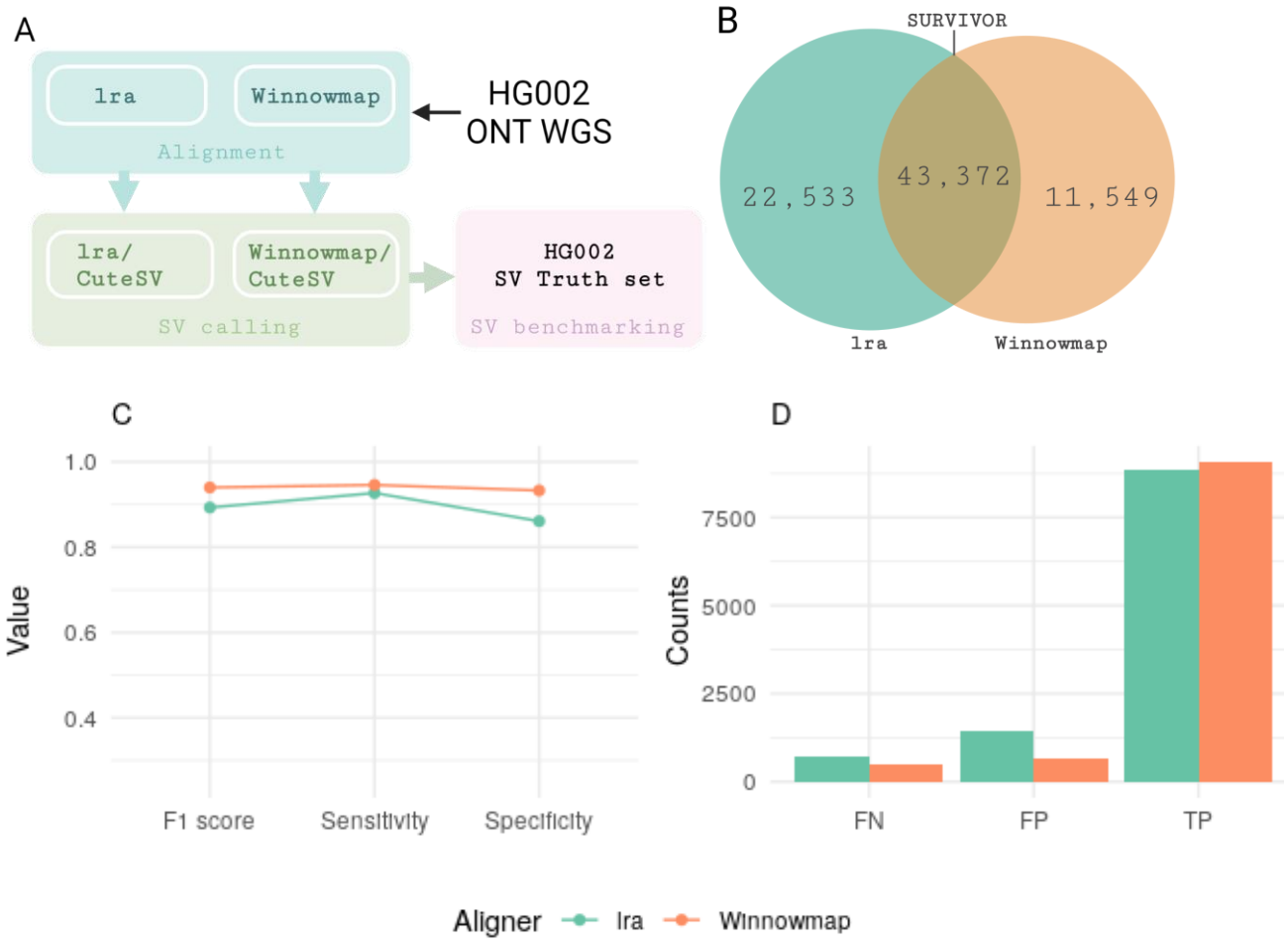


Figure 1. Benchmarking workflow and results of SV calling on GIAB HG002 truth set for lra+CuteSv and Winnowmap+CuteSV combinations. **A.** SV calling workflow. **B.** SV counts obtained after alignment with lra or Winowmap and an intersection SV call set. **C**. Value of F1 score, sensitivity and specificity. **D.** Counts of TP, FP and FN SVs.

Table2. Summary of SV callsets from Winnowmap and lra

Aligner	SV type	SV counts	Total callset size	TP SV counts	Total TP callset size
lra	DEL	45095	14.3 Mbp	4096	2.8 Mbp
	INS	21810	16.6 Mbp	4945	2.9 Mbp
Winnowmap	DEL	32697	9.3 Mbp	4127	2.9 Mbp
	INS	22224	8.8 Mbp	5161	2.9 Mbp

Singleton SVs, SVs which were called only after one of the two aligners, were analyzed separately (Figure 2). Winnowmap- and lra-specific call sets include 11,549 (5,858 DELs and 5,691 INSs) and 23,533 (5,277 INSs and 18,256 DELs) SVs, respectively. The lra-specific call set was significantly enriched in FP DELs and FN INSs (odds ratio 6.28, 4.37, p-value<0.0001). The Winnowmap-specific call set was significantly enriched in FP INSs (odds ratio 1.73, p-value<0.001).

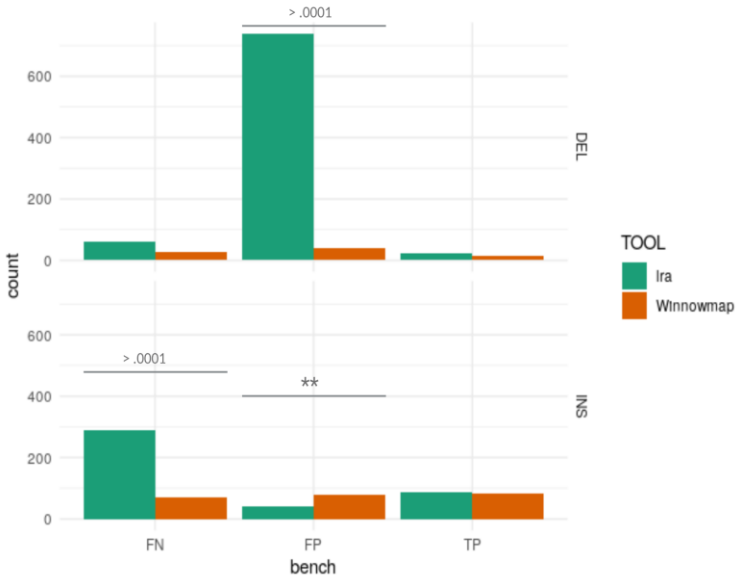


Figure 2. Distribution of TP, FP, and FN SV calls in the subset of lra and Winnowmap singletons.

4.2 SV size distribution and genome feature enrichment

The SV size may vary from 30 bp up to 100 kbp creating a demand for the high-resolution identification of the SV start and end breakpoints. SVs were sorted according to their length among the bins from 100 to 1000 bp bins with a step of 100 bps and from 1000 to 10000 bp bins with a step of 1000 bps. To investigate the effectiveness of different mapping approaches for the SV calling, benchmarking metrics were calculated for each bin and compared between two aligners (Figure 3, A-B). Sensitivity and recall for both aligners were above 75% for the majority of SV-size bins. The overall accuracy of the deletions calling of both aligners was above 88% for variants smaller than 1 kbp and above 86% for variants with a size varying between 1 kbp and 10 kbp. The general accuracy of the insertion calling of both aligners was above 80% for variants smaller than 1 kbp and above 75% for variants with a size varying between 1 kbp and 10 kbp. Both aligners showed a similar level of accuracy for deletions and insertions across the investigated length ranges with two exceptions. lra demonstrated a lower F1 score (F1 = 0.73) for the short deletions with a length up to 200 bp in comparison with Winnowmap (F1 = 0.95, Figure 3, A). The number of FP DELs in the bin of 100 bp are significantly higher than the number of SVs in the respective group called after Winnowmap (Fisher's exact test, odds ratio 0.14, p-value 2e-16). In contrast, Winnowmap shows a lower level of recall (Recall = 0.5) for large deletions with a length of around 10 kbp (Figure 3, B). Note that SV counts within the 10kbp bin were low (2 TP DELs after lra, 1 TP DEL, and 1 FN DEL after Winnowmap), which prevents any solid conclusions about the accuracy of both aligners for the mentioned bin size.

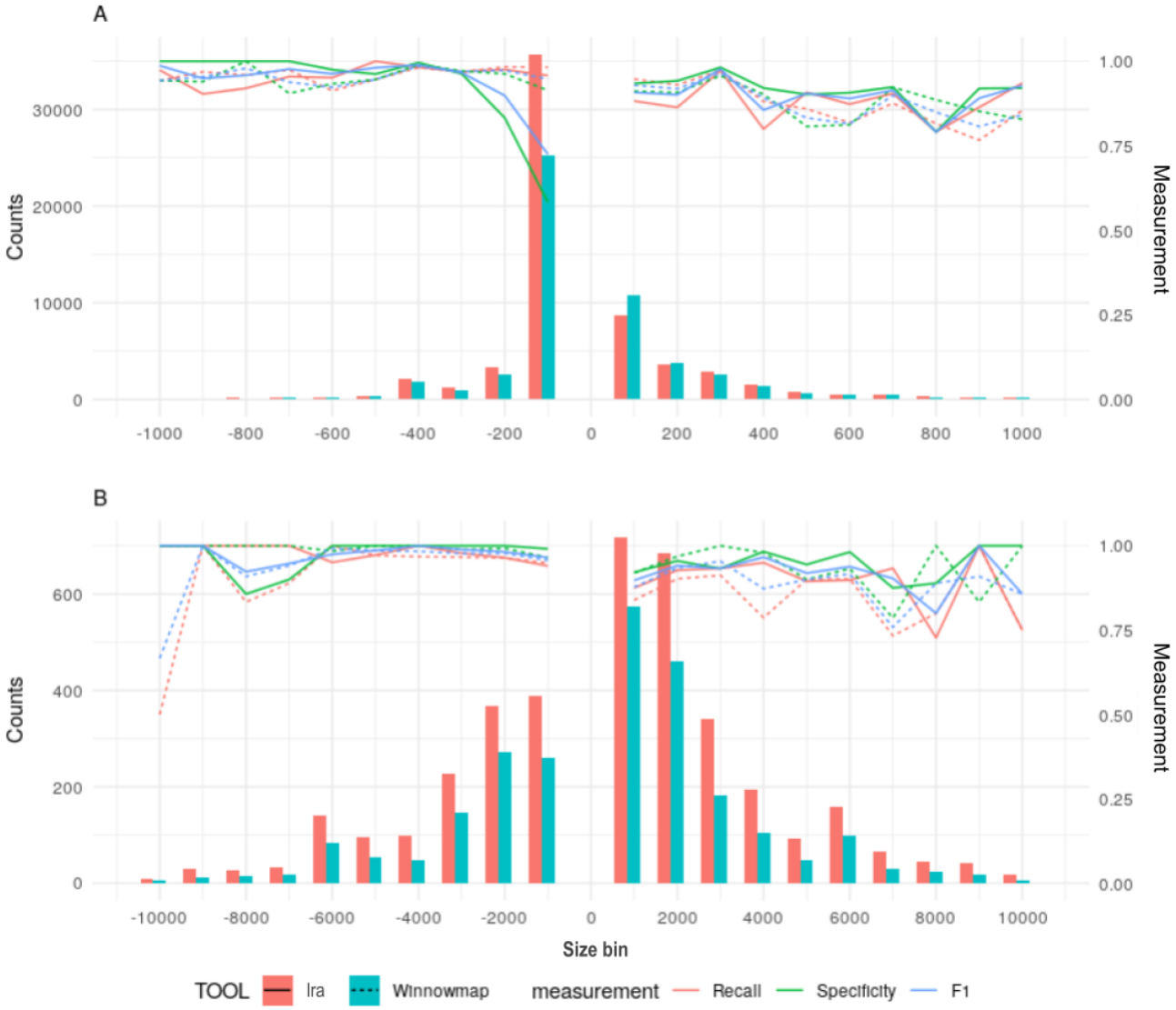


Figure 3. Distribution of SV counts and method accuracy measurements in different SV size bins. Negative and positive SV lengths refer to DEL and INS, respectively. **A**. Results are shown for SV size bins from 100bp to 1000bp. **B.** Results are shown for SV size bins from 1000bp to 10,000bp.

To investigate further lra and Winnowmap performance and specifically check for the potential SV calling biases related to the read alignment in specific genomic regions, we performed an enrichment test comparing SV distribution within a set of genomic features. A two-sided Fisher's exact test was performed within each

benchmarking group of SV calls (TP, FP, FN) for INS and DEL individually and for a unified call (Figure 4). SVs were annotated with the following genomic features: Gencode annotation (v19) of exons, introns, UTR, and intergenic regions, Repeatmasker annotation of repeats, SD regions, assembly gaps, and problematic regions (Materials and Methods Section). Fisher's exact test p-values were adjusted, controlling for the false discovery rate.

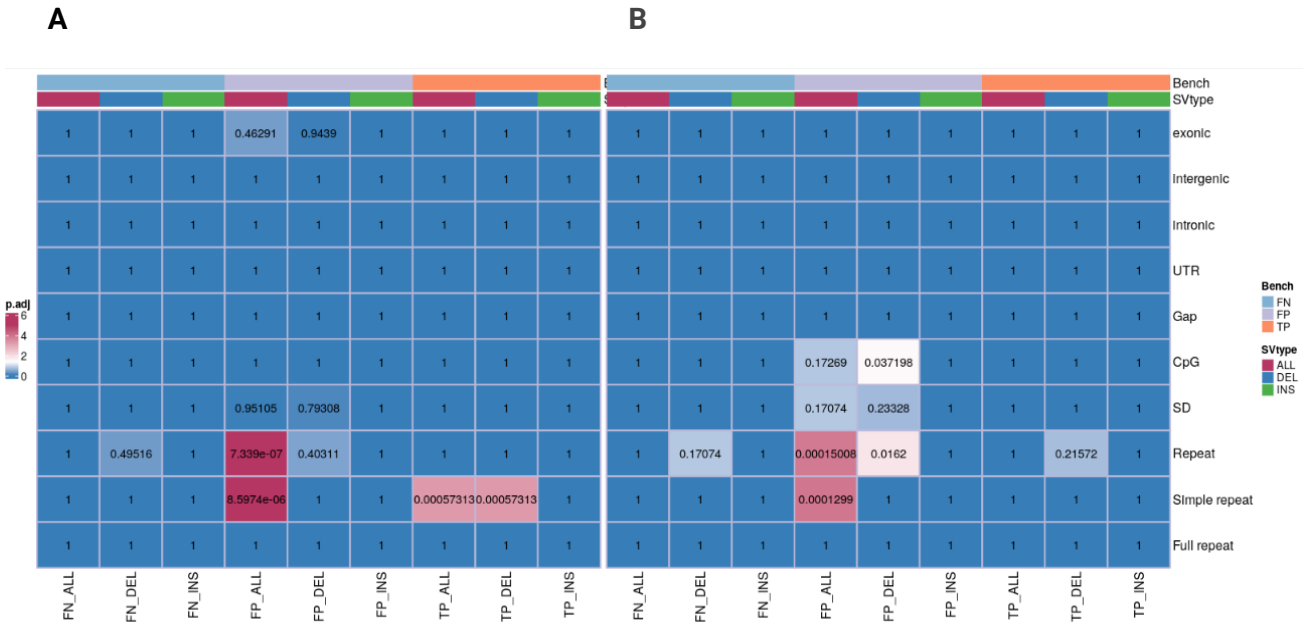


Figure 4. Heatmap demonstrating the p-value distribution for SV enrichment called after the WGS alignment with lra or Winnowmap. A. Unified SV call. B. Subset of singleton SVs.

FP SV calls from lra were enriched in simple and other repeats (odds ratio 2.09, 2.15; p-value < 0.05; Figure 4, A). At the same time, TP SVs prevail in simple repeat regions after lra mapping (odds ratio 1.16, p-value < 0.05). Other examined regions did not show statistically significant results for SV call enrichment. FP singleton SVs demonstrated the same pattern of enrichment as general FP SVs in simple and other repeats (odds ratio 1.28, 3.3; p-value < 0.05; Figure 4, B). We have additionally explored the origin of short FP DELs (SV length up to 300 bp). Short FP DELs occurred to be enriched in SINEs (short interspersed nuclear elements which include

Alu elements, primate-specific transposons) with an odds ratio reaching 2.5, p-value < 0.01. According to our observations, the Winnowmap-produced alignments decrease the number of false positive SV calls on the genome-wide level and within repetitive regions.

4.3 Discussion

Accurate SV detection is indispensable for a comprehensive investigation of genomic variation, which plays an essential role in clinical diagnostics [239][240]. LR sequence alignment-based approaches were shown to be gold standard tools for the efficient and reliable calling of simple and complex SVs [241]. Identification of SV breakpoints is highly dependent on the accurate detection of aberrant read alignment patterns, thus making a choice of the aligner to be a crucial step in the overall workflow. A recently developed long-sequence alignment tool, lra, is based on the seeding and chaining heuristic with a modified minimizer approach and outperforms most other long read aligners by the alignment metrics, variant discovery, and computational runtime [235]. However, the accuracy of SV calling within the repetitive genomic regions remained to be poorly assessed. A novel concept of the ‘weighted minimizers’ was claimed to be able to avoid excessive amounts of false-positive matches within repeats and maintain high alignment accuracy [236,242]). The algorithm was implemented as a stand-alone aligner Winnowmap designed based on the efficient seed and chaining algorithm of minimap2 [236,242]. In this study, we performed a benchmarking analysis comparing ONT SV calling after lra or Winnowmap alignment. We ran CuteSV for the variant calling because it was shown to have the best performance results after a number of state-of-art long sequences aligners including lra[243]. Both lra and Winnowmap follow a typical seed-chain-align procedure, however, the following key aspects distinguish the algorithms from each other:

1. Seeding. Winnowmap applies weights on minimizers, lra masks repetitive minimizers, and uses a second local minimizer index to refine chained anchors.

2. Chaining. Both aligners use the concave-gap penalty function, however, lra calculates the exact solution of seed chaining sparse dynamic programming, whereas Winnowmap uses a heuristic inherited from its predecessor minimap2.

Winnowmap mapped a higher read number than lra on a genome-wide level achieving at the same time a higher percentage of high-quality mappings (MAPQ > 10). The results are in agreement with the previous study where the minimap2 performed better in terms of alignment numbers than lra suggesting that the minimap2 chaining heuristic works more effectively [235,242]. The number of reported SV calls genome-wise after lra mapping prevails over the respective number of SV calls after Winnowmap, however, the Winnowmap algorithm outperformed lra by the number of TP SV calls within the benchmarking regions. In addition, lra alignments lead to a higher false positive rate both genome-wise and within repetitive regions. Since the human genome harbors 50-70% of repeats which affect different gene networks[244,245] the reduction of false positive variant signals from these regions is crucial for the sophisticated characterization of genome variation. The dissimilarity in aligner performance may be driven by the difference in dealing with minimizers coming from repeats. Recall and precision is higher for DELs in comparison to INS for both aligners which is replicated in general for other aligner-caller pairs[235]. However, this study demonstrates that the calling precision of small DELs (up to 300 bp) by cuteSV is almost 1.5 times lower after lra alignment. The variant length and origin coincide with that of the SINEs including the primate-specific Alu elements which point again to the suboptimal generation of mapping patterns within repetitive regions by lra algorithm. Observed results demonstrate that the weighted-minimizer-based approach outputs more accurate and informative alignments for the subsequent genome-wise SV calling.

5 SV detection and annotation on FOUNDIN-PD cohort: database construction

5.1 SV database construction

Structural variants vary in their size and sequence structure, causing a large impact on phenotype and triggering pathogenic processes[246]. Here we identified and annotated a set of reliable SVs across 95 iPSC lines that can be used for functional impact analyses in the context of PD. SVs calling was performed with the pipeline described and benchmarked in Chapter 4. The analysis was conducted individually for each sample, and the resulting SVs were clustered based on their type, genome position, and affected sequence length across the whole cohort (Figure 5, A). SV with missing genotypes and SVs which failed QC were filtered out before the clustering (on average 0.8% SVs per sample). This approach yielded a total of 150,809 SVs that met quality filters and became a basis for follow-up analyses.

The final SV set predominantly consisted of INSs and DELs with a size under 1 kbps (90.1%, Figure 5, B). SINE mobile elements and DNA satellite INSs and DELs peaks were marked at approximately 300 and 150 bps, respectively, as expected. A total length of non-reference INSs reaches 24 kbps (0.8% of the human genome) and is primarily represented by nonrepetitive sequences (45% of occupied sequence length), simple repeats (18%), LINEs (14%), and SINEs (11%, dominated by Alus: 10,8%). We detected consistent SV numbers and type proportions per sample genome (Figure 5, C). The median genome contained 32k SVs, with a cumulative length of the affected genome reaching 2.2%. Approximately 50% of SVs are singletons or rare variants (MAF < 5%). The prevalence of rare variants grows with the SV size: a group of SVs larger than 10 kbp is significantly enriched for the variants present in less than 5% of samples (p-value < 0.001, Chi-squared test) (Figure 5, D).

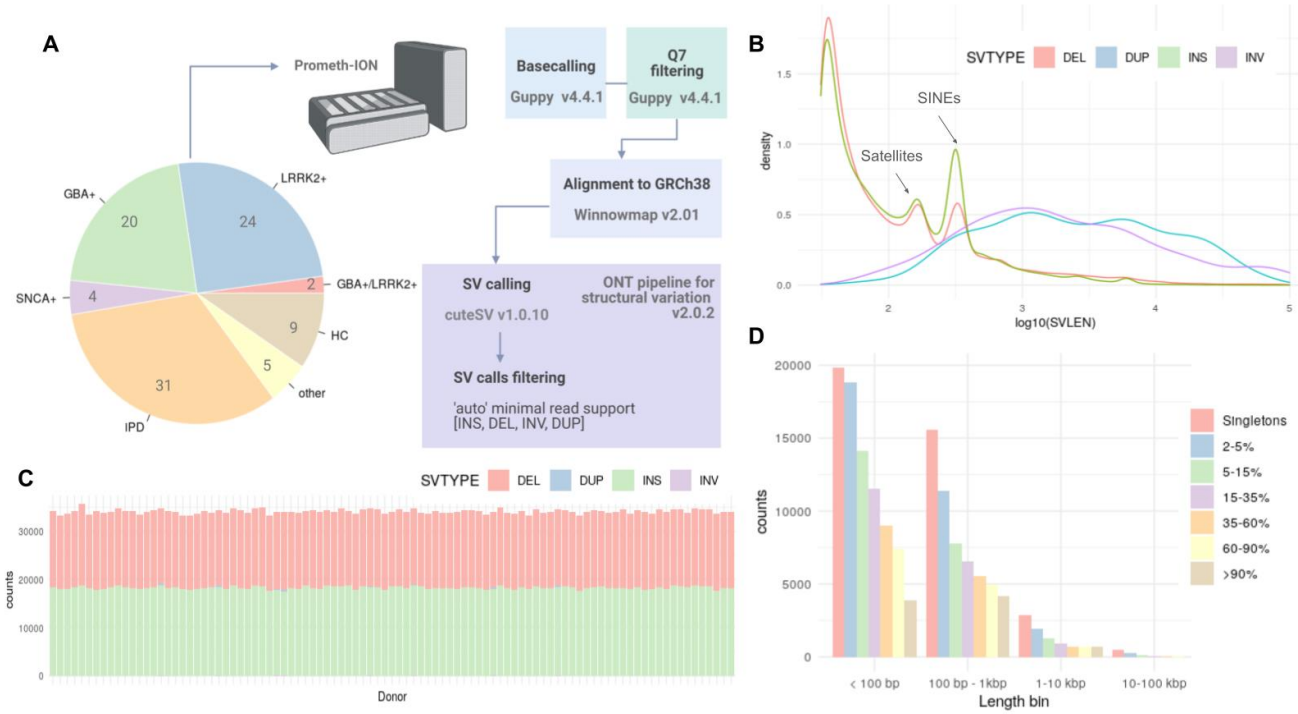


Figure 5. SV collection general results. **A.** SV calling workflow and sample PD genetic status. **B.** SV length distribution grouped by type. **C.** Per sample SV counts. **D.** SV frequency bin distribution grouped by size.

We accounted for potential FP SV calls from problematic genomic regions, such as gap-juxtaposed regions and regions within high somatic variability. We flagged variants located within segmental duplication regions and regions adjacent to gaps in GRCh38/hg38 build (12.7% and 4.8% of SV call set, respectively; Figure 6, A). DELs coincided with homopolymers, and SVs from HLA loci (cumulatively 3% of SVs) were removed from the further analyses (Figure 6, A).

Next, we assessed for putative iPSC-specific variation. Cells were collected for the LR sequencing assay on day 0; therefore, mutations originating from the mitotic divisions during the extended iPSC culturing are not expected to occur in the current dataset. Genetic changes can be introduced in iPCS lines during the reprogramming process. The reprogramming process can induce DNA double-strand breaks triggering activation of DNA repair mechanisms such as homologous recombination (HR) and non-homologous end-joining (NHEJ)[247,248]. Large chromosomal rearrangements were checked as a part of the FOUNDIN-PD iPSC line genome integrity screen [249]. However, smaller SVs (< 100 kbp) can still be present in the final SV callset and interfere with the interpretation of the results. A recent meta-study conducted extensive research collecting systematically recurrent CNVs in human iPSC lines [250]. We used 20 publicly available common abnormal regions, which cover 90.7% of all reported iPSC recurrent CNVs [250]. During the liftover from GRCh19/hg19, GRCh38/hg38 continuous coordinates from 17 regions were successfully obtained, comprising 81.4% of known iPSC-specific CNVs and spanning over a 7.9% of the human genome (Appendix table iPSC specific CNV regions GRCh38). We used the defined iPSC-specific CNV genetic hot spots to flag the variants located inside these regions. As a result, 4.5k CNVs were marked with a putative iPSC-specific variation label. No enrichment was found for a specific frequency group (common vs rare CNVs, MAF 5%) or for a specific CNV length group (long, >1kbps vs short CNVs, <1kbps) (Figure 6, B). Common CNVs under iPSC hot spots were depleted for variants affecting coding sequences and splice sites (Figure 6, C).

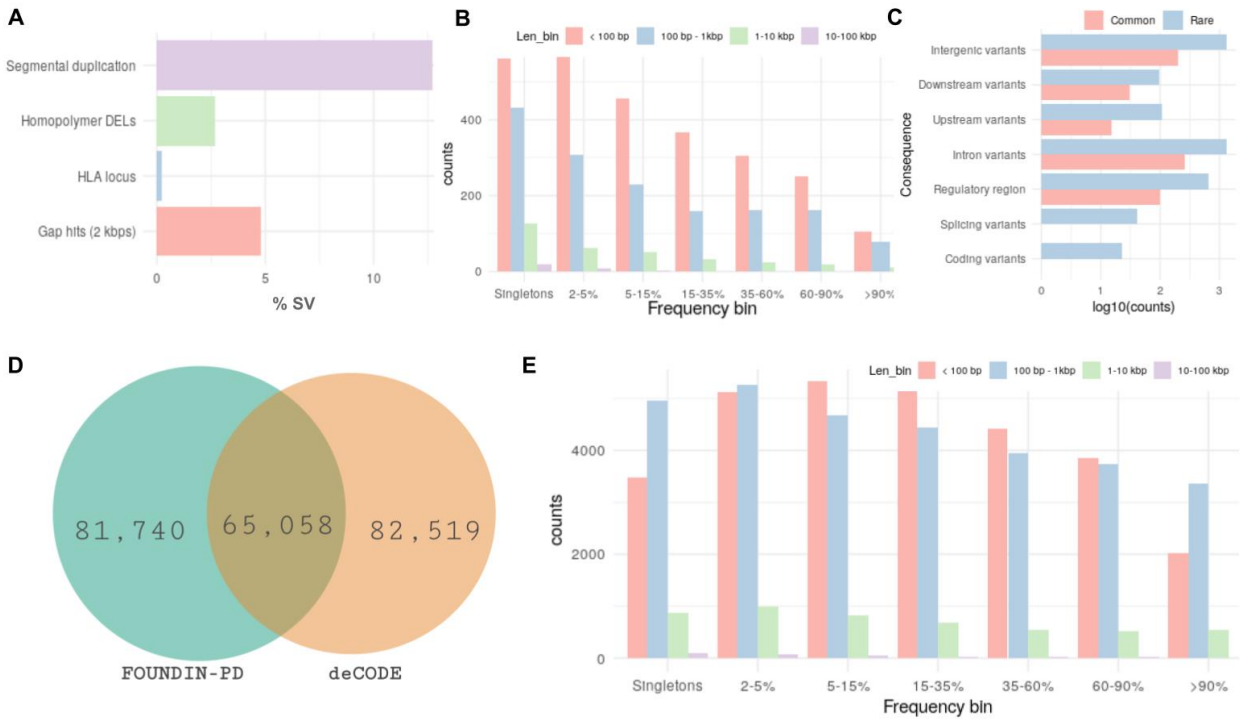


Figure 6. SV collection quality control and reference comparison. **A.** SV counts within problematic regions, including segmental duplications, homopolymer regions, HLA loci, and regions adjacent to gaps. **B.** Distribution of CNVs grouped by frequency and length under iPSC-specific CNV hot spots. **C.** Counts of CNVs under iPSC-specific CNV hot spots based on their positional annotation. **D.** SV callset comparison with deCODE SV collection: number of common and unique SVs. **E**. Distribution of deCODE/FOUNDIN-PD common SVs grouped by frequency and length.

SV collection was compared to the deCODE SVs callset where SV calling was carried out for ~3k Icelanders [183]. Both cohorts contain comparable SV numbers: 151k and 147k for FOUNDIN-PD and deCODE, respectively. After SV clustering based on their position, type, and length, we discovered 65k SVs shared by two cohorts which comprise 43-44% of SV collections (Figure 6, D). There was no detected over- or -underrepresentation of any specific frequency group or enrichment of a particular SV size group (Figure 6, E). We observed a more balanced representation of DELs vs. INSs in our cohort: 87% vs. 74% DEL/INS ratio for FOUNDIN-PD and deCODE SV call sets, respectively.

5.2 SV positional annotation and a prediction of their effect

SVs can span regions from hundreds to thousands of base pairs, thus affecting several genomic features e.g., one SV can be annotated to affect promoter, 5`UTR, and exonic regions. Positional SV annotation and variant consequence prediction was performed with Variant Effect predictor (VEP). We prioritized positional effects and assigned each SV the most severe consequence according to VEP annotation. The majority of SVs are located within intergenic or non-coding regions, including introns, non-coding exons, and regulatory regions (Figure 7, A). SVs are expected to alter the expression of multiple genes in their vicinity. Each SV affected or was annotated to, on average, 1.14 unique genes. The number of SVs affecting coding sequences (coding exons, splicing variants, frameshift variants, AF = 0.015) were notably lower than intergenic and intronic SVs (AF = 0.04) and enriched for singletons and rare variants (p-value<0.001, Chi-Square test).

SVs overlapping regulatory regions were one of the dominant groups (Figure 7, A). The number of SVs intersecting promoters, enhancers, and TF binding sites reaches ~7% of the SV callset following intronic and intergenic SVs. The majority of regulatory SVs are located within promoter flanking regions (7981 SVs, median MAF 0.04) preceding the number of SVs in CTCF binding sites (6263 SVs, median MAF 0.03) and open chromatin regions (4154 SVs, median MAF 0.05) (Figure 7, B). Notably, the median MAF of enhancer SV (0.02), promoter SVs (0.03), and SVs

within TF binding sites (including CTCF, MAF 0.03) are lower than intronic and intergenic SVs MAF, which reaches 4%.

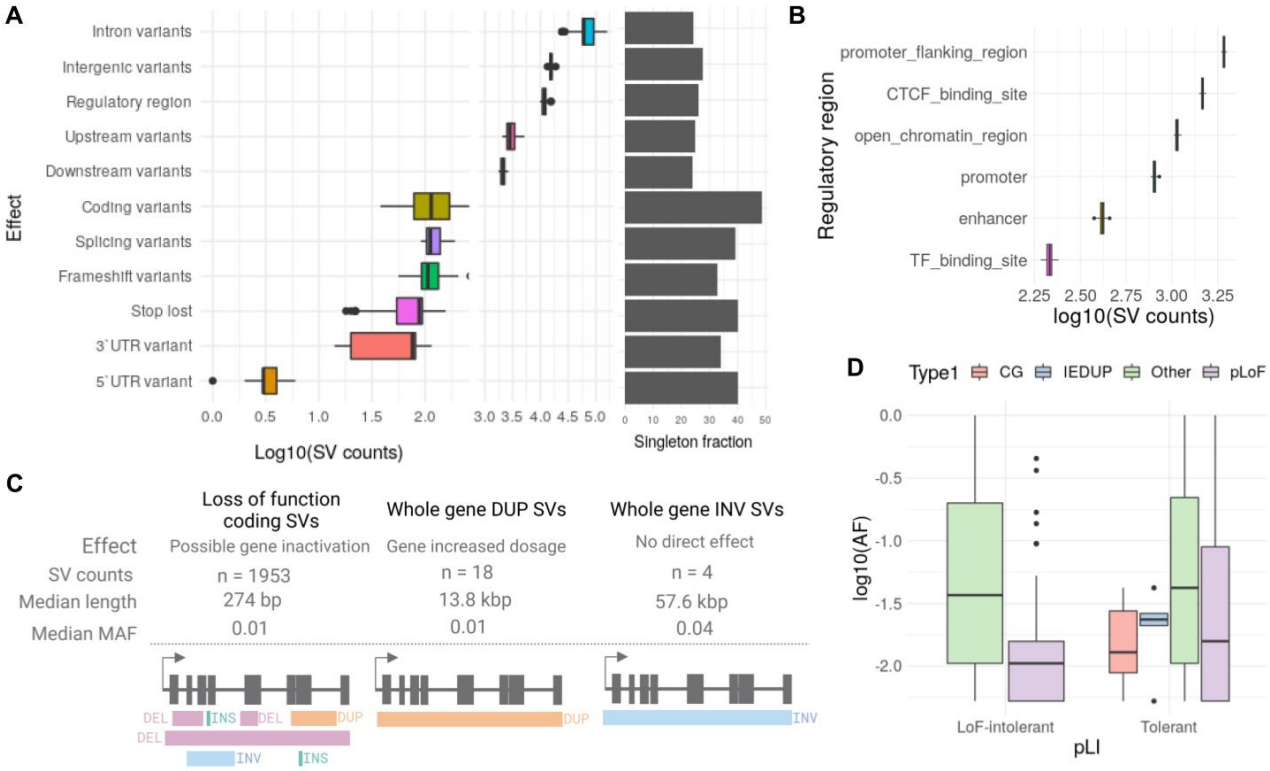


Figure 7. SV annotation according to their genome position and putative effect on gene expression. **A.** Distribution of SV effect groups and fraction of singletons within each group across 95 genomes. **B.** Distribution of SVs located within regulatory features across 95 genomes. **C.** Coding sequence altering SVs leading to gene inactivation, dosage increase, or lack of direct effect with counts of total SVs and median SV size. **D.** Distribution of SV MAF within LoF-intolerant and tolerant genes based on SV predicted effect. pLoF - loss-of-function, CG - copy gain, IEDUP - whole exon duplication, other - remaining predicted effects excluding intergenic SVs.

As it was suggested previously, any SVs altering coding nucleotides or altering ORFs are predicted to have a loss-of-function (pLoF) effect [126]. Whole gene DUPs are predicted to cause an increased gene-dosage effect whereas whole gene INVs are not expected to have any direct effects on gene expression. We detected 1953 pLoF SVs (median MAF 0.01), 19 CG (copy-gain) DUPs (median MAF 0.01), and 4 whole-gene INVs (median MAF 0.04) (Figure 7, C). When restricted to rare SVs (MAF < 0.05), we observed on average 16, 7, and 2 pLoF SVs, CG DUPs, and whole-gene INVs per genome respectively. In total, pLoF SVs affected 1863 genes, out of which 8,5% were pLoF-intolerant, including 874 protein-coding genes, 140 ncRNA genes, and 314 pseudogenes. We discovered that protein-coding pLoF SVs were significantly depleted for homozygous variants in comparison to pseudogene pLoF SVs (p value 0.0015, Chi-Square Test). On average, MAF of pLoF SVs affecting protein-coding genes and pseudogenes reached 0.015 and 0.02 respectively. The majority of pLoF SVs are represented by INS and DELs with a 1:2 ratio. Whole gene DUPs impact 38 genes including 32 protein-coding genes with an average MAF of 0.0052. We assessed SV distribution and predicted consequences in the LoF-intolerant protein-coding genes. We used the gnomAD upper bound of a 90% confidence interval of expected/observed LoF variation (LOEUF < 0.35) calculated per gene [251]. We obtained ~3.000 LoF-intolerant coding genes where SV set was compared versus variation in ~17.000 “tolerant” genes. We discovered that CG and exon duplications (IEDUP) are absent from the pLoF-intolerant genes (Figure 7, D). In addition, data shows the apparent tendency of pLoF SV MAF affecting pLoF-genes to be lower than pLoF SV MAF within tolerant genes although no significant signal was detected on the level of FDR 5% (p-value = 0.058, two-sided Mann-Whitney U Test). The fraction of SVs within intronic regions of LoF-intolerant genes reaches on average of 37.3%. In comparison, the fraction of exonic and splicing SVs affecting LoF-intolerant genes ranges on average between 0.8 and 4%.

The detrimental effect of pLoF SVs is usually avoided by the presence of the second functional gene copy unless the gene exhibits a high probability of haploinsufficiency (HI). We tested the hypothesis that genes harboring predicted loss-of-function SVs should not have a high HI probability. Indeed, the results demonstrate the decrease of pLoF SV fraction while walking up the HI gene probability percentiles with pLoF SVs being significantly depleted in the highest percentile (75-100%, p-value < 2.2e-16, Chi-Square Test) (Figure 8, A). We explored further if there is a difference in pLoF SV distribution between healthy controls and PD affected or prodromal cases (Figure 8, B). We calculated an SV prevalence for each of the two groups in the following way: variant was considered as "PD-prevalent" if healthy controls MAF < 0.05 and PD-cohort MAF > 0.05 (opposite conditions had be to meet for the "HC prevalence" and other combinations were considered as "No prevalence"). We discovered that while intronic and up/downstream variants do not show any difference in their target gene HI probability distribution among SV prevalence groups, rare in general cohort PD-pLoF SVs are observed in genes with a significantly high HI probability (p-value 0.006, two-sided Mann-Whitney U Test). The most significant molecular pathways enriched with genes affected by the given SV subset are signaling transduction, nervous system development, and axonal guidance, while cellular compartments are highly enriched for the neuron-specific subcellular parts such as axons, synapses, and dendrites (Supplementary Figure S1).

Human-lineage-specific regions defined as constrained non-conserved regions (CNCR) were shown to be enriched for the neurological and neurodevelopmental diseases associated variation. Exploration of FOUNDIN-PD SV datasets reveals that 8.35 % of SV overlap CNCRs including 282 pLoF SVs and 7 CG DUPs. We found a significant depletion of pLoF SVs (p-value < 0.00001, Chi-Square Test). In addition, we discovered a significant correlation between the CNCR score of SVs and the target gene biotypes as well as regulatory features (Figure 8, C). The protein coding gene and lncRNA biotypes were depleted for the highest percentile of the CNCR score, while interestingly the enhancer and CTCF sites were enriched for the same CNCR score range.

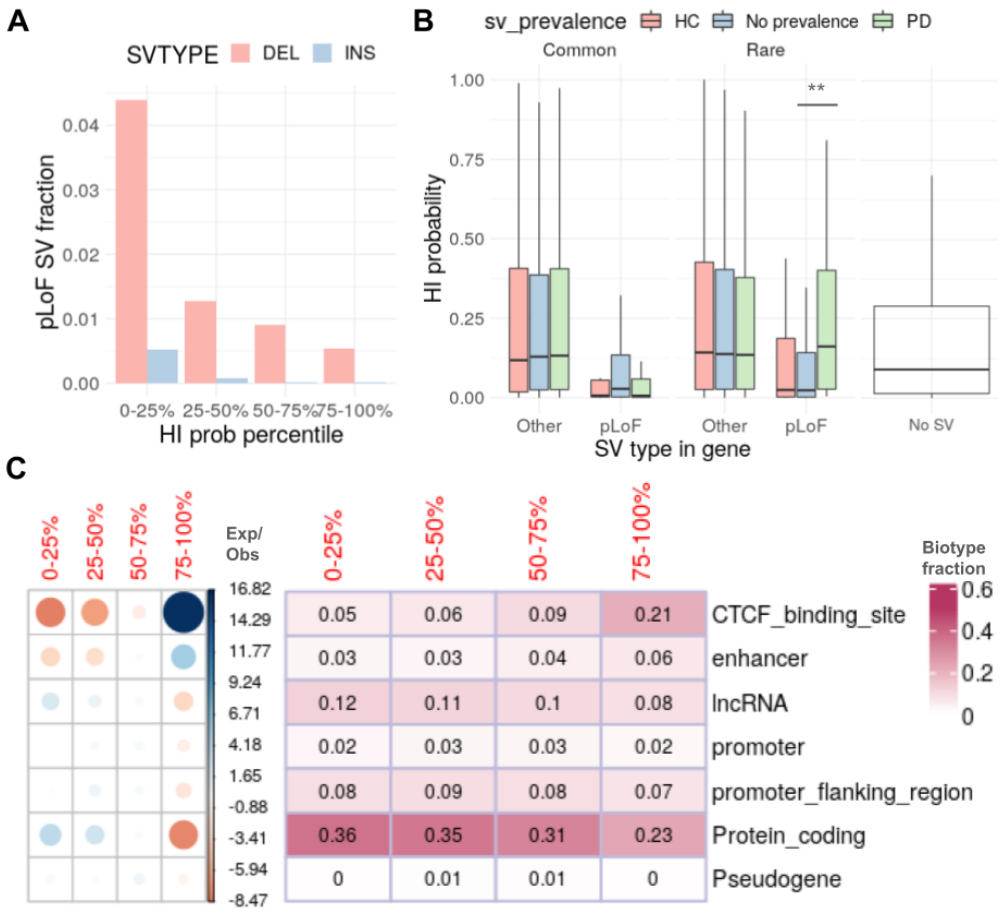


Figure 8. Properties of SVs affecting haploinsufficient genes and human lineage-specific regions. **A.** Presence of pLoF SVs within genes with different HI probability percentiles. **B.** HI probability distribution for genes harboring pLoF and Other (non-pLoF, excluding intergenic) mutations and for genes without observed SVs across HC and PD SV prevalence groups. The variant was considered as "PD-prevalent" if HC MAF < 0.05 and PD-cohort MAF > 0.05 (opposite conditions had be to meet for the "HC prevalence" and other combinations were considered as "No prevalence") **C.** Observed vs expected ratio of biotypes with SVs overlapping CNCRs and their fraction across different CNCR score percentiles.

During our next step, we focused on SVs localized within PD-associated loci. Since all individuals from the FOUNDIN-PD cohort have European ancestry, we used the largest PD meta GWAS which discovered 90 loci in the European population [50]. We obtained 542 SVs (322 INS, 245 DEL, and 2 DUP, 0.37% of total callset, 223,407 bps of cumulative length) within the LD regions of 90 GWAS loci (R2 > 0.5 in EU populations). A major part of the SV callset within PD GWAS hits was represented by intronic and regulatory region SVs, followed by intergenic, intronic, and coding variants (Figure 9, A). The predicted variation which drives the GWAS signal should be common and prevalent in the PD cases cohort. We explored the SV MAF for PD and HC prevalence SV groups. The groups were determined as explained above, briefly, PD-prevalence means that the variant is common in the PD-affected and PD high risk mutation carriers and rare in healthy controls, while for the HC-prevalence the opposite conditions have to be true. In general, we did not catch a significant difference for most SV consequences except for intronic SVs, those MAF within the PD cohort were significantly higher under PD GWAS regions (p-value 0.003523, two-sided Mann-Whitney U Test). We specifically focused on the PD-prevalent SVs, investigating the consequence, and affected gene biotypes (Figure 9, C). The prevailing consequences of the given SV subset were intronic and regulatory region variation. Interestingly, both common and rare intronic SVs in the general cohort targeted protein-coding genes, but only rare SVs were found in the introns of lncRNAs. The coding variants were represented by one DEL overlapping a CTCF binding site and a pseudogene exonic region. The regulatory regions affected by PD prevalent SVs were mainly promoter and promoter flanking regions of protein-coding, RNA genes, and pseudogenes.

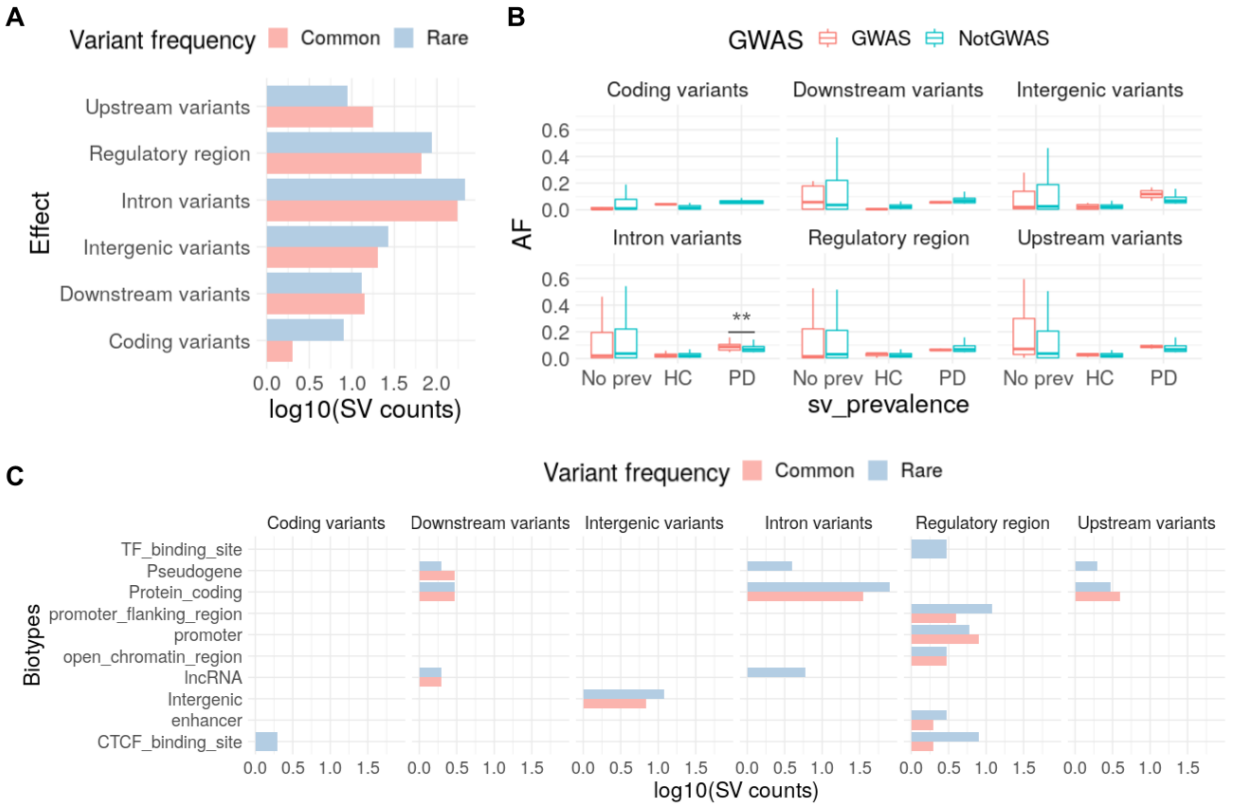


Figure 9. SVs under PD GWAS hits. **A.** SV consequence and variant frequency in the FOUNDIN-PD cohort. **B.** Distribution of SV MAF within HC and PD prevalence cohorts under and outside PD GWAS regions (R2 > 0.5 in EU populations). The variant was considered as "PD-prevalent" if healthy controls MAF < 0.05 and PD-cohort MAF > 0.05 (opposite conditions had to be met for the "HC prevalence" and other combinations were considered as "No prevalence"). **C.** SV from the PD-prevalence subset (MAF PD > 0.05) and their highest gene consequences and affected gene biotypes. Common and rare (MAF < 0.05) in FOUNDIN-PD cohort groups were compared.

The previous SR-based SV metastudy (which incorporates a subset of FOUNDIN-PD patients) identified eight SVs from WGS of patient blood under PD GWAS hits (Billingsley et al., 2022, unpublished manuscript). Three SVs were *in silico* confirmed in the LR-based SV callset from blood DNA sequencing and detected in the current study. Three DELs are Alu mobile element deletions located upstream of gene *ZSCAN9*, downstream of gene *NEK1*, and in the third intron of gene *LRRN4*.

5.3 Discussion

Genetic research capturing the complete variation landscape in individual genomes from patients with different phenotypes is essential for profound clinical diagnostics. In this thesis, we have conducted whole genome SV calling in the cohort enriched for PD phenotype which includes both familiar and idiopathic cases. The use of LR sequencing and high (more than 28x) coverage allowed us to map common and rare SVs at high genomic resolution and predict the SV effect highlighting putatively deleterious variants. We identified on average 32k SVs per genome, which exceeds the SV number in the recent large-scale LR-based dataset (median 22k per genome) [68,183]. This observation can be explained by the different minimal SV sizes used for the final SV callset generation: we lowered boundaries defined for SVs and started with 30bp length to capture small repetitive DELs and INSs, which are normally filtered out during the NGS InDel calling process. One of the major classes of SVs occurred to be CNVs (92 %) which agrees with the previous results [126,252]. The majority of SVs increasing genome length is INSs which contrasts with SR-based SV studies where numbers of INSs (predominantly mobile element INSs) and DUPs are similar, pointing out a potential LR-biased misclassification of DUPs as INSs due to the basecalling errors which artificially decreases the sequence similarity. We detected more INSs than DELs which aligns well with the results from LR-based SV studies [68,183] but differs from the results based on SR SV calling [251]. The observations implicate that NGS technologies allow easier identification of DEL breakpoints in comparison to INSs which might arise due to the repetitive nature of inserted sequences and insufficient long read length to cover

and identify the non-reference sequences. We replicated the finding that the SV length grows with the drop of MAF [55,68,246,251]: less than 200 SVs >10kbps were identified which were enriched for rare variants. The finding is a piece of clear evidence for the strong positive correlation between the DNA rearranged amount and the natural negative selection.

We systematically analyzed predicted SV effects on genomic features to facilitate further SV functional annotation and genotype-phenotype association. A negative selection leads to low frequencies of deleterious variants, enabling the use of frequency estimates as a proxy factor for the assessment of a variant's negative impact on the phenotype [251]. The dataset is highly enriched for intergenic and non-coding SVs, with the fraction of singletons rising from non-coding to coding loci due to the natural negative selection. We showed that SVs within coding regions and SVs overlapping promoters, enhancers, and CTCF binding sites were observed at lower frequencies than other non-coding SVs supporting earlier findings, which prioritizes protein coding exonic and regulatory region SVs based on their profound detrimental effect [252]. We also identified that the whole gene/locus DUPs are mostly rare variants and depleted for common and homozygous SVs. This finding is in line with the suggested deleterious effect of large DUPs, which affect and rearrange the TADs, thus interfering with the gene expression regulation [253].

Coding SVs with a predicted LoF effect were enriched for singletons and rare variants, pointing out the highly detrimental effect and high penetrance of SVs overlapping protein-coding exons. This conclusion is also supported by the fact that LoF-intolerant genes are affected by rarer pLoF variants than tolerant genes. The LoF intolerance was calculated from the expected vs. observed number of non-synonymous SNPs. Obtained results tell us that genes intolerant to LoF small variants are intolerant to LoF-SVs as well. However, it is assumed that SNP-tolerant genes may be not tolerant to pLoF SV, thus to assess SV-specific intolerance, LR-based SV calls from larger cohorts, including related individuals and different populations, have to be explored [126]. We also showed that the fraction of pLoF variants drops drastically with the growth of gene probability to be haploinsufficient

since negative selection reveals itself more intensively for the genes where the unaffected copy cannot fully compensate for the lost function.

Further investigation of SV rare in the general cohort but common in the PD affected and PD high risk mutation-positive individuals revealed that pLoF SVs, on average, affect HI genes with a significantly higher MAF. Affected molecular pathways and cellular compartments with enrichment for these genes are related to the nervous system development, organization, and functioning. Obtained results are expected to appear regarding HI genes in general. However, an increased frequency of detrimental variation in HI genes, which we observe in the PD cohort, implies the presence of rare pLoF SVs in a unique combination in PD-affected and prodromal cases leading to the decrease of genome fitness and increased vulnerability to pathogenic processes.

Identification of human-lineage-specific elements provided more insights into human neurological disease genetics, demonstrating that combined usage of constraint and non-conserved metrics significantly increases the information gain about the functionally important genomic features [211]. The study which introduced CNCRs already showed that these genomic regions are enriched for lncRNA genes and regulatory elements [211]. It was shown that the proportion of protein-coding exons overlapping CNCRs remains the same, and the ratio of protein-coding intron and lncRNA exons rises from low to high CNCR score distribution percentiles [211]. However, we observed that protein-coding genes and lncRNA genes harboring exonic and intronic SVs are significantly depleted from the high percentiles of the CNCR score distribution. This finding indicates that protein-coding intronic regions and lncRNA genes are under human lineage specific negative selection and highlights intronic and lncRNA SV overlapping CNCRs as candidates for further disease association.

Recent PD GWA studies identified 90 loci in European and 2 loci in Asian populations. However, we still have a limited understanding of the genetic variants that drive the association signals. In contrast to familiar cases, variants under GWAS hits are expected to have low risk effect size and cumulative contribution to the disease development. We discovered that the frequency of PD-prevalent intronic

SVs is significantly increased under PD GWAS hits, which supports the hypothesis of the several causal intronic SVs with a small risk effect size. We expanded a list of candidate SNPs and InDels with SVs located under PD GWAS hits that could potentially drive the observed association signal. Previously identified SV hits within PD associated loci (Billingsley et al., 2022, unpublished manuscript) were found in our study. Follow-up functional annotation and replication of findings have to be performed for further SV candidate prioritization.

6 SV functional annotation and PD risk factor prioritization

6.1 SV impact on nearby gene expression

The advent of SR and LR genome sequencing technologies facilitated SV detection in different populations on a large scale [126,183]. Despite this progress, the functional impact of SVs remains to be understudied. The sophisticated SV functional annotation is essential to understand the role of SVs in phenotypic manifestation in general and assess functional evidence of candidate variants in the scope of PS3/BS3 ACMG/AMP guidelines criterion for variant clinical interpretation [254]. SVs can affect the expression of nearby genes by altering the coding or cis-regulatory sequence of a target gene [103]. To explore the impact of SVs on gene expression, we used bulk RNA sequencing data from FOUNDIN-PD collection performed on day 65 of iPSC lines differentiation into a dopaminergic neuronal-like state [249].

To quantify the transcriptional consequences of SV, we first assess the gene relative expression as the median expression of SV carriers divided by noncarrier expression defined in counts per million (cpm) and converted to logFC under a hypothesis that SV with a greater probability will affect the nearest gene. First, we investigated the difference in gene expression harboring pLoF and CG SVs (Figure 10, A-B). Here we also investigated the functional effect of CNCR regions, testing a hypothesis that variation interfering with CNCRs should cause a larger impact. Indeed, we observed that pLoF SVs overlapping CNCRs on average show significant downregulation of the target genes (mean logFC = -0.2, p-value < 0.00001, two-sided Mann-Whitney U Test). CG DUPs not overlapping with CNCRs demonstrated an insignificant upregulation (logFC = 0.21), whereas logFC distribution shifts of genes with other SVs, including intronic, regulatory, and up/downstream variants equal to 0 (Figure 10, A). We break further these SV groups studying the putative effect of SV types and consequences (Figure 10, B) separately. We classified pLoF variants into coding

(purely exonic) and splicing (affecting exon-intron sites) SVs. We discovered that observed downregulation associated with pLoF SVs was specifically driven by coding (intra-exon) variation, while genes harboring splicing variants were up- and downregulated. In addition, we observed that transcript ablation variant genes were on average downregulated (logFC = -0.56) regardless of whether the associated SV overlaps the CNCR or not. Interestingly, no INS were found to be within CNCRs, and no observed effect was detected of any of the INS consequences.

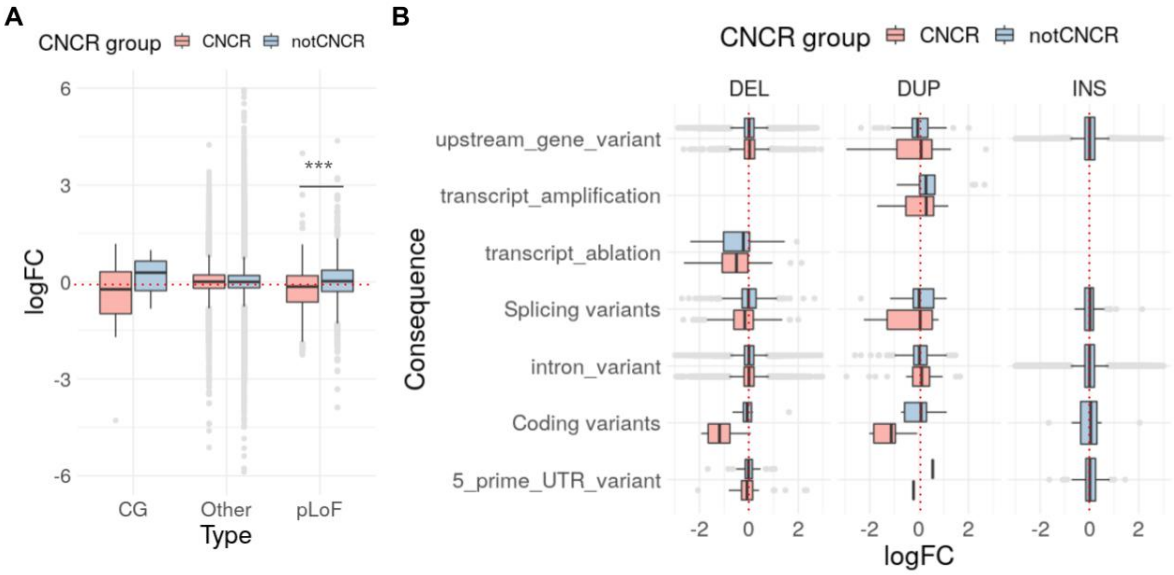


Figure 10. Change in target gene average expression between SV carriers and non-carriers expressed in logFC. **A.** Effect of different SV groups overlaid with the information of CNCR overlap. **B.** Effect of different SV types and consequences on gene expression based on CNCR overlap. pLoF - loss-of function SVs including coding and start/stop codon loss SVs; CG - whole gene duplications; Other - other SVs including intronic variants and up/downstream SVs. The dashed red line indicates logFC 0.

Next, we set out to explore the changes in gene expression that associated regulatory elements were affected by SVs. We started with the promoters of the protein-coding genes (Figure 11, A). A high confidence set of promoters from the ENCODE database was used. We discovered 92 unique SVs overlapping promoters of 89 unique genes [255]. On average genes with transcripts ablation, splicing, exonic and

5`UTR variants showed a consistent downregulation (logFC -2.3 to -1.1), while genes with transcript amplification and variants in the introns of their other transcripts were on average upregulated (logFC 0.22-0.55).

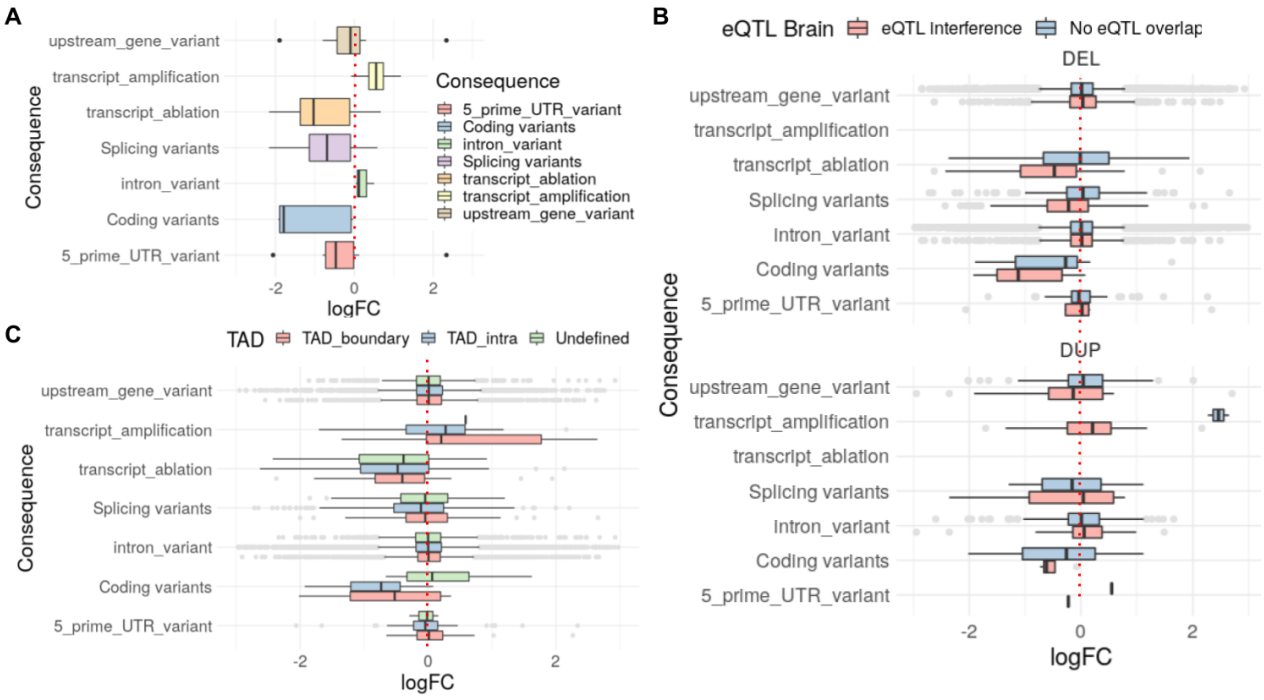


Figure 11. SV disruption of regulatory regions and 3D chromatin structure. **A.** Effect of SV overlapping promoter regions of protein-coding genes on the target gene expression. **B.** Predicted consequence and gene expression effect of SVs based on their overlap with cis-eQTLs active in brain regions. **C.** Predicted consequence and gene expression effect of SVs based on their TAD intra or boundary localization. The dashed red line indicates logFC 0.

We explored SVs affecting cis-eQTLs under a hypothesis that deletion or duplication of associated cis-eQTL should also affect the expression of the cis-eQTL-associated gene. For this purpose, we used eQTLs GTEx V8 release, which

includes significant eQTL-gene pairs for 18,262 protein-coding and 5,006 lncRNA genes. We discovered 3,635 DELs and DUPs interfering with cis-eQTLs active in different brain regions, including the substantia nigra, basal ganglia circuit, and cerebellar cortex. On average, the median MAF of SVs interfering with cis-eQTLs (median MAF = 0.01) is lower than the MAF of eQTL-nonoverlapping SVs (median MAF = 0.03), which is replicated for all SVs types. We checked if any expression changes effects were observed for the CNVs overlapping brain cis-eQTLs (Figure 11, B). On average, genes with whole transcript deletions and splicing DELs and DUPs overlapping eQTLs were significantly downregulated (logFC -0.3 - (-0.8), p-value < 0.01, two-sided Mann-Whitney U Test). Specifically, genes with splice sites affecting DELs overlapping eQTLs were significantly downregulated than genes with eQTL-non-overlapping DELs (p-value 2.42e-07, two-sided Mann-Whitney U Test). This observation was not replicated for the genes with splice affecting DUPs, which demonstrated various expression changes regardless of the cis-eQTL overlap. Genes with exonic CNVs were systematically downregulated regardless of eQTL interference. Genes with intronic and up/downstream CNVs did not show any consistent up or downregulation, with an average logFC equaling 0.

SVs are expected to disrupt gene expression regulation through direct interference with regulatory elements and via perturbation of TADs and rewiring associated gene-regulatory region pairs. We explored if there is an observable difference in gene expression with SV disrupting TADs (overlapping the TAD boundary) and located inside the TADs (Figure 11, C). We used estimated TAD regions from Hi-C sequencing of A549 and Caki2 human cell lines [207]. The SV was predicted to disrupt TAD if it overlapped the TAD boundary in both datasets. SV was predicted to localize within the TAD region if both datasets supported it. Otherwise, SV was considered to have an undefined impact on TAD. On average, genes with CNVs inside TADs and on their boundaries did not demonstrate a significant change in the expression level in comparison to genes with CNVs with undefined impact on TADs, except for genes with coding CNVs inside TADs and on their boundaries which were significantly downregulated in comparison to TAD undefined CNVs (p-value 0.02993, two-sided Mann-Whitney U Test).

Exploratory analysis of gene expression changes revealed interesting patterns based on the consequences of present SVs, however further statistical analysis has to be performed to infer the real differential signal accounting for the data variability. We run a differential expression analysis to capture statistically significant SV-gene expression change association. The candidate SVs were chosen according to the following criteria:

- SV presence (0/1, 1/1) and absence (0/0) at least in three samples
- SV candidate for a cis-regulation: intergenic SVs excluded

For this analysis, we run ~60,000 DEAs to test candidate SVs individually (Figure 12, A). We identified 177 candidate SVs predicted to alter gene expression directly, eSVs (Figure 12, A), including 76 INSs, 94 DELs, 4 INVs, and 3 DUPs. Most hits were intronic and upstream SVs, mainly affecting their protein-coding gene targets (Figure 12, B). Predicted eSVs were not enriched in any variant consequences. We checked the general patterns of logFC distribution for the DE hits (Figure 12, C). Interestingly, while protein-coding genes with coding, splicing and 5`UTR variants were downregulated, lncRNAs with intronic and splicing variants were on average upregulated.

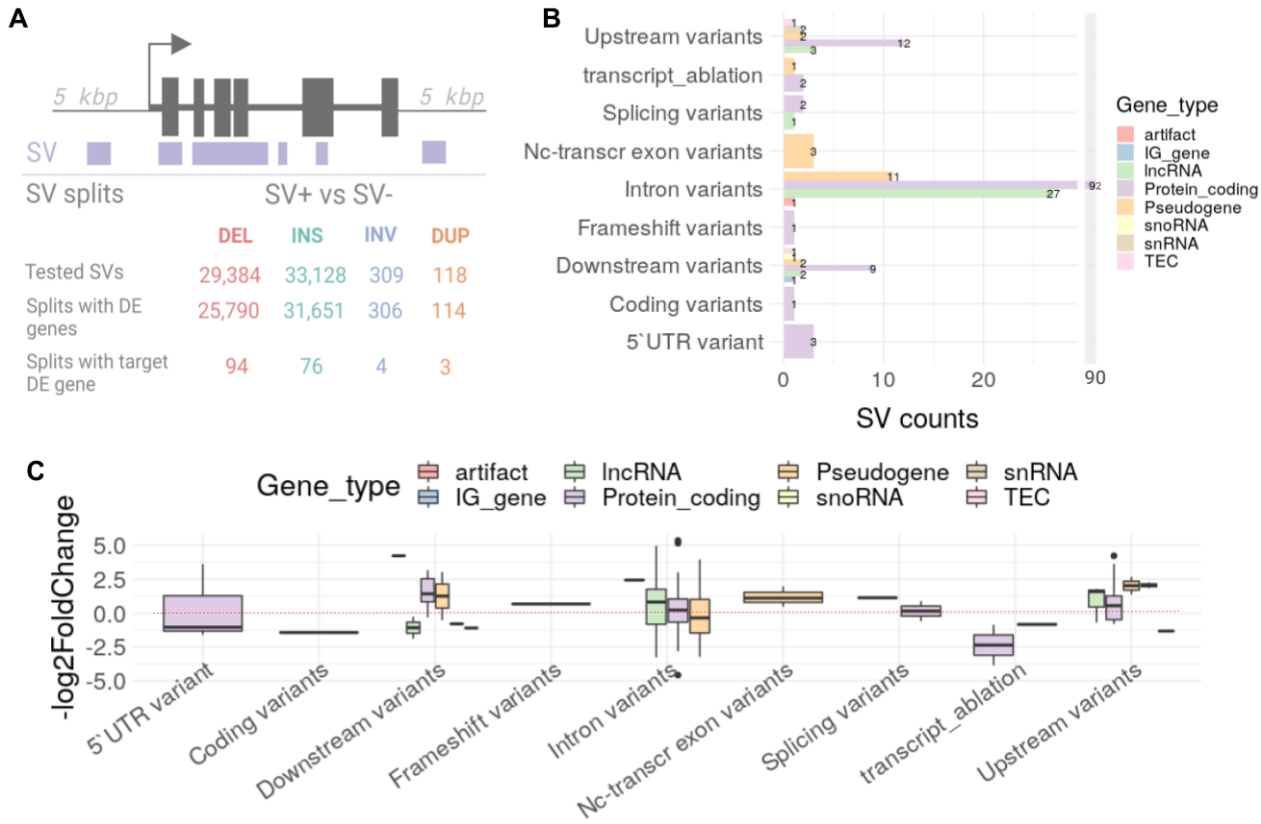


Figure 12. Design of DE analysis for SV functional annotation. **A.** Criteria for SVs to be included into the DE analysis according to their positional annotation and resulting numbers of SV types. **B.** Effect of SVs for which DEA revealed target genes to be differentially expressed (hit SVs). **C.** logFC distribution for the differentially expressed target genes based on the consequence of associated SV.

We focused on SVs whose genotype split coincided with the patient PD phenotype, comparing healthy controls with PD-affected individuals and carriers of PD mendelian mutations. We identified three hits, which can be prioritized as candidates to disturb molecular pathways implicated in PD and PD-related disorders.

The first hit - DEL (CNCR = 4, 50-75% percentile, MAF in PD cohort 0.04) located in the promoter/5'UTR region of gene *CBR1*, mitochondrial protein Carbonyl Reductase 1 (Figure 13, A). CBR1 was significantly downregulated in 1 1/1 and 4 0/1 DEL carriers with logFC = -1 (Figure 13, B). We leveraged a large FOUNDIN-PD database of assays available for the studied samples which include scRNA-seq, bulk ATAC-seq, and methylation profiling performed on day 65 of iPSC lines differentiation. We observed the same pattern of CBR1 expression for 1/1, 0/1, and 0/0 DEL carriers in the dopaminergic neuronal cluster from the scRNA seq dataset (Figure 13, C). We detected that the first CBR1 exon was less accessible in the DEL carriers according to the analysis of ATAC-seq dataset (0/1 and 1/1 genotypes) (Figure 13, D). In addition, the upstream region and CBR1 gene body showed a higher level of methylation in DEL carriers compared to DEL non-carriers (Figure 13, E). Other than the CBR1, another 3 protein-coding genes were differentially expressed in the CBR1-DEL GT: CLR1, GJA8, CLEC3A. Their protein products are involved in the cell adhesion activity and gap and synapse junction organization.

Two other hits, DEL and INS, are located in the intronic regions of PTPRN2 (Supplementary Figure S2) and PTPRG (Supplementary Figure S3) genes. PTPRN2 and PTPRG were significantly downregulated for the DEL (logFC = -1, MAF = 0.03) and INS (logFC = -0.85, MAF = 0.052) hit carriers, respectively. The results were not replicated for the dopaminergic neuronal clusters (Supplementary figures S2-S3). Analysis of epigenetics in the corresponding regions did not reveal any differences between SV hit carriers and non-carriers. We investigated the regions affected by these hits to account for potential known genetic features discovered in large cohorts or in samples with similar phenotypes. We discovered that PTPRN2-intronic-DEL intersects a differentially methylated region in AD patients according to the recent genome-wide histone 3 lysine 27 acetylation (H3K27ac) profiling [256]. Gene ontology enrichment analysis for the PTPRN2-DEL and PTPRG-INS revealed enrichment for synaptic signaling and mRNA alternative splicing biological processes, respectively (Supplementary Figure S4).

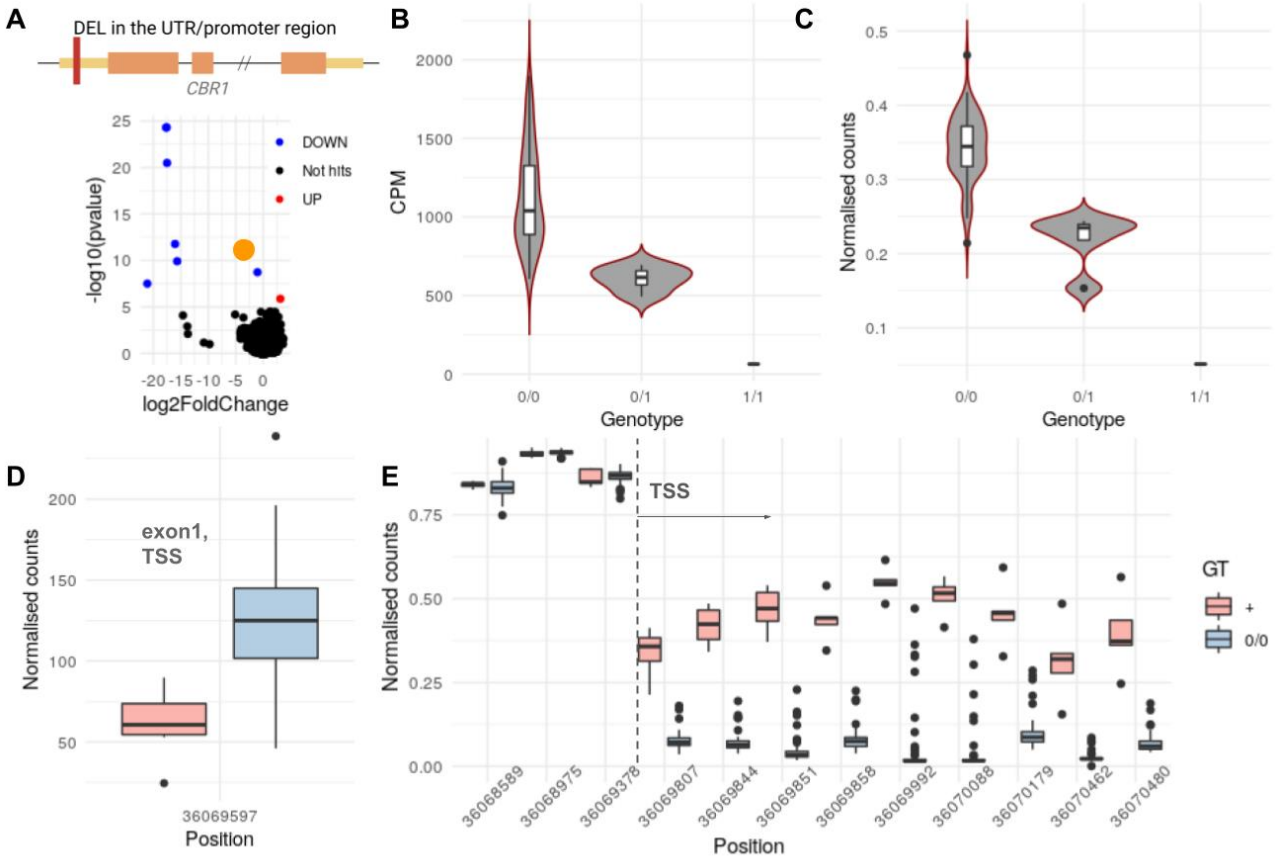


Figure 13. DEL (DEL-CBR1) in promoter/UTR region of *CBR1 and CBR1* expression in multi omics datasets by DEL-CBR1 GT groups. **A.** Schematic representation of DEL localization and volcano plot of DEA results for DEL-CBR1 GT split. **B.** *CBR1* expression in bRNA-seq (day 65). **C**. *CBR1* expression in DA clusters of scRN-seq (day 65). **D**. Chromatin availability detected by bATAC-seq in the region of *CBR1* TSS/exon1 (day 65). **E**. Methylation profile of *CBR1* gene body and upstream region (day 65).

6.2 Expression outlier analysis

Most studies in the functional variant annotation focus on common variants for which statistical power can be gained to account for the sample group variability and identify the direct association between the variant and its transcriptional consequences. However, a recent analysis revealed a substantial role of rare and ultra-rare regulatory variants in gene expression by identifying gene expression outliers (eOutliers) [257].

We ran eOutlier detection analysis using autoencoder to account for batch effects and unknown covariation [233] (Supplementary Figure S5). After filtering for lowly expressed genes, we ended up with 72,655 coding genes and lnRNAs from 92 samples which were input in the eOutlier analysis. As a result, we identified 124 aberrantly expressed genes from 47 samples. On average, we detected one aberrantly expressed gene per sample with three aberrant genes/sample being in the 90th percentile (Figure 14, A). During the next step, we explored SV presence in the vicinity of or within the detected genes matching the SV genotypes. We identified a total of 96 unique SVs (median MAF 0.3, median size 161.5 bps), including 27 rare SVs located in the gene body or within 5 kb flanking regions of 43 gene eOutliers. Most eOutlier associated SVs (eoSVs) were intronic variants, followed by upstream SVs. However, we discovered that 17% of intronic SVs were localized within transcribed pseudogenes (Figure 14, B). SVs affecting coding sequences and 5`UTR regions were represented solely by rare SVs interfering with protein-coding genes.

We checked the distribution of gene expression for the subset of genes with eo-SVs. The transcriptional consequences were specifically notable for the genes harboring rare SVs (Figure 14, C-D). Genes with rare DELs removing one or both copies of cis-eQTLs were significantly downregulated: the median logFC of gene expression was -0.98 (p-value 6.994e-08, two-sided Mann-Whitney U Test). In addition, eOutlier genes with rare pLoF SVs were significantly downregulated with median logFC -0.9 (p-value 1.404e-05, two-sided Mann-Whitney U Test).

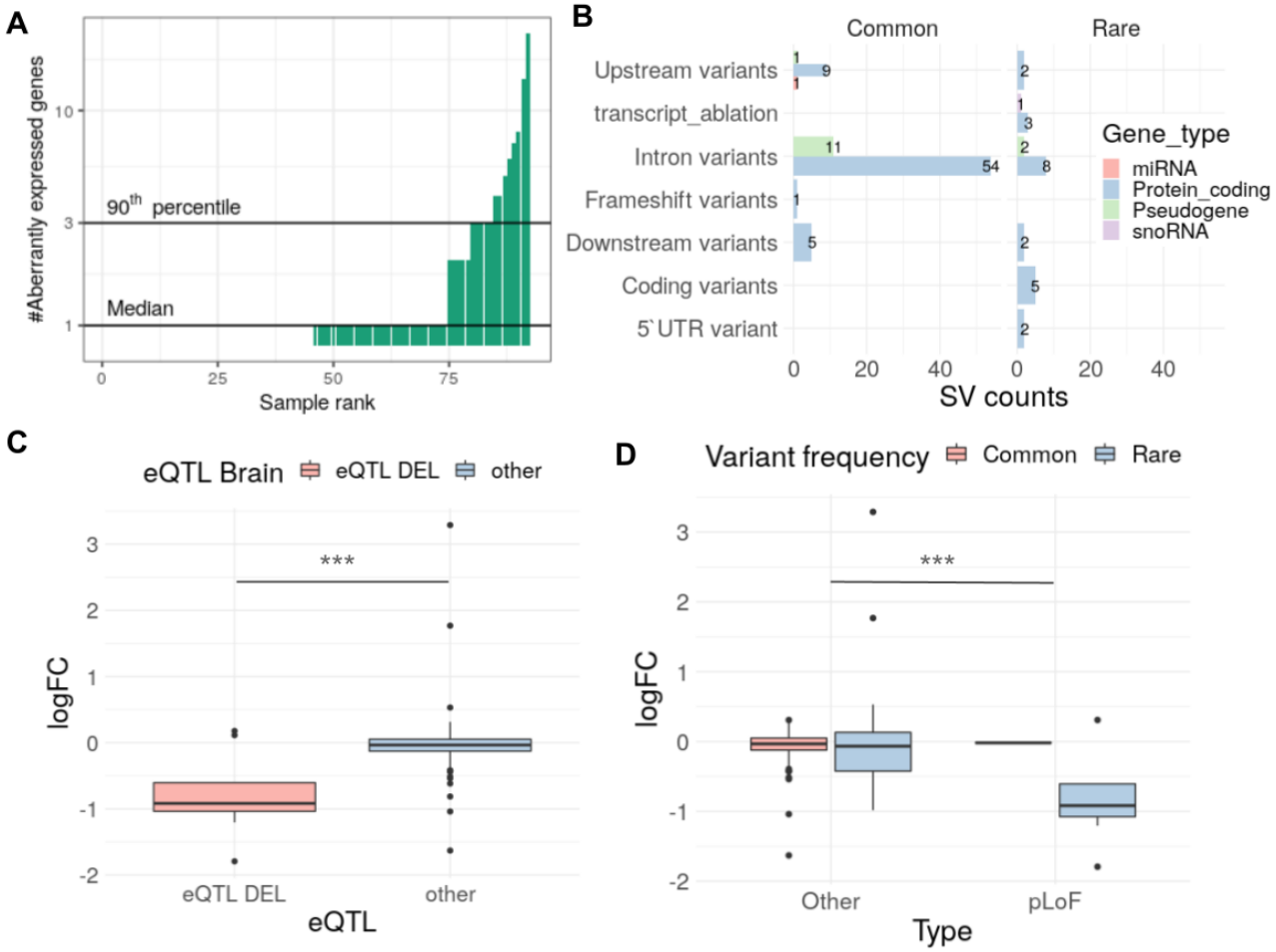


Figure 14. Overview of SVs (eoSV) associated with eOutliers. **A.** Number of aberrantly expressed genes per sample. **B.** Positional annotation and predicted effect of SVs localized in the vicinity of aberrantly expressed genes based on their predicted effect on the gene. **C.** Box plots demonstrating change in target gene average expression between eoSV carriers and non-carriers expressed in logFC for eoSVs affecting brain cis-eQTLs and **D.** for eoSVs pLoF - loss-of function SVs including coding, 5`UTR upstream and start/stop codon loss SVs; SP - SVs affecting splicing sites; Other - other SVs including intronic variants and up/downstream SVs.

We identified 18 rare eoSVs, including ten pLoF SVs which could trigger aberrant gene expression in PD affected samples or samples with high penetrance variants.

We replicated our finding of CBR1 gene downregulation in the promoter DEL carriers: CBR1 was detected to be an eOutlier in the 1/1 CBR1-DEL carrier (LRRK2 + unaffected individual).

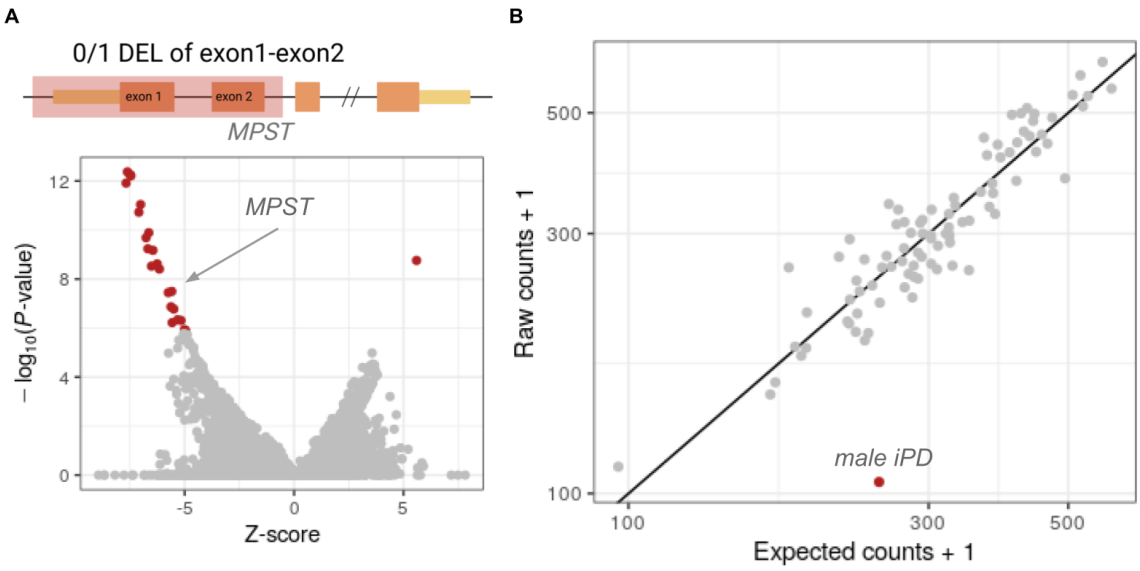


Figure 15. Heterozygous singleton DEL associated with MPST aberrant expression. A. Schematic representation of MPST region affected by the DEL and volcano plot showing differentially expressed genes for the eOutlier male iPD sample. B. MPST raw counts (adjusted for covariates) vs expected counts for MPST within FOUNDIN-PD cohort.

Another detected hit is a large singleton heterozygous DEL affecting 5'UTR and the first two exons of gene MPST (Figure 15, A). The DEL is specific for a sample from an iPD male individual. MPST was significantly downregulated in the given sample (z score = -5.57, p-value 2.448537e-02, Figure 15, B)). Apart from MPST, there were 22 aberrantly expressed genes detected for the given sample. Pathway enrichment analyses revealed dysregulation in iron-sulfur metabolism and export from mitochondria as well as protein metabolism and chromatin organization (Supplementary Figure S6).

Several eoSVs were identified to be associated with expression changes of genes previously associated with PD or other neurodegenerative disorders: ZNF543[258], FEM1A[259], UEVLD [260]. A complete table of eoSVs hits absent in the samples from healthy individuals can be found in Supplementary materials (Supplementary table 2).

6.3 Differential transcript usage analysis

Genetic variation changes transcript structure and generates transcript diversity playing an essential role in disease manifestation. We generated a LR PCR-cDNA RNA-seq dataset from 10 samples of the FOUNDIN-PD cohort. Reference-guided transcript annotation and quantification were performed with bambu (v 0.3.0) based on GENCODE (v.29) expanded with LNCipedia (v 5.2). The resulting transcript set comprised 45,539 annotated genes and 228,775 transcripts, including 46,657 (20.4%) novel transcripts. We compared FOUNDIN-PD annotation with GENCODE (v.26) reference expanded with GTEx LR-annotated novel transcripts [214]. The number of transcripts with a complete and exact intron chain match reached 23,813 (13% of the annotated transcript set). Most transcripts (~50%) have at least one matched intron junction between the reference and query annotation set.

We explored the presence of intron-retainment events. The group of transcripts with retained introns where all or several introns are matched comprised 6720 transcripts (~ 3%). In addition, we identified 3081 transcripts (56.7% novel transcripts) fully contained in the reference introns. A group of transcripts annotated on the opposite strand included 5703 features, with 65% being novel transcripts.

Next, we explored the transcript length distribution and level of expression across different comparison classes. Classes with transcripts with at least one intron junction match and classes with partially overlapped exons, overlapped introns on another strand, or no overlaps, were compared. We identified that transcripts belonging to the first group are significantly longer (median length 1029 and 841 bps for group 1 and 2 respectively, p-value < 0.05, two-sided Mann-Whitney U Test) and expressed on a significantly higher level (median normalized counts 6.88 and

2.9 for group 1 and 2 respectively, p-value < 0.05, two-sided Mann-Whitney U Test) than transcripts from the second group. We have used annotated and novel transcripts with at least one intron junction match on the same strand with the reference annotation for further downstream analysis.

SVs were screened for potential association with differentially used transcripts. For this purpose, we run a differential transcript usage (DTU) analysis to capture statistically significant SV-transcript expression association.

The candidate SVs were chosen according to the following criteria:

- SV presence (0/1, 1/1) and absence (0/0) at least in three samples for which we had long read RNAseq data
- SV candidate for a cis-regulation: intergenic SVs excluded.

For this analysis, we run ~30,000 DTU analysis runs to test candidate SVs individually. As a result, we obtained ~15,000 GT splits where DU transcripts were detected. Exploratory analysis of gene bodies and 5 kbp gene flanking regions revealed 46 SVs associated with 37 unique genes with 51 DU transcripts. A major part of DU transcripts was represented by protein-coding transcripts (78.9%). The remaining isoforms were transcripts with retained introns (13.1%), transcripts directed to nonsense-mediated decay (5.4%), and lncRNA transcripts (2.6%).

Similar to the DEA results exploration, we focused on SVs whose genotype split coincided with the patient PD phenotype, comparing healthy controls with PD-affected individuals and carriers of PD mendelian mutations. Out of 46 discovered hits, we found 1 DEL, which was absent in the healthy controls: heterozygous DEL (DEL-DUT, MAF = 0.14, absent in healthy controls) located in the upstream/promoter region of gene *DUT*, Deoxyuridine Triphosphatase (Figure 16, A). DUT is an enzyme involved in the metabolism of nucleotides and DNA preparation, catalyzing the hydrolysis reaction of dUTP to dUMP and pyrophosphate [261]. The two main isoforms encoded by DUT are the following: the MANE transcript encodes the mitochondrial isoform (ENST00000331200.7, DUT-M) and the second transcript encodes nucleus isoform (ENST00000455976.6, DUT-N) (Ensembl release 107). We identified that while the DUT-M expression level

remains similar across the DEL GT groups, DUT-N is significantly downregulated in the DEL carriers. DEL-DUT genes with DUTs were enriched in the regulation of RNA splicing, mitotic DNA integrity, damage checkpoint signaling, and dUMP biosynthetic process (Figure 16, B).

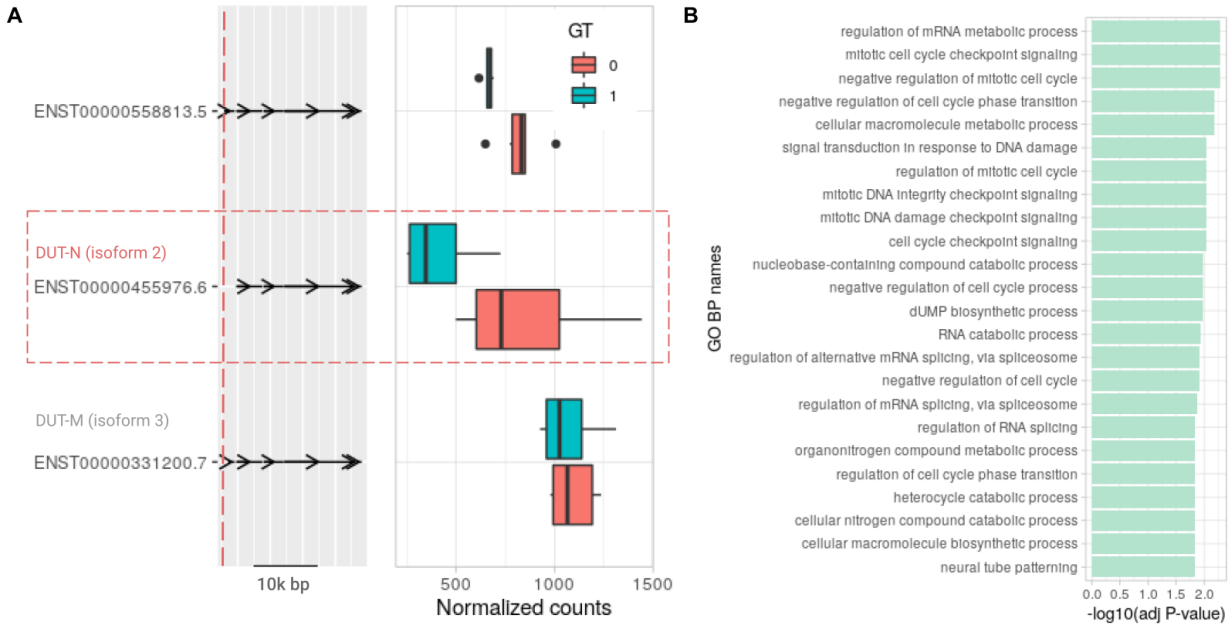


Figure 16. Differential usage of DUT transcripts. **A.** Exon-intron structure of the top expressed DUT transcripts, red rectangle indicates DU transcript (ENST00000455976.6), red dashed line indicates the position of associated DEL. **B.** Pathway analysis of genes with DTUs obtained from DEL_DUT- vs DEL_DUT+ comparison.

6.4 Discussion

SVs comprise a substantially larger fraction of gene expression-altering genetic variants compared to other forms of genetic and genomic variation [103,262]. Due to the complexity of SVs the assessment of their consequences depends heavily on available functional assays and cannot be inferred reliably from only positional genomic annotation [263]. The integration of genomic, transcriptomic, and epigenomic data from the same iPSC lines provided a unique opportunity for large-scale characterization of the SV functional impact on gene expression. We screened through the annotated SVs with predicted consequences obtained in the previous chapter in order to observe to which extent the variant effect directly reveals itself in the change of the target gene expression. Our results show that genes with coding DELs and DUPs are predictably downregulated. However, genes with coding INSs do not demonstrate a certain direction in their expression change which is in line with findings of the largest multi tissue SV-eQTL study of common and rare SVs [103]. Our findings suggest that genes with coding DELs and DUPs overlapping human lineage-specific regions and known SNP-eQTLs tend to have a more pronounced change in their expression. However, again these results were not replicated for coding INSs. As was already shown in the previous SR and LR-based SV-eQTL studies, only a minor set of INSs were associated with the gene expression demonstrating bidirectional behavior of the allele effect size [103]. INSs work through a different functional mechanism in comparison to other CNVs, which indicates the importance of additional information other than reference sequence context such as assessment of non-reference INS sequence pathogenicity via a GC content, repetitiveness and classes of present repeats, and presence of potential regulatory elements and TF factor binding sites which may interfere with the gene expression regulation.

Our study demonstrated that 34% of gene eOutliers could be explained by the presence of SVs within the gene body or in the vicinity. These results are lower than previous findings [103], which can be explained by the lower number of samples and higher variability in the gene expression due to the presence of heterogeneous PD genetic and idiopathic groups. We also showed the importance of the 3D

chromatin organization information, demonstrating that genes with CNVs localized within or on the boundaries of TADs recurrent in several cell lines are significantly downregulated. These results should be further refined with the recalculation and assessment of TAD boundaries for different SV genotype groups of the actual SV carriers and non-carriers to capture the dynamics of chromatin organization in response to the genomic variation and distinguish between intra-, extra, and boundary TAD CNVs.

Returning to the coding DELs and DUPs, we found that genes with exonic DEL and DUPs are notably downregulated compared to genes with splice-site affecting CNVs. This result implies that while exonic SVs are expected to impact the general gene expression, the effect of splicing SVs can be hidden on the gene level and revealed only on the level of individual transcript expression. LR transcriptome data provides important insights into how rare and common SVs alter transcript expression and modify the risk of disease development.

In this study, we also leveraged the availability of matched LR RNA sequencing data to explore the potential effect of SVs on transcript usage changes. We identified 228,775 transcripts, a middle between one sample-based and multi-sample multi-tissue meta-LR RNA studies [214,264,265]. FOUNDIN-PD transcript dataset includes novel transcripts with intron retention, which indicates the presence of pre-mRNA. However, the number of these transcripts is considerably lower in comparison to the findings of a recent LR multi-tissue RNA study [214]. Most transcripts annotated on the opposite strand in the current study were novel, implicating several points: discovery of novel genetic features such as lncRNA and antisense RNA genes or/and artifacts occurring during the computational process such as issues with the read reorientation and incorrect read strand assignment. Further benchmarking and novel transcripts validation analyses have to be performed to test suggested hypotheses.

We identified several SV hits prioritized as risk factors to perturb molecular pathways leading to the PD or PD-related clinical phenotype in the current study. While the major expression changes were observed among the genes altered by coding SVs, almost all prioritized SV occurred to be intronic and regulatory common

and rare CNVs supporting on one hand side the low effect risk variants and "common disease common variant" theory underlying the PD genetics basis, and on the other hand - revealing the presence of rare and singleton potentially moderately pathogenic CNVs, the rareness of which can be explained by their novelty.

A promoter/5`UTR CBR1-DEL absent in healthy individuals was associated with the CBR1 downregulation on transcriptomic and epigenomic levels. CBR1, NADPH-dependent carbonyl reductase, reduces reactive quinones and lipid aldehydes, playing a protective role against ROS-induced cellular damage and neurodegeneration [266,267]. In addition, it was shown that the CBR1 promoter possesses at least one antioxidant response element (ARE) and that CBR1 transcription is regulated by NRF2, a TF that regulates the expression of antioxidant proteins, and overexpression of this enzyme is crucial for the survival of cancer cells [268]. These findings indicate a major role of CBR1 in response to oxidative stress; those dysregulations can be critical for neuronal cell survival. Interestingly, in our cohort, the largest downregulation was observed in the homozygous CBR1-DEL carrier, a LRRK2 positive individual unaffected with PD by the time the patient metadata was collected when the patient was younger than 65 years. On one hand, it might be an indication of the cumulative effect of the CBR1-DEL alleles. On the other hand, we see that the variant penetrance is not high. However, it is expected to contribute to the organism's vulnerability to neurodegeneration, thus increasing the overall risk of developing the pathogenic process.

An intronic (ACA)-repeat INS was associated with the PTPRN2 downregulation on the transcriptomic level. Even though we did not observe any methylation changes in the INS carriers, we discovered that the PTPRN2-INS is located within the island of increased acetylation in the AD patients, which points out its potential role in the chromatin accessibility in the given region and subsequent regulation of PTPRN2 expression. PTPRN2, Protein Tyrosine Phosphatase Receptor Type N2, is an important protein in the presynaptic density and the cellular membrane of the pancreatic islets, which is involved in the vesicle-mediated secretory processes controlling the release of catecholamine neurotransmitters and insulin [269]. It is worth mentioning that recent metastudy demonstrated and supported the increased risk of PD development in patients diagnosed with diabetes mellitus [270]. PTPRN2

dysregulation was associated with diabetes type I and II, and with childhood obesity [271]. Increased methylation within the PTPRN2 gene body was previously associated with faster motor progression in PD affected individuals [272]. Collectively, these findings highlight the important link between neurodegeneration and insulin signaling in the brain expanding potential treatment targets for PD which can be approached with antidiabetic medications.

An intronic DEL was associated with the PTPRG downregulation on the transcriptome level. Receptor PTPRG is a Protein Tyrosine Phosphatase Receptor Type G, which has a high probability of HI (p(HI) = 0.94). It is located in one of the loci associated with AD in the family based GWAS [273]. PTPRG is involved in the mitochondrial autophagy process regulation, which was demonstrated in the cohort of AD patients [274]. PTPRG was also determined to be a causal gene in other neurological and neuropsychiatric disorders [275][274].

A heterozygous DEL removing promoter and first exons of MPST was associated with the MPST low expression in one sample from one iPD individual. The p(HI) is 0.48. However, in this case, we observe a clear haploinsufficiency mode of SV pathogenicity. This example indicates that SNP and InDel-based HI probability metrics have to be recalculated and updated for the larger sizes of genetic variability. MPST encodes for mercaptopyruvate sulfurtransferase, an enzyme that catalyzes the transfer of a sulfur ion from 3-mercaptopyruvate to thiol compounds. MPST was previously proven to play an essential role in brain aging and neurodegeneration through a process of persulfidation[276]. Depletion of MPST leads to reduced metabolic rate and impaired mitochondrial protein transport in mice [277]. Thus, the discovered DEL impacting one of the MPST copies can be considered as one of the risk factors contributing to the development of PD in the given idiopathic patient.

A common DEL located in the upstream/promoter region of gene DUT, a deoxyuridine triphosphatase, was linked to the differential transcript usage of the DUT isoform which is specifically active in the nucleus (DUT-N) in comparison to the mitochondrial isoform (DUT-M) which was not dysregulated in our dataset. Interestingly, that gene level expression for the DUT-DEL GT split did not demonstrate any differences highlighting the importance of transcript level

expression assessment. DUT hydrolyses dUTP to dUMP and pyrophosphate, removing dUTP from the nucleotide pool during the processes of DNA reparation and replication [261]. Notably, neurons as non-dividing cell types are particularly dependable on the unimpacted genome integrity [278]. Dysregulation of DUT-N suggests that while the mitochondrial DNA reparation process is intact, nucleus DNA reparation is dealing with insufficient concentration of the DUT which leads to elevated incorporation of U into DNA molecules and triggering of active DNA reparation, occurrences of double-strand breaks, and subsequent apoptosis.

Prioritized hits are subjects for further validation to prove that the SVs are indeed present in the donor patients. The upcoming large study of LR DNA sequencing of PPMI patients will be an ideal data source for the hit SV validation. Additional analysis on the larger PD patient and control cohort has to be conducted to validate the observed variant-expression associations and to explore the allele effect size of the prioritized hits.

7 Conclusions

In conclusion, we generated and systematically annotated the largest LR-based SV dataset for the cohort enriched for familial and idiopathic PD. We showed that the usage of LR DNA sequencing dataset of sufficient coverage and read length allows one to obtain an accurate SV callset genome-wide including the low complexity regions. We demonstrated that a combined genome-sequencing and omics assay performed for the same samples is essential to interpret further the downstream consequences of SVs, to improve predicted variant annotation, and assess the actual functional effect of the observed variation. Our results indicate the importance of the transcriptome analysis not only on the level of genes but rather on a level of transcript structure and expression. The diploid nature of the human genome raises a complex interplay among the alleles and diminishes the effect size of potentially deleterious variation which acts in the recessive mode. Obtained results have to be refined with the SV-eQTL analysis as well as gene LoF-intolerance and HI SV-based probability calculations once the statistical power for these kinds of methods is sufficient.

Many genes and GWAS loci were associated with the development of PD and modulation of the disease onset age and progression. The current study suggests that the ultimate causal PD risk SVs might be less common than we expected, thus requiring larger cohorts to be analyzed in order to capture this variation and resolve the gap of PD missing heritability. Here, we also highlight a group of molecular pathways, which should be carefully examined in the context of PD in a row with mitochondrial and lysosomal malfunctions. The group includes vesicular release signaling processes, oxidative stress response, and neuronal DNA reparation. Novel and existing population-wide sequencing studies have to be actively used to quantify and validate the pathogenicity of discovered variants and nominated causal genes.

The genetics of PD is complex and heterogeneous. PD cases should be carefully stratified based on their ethnicity, age of onset, symptoms, and disease progression. Based on the simple calculations, researchers should aim for a minimal cohort size of 60 to 300 individuals to be able to capture a significant variant-trait association

assuming the target risk SV MAF is around 0.05 to 0.01. It is essential to leverage state-of-art technologies, such as 3GS, to ensure the maximal yield of genetics studies covering a wide spectrum of DNA variation. Given the high heterogeneity of molecular process triggering PD, one should focus on the tissue and cell-specific level of variant functional annotation for a more homogeneous sample group and a clearer result interpretation, which is actively pioneered and demonstrated by the FOUNDIN-PD project. We anticipate that future expansion of SV datasets accompanied with the sample-matching omics data will provide a better opportunity to predict SV phenotypic impact and assess their role in the development of PD and other complex disorders.

The dataset generated in the current study is publicly available to facilitate the further discovery of SV-PD risk association to expand and support findings in the genetics of neurodegenerative disorders.

8 References

1. Balestrino R, Schapira AHV. Parkinson disease. Eur J Neurol. 2020;27: 27–42. doi:10.1111/ene.14108

2. Enders D, Balzer-Geldsetzer M, Riedel O, Dodel R, Wittchen H-U, Sensken S-C, et al. Prevalence, Duration and Severity of Parkinson's Disease in Germany: A Combined Meta-Analysis from Literature Data and Outpatient Samples. Eur Neurol. 2017;78: 128–136. doi:10.1159/000477165

3. Dorsey ER, Bloem BR. The Parkinson Pandemic-A Call to Action. JAMA Neurol. 2018;75: 9–10. doi:10.1001/jamaneurol.2017.3299

4. Kalia LV, Lang AE. Parkinson's disease. Lancet. 2015;386: 896–912. doi:10.1016/s0140-6736(14)61393-3

5. Jellinger KA, Logroscino G, Rizzo G, Copetti M, Arcuti S, Martino D, et al. Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis. Neurology. 2016;87: 237–238. doi:10.1212/WNL.0000000000002876

6. Tolosa E, Garrido A, Scholz SW, Poewe W. Challenges in the diagnosis of Parkinson's disease. Lancet Neurol. 2021;20: 385–397. doi:10.1016/S1474-4422(21)00030-2

7. Noyce AJ, Bestwick JP, Silveira-Moriyama L, Hawkes CH, Giovannoni G, Lees AJ, et al. Meta-analysis of early nonmotor features and risk factors for Parkinson disease. Ann Neurol. 2012;72: 893–901. doi:10.1002/ana.23687

8. Blauwendraat C, Nalls MA, Singleton AB. The genetic architecture of Parkinson's disease. Lancet Neurol. 2020;19: 170–178. doi:10.1016/S1474-4422(19)30287-X

9. Hardy J, Lewis P, Revesz T, Lees A, Paisan-Ruiz C. The genetics of Parkinson's syndromes: a critical review. Curr Opin Genet Dev. 2009;19: 254–265. doi:10.1016/j.gde.2009.03.008

10. Karimi-Moghadam A, Charsouei S, Bell B, Jabalameli MR. Parkinson Disease from Mendelian Forms to Genetic Susceptibility: New Molecular Insights into the Neurodegeneration Process. Cell Mol Neurobiol. 2018;38: 1153–1178. doi:10.1007/s10571-018-0587-4

11. Polymeropoulos MH, Lavedan C, Leroy E, Ide SE, Dehejia A, Dutra A, et al. Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. Science. 1997;276: 2045–2047. doi:10.1126/science.276.5321.2045

12. Wong YC, Krainc D. α-synuclein toxicity in neurodegeneration: mechanism and therapeutic strategies. Nat Med. 2017;23: 1–13. doi:10.1038/nm.4269

13. Winner B, Jappelli R, Maji SK, Desplats PA, Boyer L, Aigner S, et al. In vivo demonstration that α-synuclein oligomers are toxic. Proceedings of the National Academy of Sciences. 2011. pp. 4194–4199. doi:10.1073/pnas.1100976108

14. Le S, Fu X, Pang M, Zhou Y, Yin G, Zhang J, et al. The Antioxidative Role of Chaperone-Mediated Autophagy as a Downstream Regulator of Oxidative Stress in Human Diseases. Technol Cancer Res Treat. 2022;21: 15330338221114178. doi:10.1177/15330338221114178

15. Xilouri M, Brekk OR, Stefanis L. Autophagy and Alpha-Synuclein: Relevance to Parkinson's Disease and Related Synucleopathies. Mov Disord. 2016;31: 178–192. doi:10.1002/mds.26477

16. Petrucelli L, O'Farrell C, Lockhart PJ, Baptista M, Kehoe K, Vink L, et al. Parkin protects against the toxicity associated with mutant alpha-synuclein: proteasome dysfunction selectively affects catecholaminergic neurons. Neuron. 2002;36: 1007–1019. doi:10.1016/s0896-6273(02)01125-x

17. Yung C, Sha D, Li L, Chin L-S. Parkin Protects Against Misfolded SOD1 Toxicity by Promoting Its Aggresome Formation and Autophagic Clearance. Mol Neurobiol. 2016;53: 6270–6287. doi:10.1007/s12035-015-9537-z

18. Stefanis L. α-Synuclein in Parkinson's disease. Cold Spring Harb Perspect Med. 2012;2: a009399. doi:10.1101/cshperspect.a009399

19. Paisán-Ruíz C, Jain S, Evans EW, Gilks WP, Simón J, van der Brug M, et al. Cloning of the gene containing mutations that cause PARK8-linked Parkinson's disease. Neuron. 2004;44: 595–600. doi:10.1016/j.neuron.2004.10.023

20. Zimprich A, Biskup S, Leitner P, Lichtner P, Farrer M, Lincoln S, et al. Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. Neuron. 2004;44: 601–607. doi:10.1016/j.neuron.2004.11.005

21. Kestenbaum M, Alcalay RN. Clinical Features of LRRK2 Carriers with Parkinson's Disease. Adv Neurobiol. 2017;14: 31–48. doi:10.1007/978-3-319-49969-7_2

22. Bryant N, Malpeli N, Ziaee J, Blauwendraat C, Liu Z, AMP PD Consortium, et al. Identification of LRRK2 missense variants in the accelerating medicines partnership Parkinson's disease cohort. Hum Mol Genet. 2021;30: 454–466. doi:10.1093/hmg/ddab058

23. Berwick DC, Heaton GR, Azeggagh S, Harvey K. LRRK2 Biology from structure to dysfunction: research progresses, but the themes remain the same. Mol Neurodegener. 2019;14: 49. doi:10.1186/s13024-019-0344-2

24. Abdel-Magid AF. LRRK2 Kinase Inhibitors as Possible Therapy for Parkinson's Disease and Other Neurodegenerative Disorders. ACS Med Chem Lett. 2019;10: 846–847. doi:10.1021/acsmedchemlett.9b00216

25. Zimprich A, Benet-Pagès A, Struhal W, Graf E, Eck SH, Offman MN, 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. doi:10.1016/j.ajhg.2011.06.008

26. Ando M, Funayama M, Li Y, Kashihara K, Murakami Y, Ishizu N, et al. VPS35 mutation in Japanese patients with typical Parkinson's disease. Mov Disord. 2012;27: 1413–1417. doi:10.1002/mds.25145

27. Kumar KR, Weissbach A, Heldmann M, Kasten M, Tunc S, Sue CM, et al. Frequency of the D620N mutation in VPS35 in Parkinson disease. Arch Neurol. 2012;69: 1360–1364. doi:10.1001/archneurol.2011.3367

28. Rahman AA, Morrison BE. Contributions of VPS35 Mutations to Parkinson's Disease. Neuroscience. 2019;401: 1–10. doi:10.1016/j.neuroscience.2019.01.006

29. Williams ET, Chen X, Otero PA, Moore DJ. Understanding the contributions of VPS35 and the retromer in neurodegenerative disease. Neurobiol Dis. 2022;170: 105768. doi:10.1016/j.nbd.2022.105768

30. Stockman JA. Multicenter Analysis of Glucocerebrosidase Mutations in Parkinson's Disease. Yearbook of Pediatrics. 2011. pp. 419–420. doi:10.1016/s0084-3954(10)79717-7

31. Blauwendraat C, Bras JM, Nalls MA, Lewis PA, Hernandez DG, Singleton AB, et al. Coding variation in GBA explains the majority of the SYT11-GBA Parkinson's disease GWAS locus. Mov Disord. 2018;33: 1821–1823. doi:10.1002/mds.103

32. Rivas MA, Avila BE, Koskela J, Huang H, Stevens C, Pirinen M, et al. Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population. PLoS Genet. 2018;14: e1007329. doi:10.1371/journal.pgen.1007329

33. Blauwendraat C, Reed X, Krohn L, Heilbron K, Bandres-Ciga S, Tan M, et al. Genetic modifiers of risk and age at onset in GBA associated Parkinson's disease and Lewy body dementia. Brain. 2020;143: 234–248. doi:10.1093/brain/awz350

34. Bo R-X, Li Y-Y, Zhou T-T, Chen N-H, Yuan Y-H. The neuroinflammatory role of glucocerebrosidase in Parkinson's disease. Neuropharmacology. 2022;207: 108964. doi:10.1016/j.neuropharm.2022.108964

35. Pickrell AM, Youle RJ. The roles of PINK1, parkin, and mitochondrial fidelity in Parkinson's disease. Neuron. 2015;85: 257–273. doi:10.1016/j.neuron.2014.12.007

36. Buhlman LM. Mitochondrial Mechanisms of Degeneration and Repair in Parkinson's Disease. Springer; 2016. Available: https://play.google.com/store/books/details?id=-uwQDQAAQBAJ

37. Imberechts D, Kinnart I, Wauters F, Terbeek J, Manders L, Wierda K, et al. DJ-1 is an essential downstream mediator in PINK1/parkin-dependent mitophagy. Brain. 2022. doi:10.1093/brain/awac313

38. Lohmann E, Coquel A-S, Honoré A, Gurvit H, Hanagasi H, Emre M, et al. A new F-box protein 7 gene mutation causing typical Parkinson's disease. Mov Disord. 2015;30: 1130–1133. doi:10.1002/mds.26266

39. Lesage S, Lunati A, Houot M, Romdhan SB, Clot F, Tesson C, et al. Characterization of Recessive Parkinson Disease in a Large Multicenter Study. Ann Neurol. 2020;88: 843–850. doi:10.1002/ana.25787

40. Kasten M, Hartmann C, Hampf J, Schaake S, Westenberger A, Vollstedt E-J, et al. Genotype-Phenotype Relations for the Parkinson's Disease Genes Parkin, PINK1, DJ1: MDSGene Systematic Review. Mov Disord. 2018;33: 730–741. doi:10.1002/mds.27352

41. Bradshaw AV, Campbell P, Schapira AHV, Morris HR, Taanman J-W. The PINK1-Parkin mitophagy signalling pathway is not functional in peripheral blood mononuclear cells. PLoS One. 2021;16: e0259903. doi:10.1371/journal.pone.0259903

42. Grünewald A, Kumar KR, Sue CM. New insights into the complex role of mitochondria in Parkinson's disease. Prog Neurobiol. 2019;177: 73–93. doi:10.1016/j.pneurobio.2018.09.003

43. Bento CF, Ashkenazi A, Jimenez-Sanchez M, Rubinsztein DC. The Parkinson's disease-associated genes ATP13A2 and SYT11 regulate autophagy via a common pathway. Nat Commun. 2016;7: 11803. doi:10.1038/ncomms11803

44. Funayama M, Ohe K, Amo T, Furuya N, Yamaguchi J, Saiki S, et al. CHCHD2 mutations in autosomal dominant late-onset Parkinson's disease: a genome-wide linkage and sequencing study. Lancet Neurol. 2015;14: 274–282. doi:10.1016/S1474-4422(14)70266-2

45. Edvardson S, Cinnamon Y, Ta-Shma A, Shaag A, Yim Y-I, Zenvirt S, et al. A deleterious mutation in DNAJC6 encoding the neuronal-specific clathrin-uncoating co-chaperone auxilin, is associated with juvenile parkinsonism. PLoS One. 2012;7: e36458. doi:10.1371/journal.pone.0036458

46. Paisan-Ruiz C, Bhatia KP, Li A, Hernandez D, Davis M, Wood NW, et al. Characterization of PLA2G6 as a locus for dystonia-parkinsonism. Ann Neurol. 2009;65: 19–23. doi:10.1002/ana.21415

47. Lesage S, Drouet V, Majounie E, Deramecourt V, Jacoupy M, Nicolas A, et al. Loss of VPS13C Function in Autosomal-Recessive Parkinsonism Causes Mitochondrial Dysfunction and Increases PINK1/Parkin-Dependent Mitophagy. Am J Hum Genet. 2016;98: 500–513. doi:10.1016/j.ajhg.2016.01.014

48. Lesage S, Mangone G, Tesson C, Bertrand H, Benmahdjoub M, Kesraoui S, et al. Clinical Variability of -Associated Early-Onset Parkinsonism. Front Neurol. 2021;12: 648457. doi:10.3389/fneur.2021.648457

49. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet. 2003;33: 177–182. doi:10.1038/ng1071

50. Nalls MA, Blauwendraat C, Vallerga CL, Heilbron K, Bandres-Ciga S, Chang D, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet Neurol. 2019;18: 1091–1102. doi:10.1016/S1474-4422(19)30320-5

51. Foo JN, Chew EGY, Chung SJ, Peng R, Blauwendraat C, Nalls MA, et al. Identification of Risk Loci for Parkinson Disease in Asians and Comparison of Risk Between Asians and Europeans: A Genome-Wide Association Study. JAMA Neurol. 2020;77: 746–754. doi:10.1001/jamaneurol.2020.0428

52. Blauwendraat C, Heilbron K, Vallerga CL, Bandres-Ciga S, von Coelln R, Pihlstrøm L, et al. Parkinson's disease age at onset genome-wide association study: Defining heritability, genetic loci, and α-synuclein mechanisms. Mov Disord. 2019;34: 866–875. doi:10.1002/mds.27659

53. Grenn FP, Kim JJ, Makarious MB, Iwaki H, Illarionova A, Brolin K, et al. The Parkinson's Disease Genome-Wide Association Study Locus Browser. Mov Disord. 2020;35: 2056–2067. doi:10.1002/mds.28197

54. Bandrés-Ciga S, Ruz C, Barrero FJ, Escamilla-Sevilla F, Pelegrina J, Vives F, et al. Structural genomic variations and Parkinson's disease. Minerva Med. 2017;108: 438–447. doi:10.23736/S0026-4806.17.05246-6

55. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526: 75–81. doi:10.1038/nature15394

56. Peng Z, Zhou W, Fu W, Du R, Jin L, Zhang F. Correlation between frequency of non-allelic homologous recombination and homology properties: evidence from homology-mediated CNV mutations in the human genome. Hum Mol Genet. 2015;24: 1225–1233. doi:10.1093/hmg/ddu533

57. Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. Pathogenetics. 2008;1: 4. doi:10.1186/1755-8417-1-4

58. Stankiewicz P, Lupski JR. Genome architecture, rearrangements and genomic disorders. Trends Genet. 2002;18: 74–82. doi:10.1016/s0168-9525(02)02592-1

59. Robberecht C, Voet T, Zamani Esteki M, Nowakowska BA, Vermeesch JR. Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. Genome Res. 2013;23: 411–418. doi:10.1101/gr.145631.112

60. Li Y-C, Chien S-C, Setlur SR, Lin W-D, Tsai F-J, Lin C-C. Prenatal detection and characterization of a psu idic(8)(p23.3) which likely derived from nonallelic homologous recombination between two MYOM2-repeats. Journal of the Formosan Medical Association. 2015. pp. 81–87. doi:10.1016/j.jfma.2011.05.015

61. Fujimoto A, Wong JH, Yoshii Y, Akiyama S, Tanaka A, Yagi H, et al. Whole-genome sequencing with long reads reveals complex structure and origin of structural variation in human genetic variations and somatic mutations in cancer. Genome Med. 2021;13: 65. doi:10.1186/s13073-021-00883-1

62. Gu S, Yuan B, Campbell IM, Beck CR, Carvalho CMB, Nagamani SCS, et al. Alu-mediated diverse and complex pathogenic copy-number variants within human chromosome 17 at p13.3. Hum Mol Genet. 2015;24: 4061–4077. doi:10.1093/hmg/ddv146

63. Gabriel A. Retrotransposons And Human Disease: L1 Retrotransposons As A Source Of Genetic Diversity. World Scientific; 2022. Available: https://books.google.com/books/about/Retrotransposons_And_Human_Disease_L1_Re.html?hl=&id=vTOJEAAAQBAJ

64. Currall BB, Chiang C, Talkowski ME, Morton CC. Mechanisms for Structural Variation in the Human Genome. Curr Genet Med Rep. 2013;1: 81–90. doi:10.1007/s40142-013-0012-8

65. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, et al. Recent segmental duplications in the human genome. Science. 2002;297: 1003–1007. doi:10.1126/science.1072047

66. Carvalho CMB, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016;17: 224–238. doi:10.1038/nrg.2015.25

67. Fudenberg G, Pollard KS. Chromatin features constrain structural variation across evolutionary timescales. Proc Natl Acad Sci U S A. 2019;116: 2175–2180. doi:10.1073/pnas.1808631116

68. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, et al. Characterizing the Major Structural Variant Alleles of the Human Genome. Cell. 2019;176: 663–675.e19. doi:10.1016/j.cell.2018.12.019

69. Almarri MA, Bergström A, Prado-Martinez J, Yang F, Fu B, Dunham AS, et al. Population Structure, Stratification, and Introgression of Human Structural Variation. Cell. 2020;182: 189–199.e15. doi:10.1016/j.cell.2020.05.024

70. Quan C, Li Y, Liu X, Wang Y, Ping J, Lu Y, et al. Characterization of structural variation in Tibetans reveals new evidence of high-altitude adaptation and introgression. Genome Biol. 2021;22: 159. doi:10.1186/s13059-021-02382-3

71. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev Genet. 2009;10: 551–564. doi:10.1038/nrg2593

72. Jakubosky D, D'Antonio M, Bonder MJ, Smail C, Donovan MKR, Young Greenwald WW, et al. Properties of structural variants and short tandem repeats associated with gene expression and complex traits. Nat Commun. 2020;11: 2927. doi:10.1038/s41467-020-16482-4

73. Spielmann M, Lupiáñez DG, Mundlos S. Structural variation in the 3D genome. Nat Rev Genet. 2018;19: 453–467. doi:10.1038/s41576-018-0007-0

74. D'haene E, Vergult S. Interpreting the impact of noncoding structural variation in neurodevelopmental disorders. Genet Med. 2021;23: 34–46. doi:10.1038/s41436-020-00974-1

75. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015;161: 1012–1025. doi:10.1016/j.cell.2015.04.004

76. Kaminsky EB, Kaul V, Paschall J, Church DM, Bunke B, Kunig D, et al. An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. Genet Med. 2011;13: 777–784. doi:10.1097/GIM.0b013e31822c79f9

77. Cruz-Martínez A, Bort S, Arpa J, Palau F. Hereditary neuropathy with liability to pressure palsies (HNPP) revealed after weight loss. Eur Neurol. 1997;37: 257–260. doi:10.1159/000117463

78. Rosen SA, Wang H, Cornblath DR, Uematsu S, Hurko O. Compression syndromes due to hypertrophic nerve roots in hereditary motor sensory neuropathy type I. Neurology. 1989;39: 1173–1177. doi:10.1212/wnl.39.9.1173

79. Liehr T, Rautenstrauss B, Grehl H, Bathke KD, Ekici A, Rauch A, et al. Mosaicism for the Charcot-Marie-Tooth disease type 1A duplication suggests somatic reversion. Hum Genet. 1996;98: 22–28. doi:10.1007/s004390050154

80. Singleton AB, Farrer M, Johnson J, Singleton A, Hague S, Kachergus J, et al. alpha-Synuclein locus triplication causes Parkinson's disease. Science. 2003;302: 841. doi:10.1126/science.1090278

81. Book A, Guella I, Candido T, Brice A, Hattori N, Jeon B, et al. A Meta-Analysis of α-Synuclein Multiplication in Familial Parkinsonism. Front Neurol. 2018;9: 1021. doi:10.3389/fneur.2018.01021

82. Lesage S, Magali P, Lohmann E, Lacomblez L, Teive H, Janin S, et al. Deletion of the parkin and PACRG gene promoter in early-onset parkinsonism. Hum Mutat. 2007;28: 27–32. doi:10.1002/humu.20436

83. Chawner SJRA, Doherty JL, Anney RJL, Antshel KM, Bearden CE, Bernier R, et al. A Genetics-First Approach to Dissecting the Heterogeneity of Autism: Phenotypic Comparison of Autism Risk Copy Number Variants. Am J Psychiatry. 2021;178: 77–86. doi:10.1176/appi.ajp.2020.20010015

84. Marshall CR, Howrigan DP, Merico D, Thiruvahindrapuram B, Wu W, Greer DS, et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nat Genet. 2017;49: 27–35. doi:10.1038/ng.3725

85. Modenato C, Kumar K, Moreau C, Martin-Brevet S, Huguet G, Schramm C, et al. Effects of eight neuropsychiatric copy number variants on human brain structure. Transl Psychiatry. 2021;11: 399. doi:10.1038/s41398-021-01490-9

86. Willsey AJ, Morris MT, Wang S, Willsey HR, Sun N, Teerikorpi N, et al. The Psychiatric Cell Map Initiative: A Convergent Systems Biological Approach to Illuminating Key Molecular Pathways in Neuropsychiatric Disorders. Cell. 2018;174: 505–520. doi:10.1016/j.cell.2018.06.016

87. Park E, Pan Z, Zhang Z, Lin L, Xing Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. Am J Hum Genet. 2018;102: 11–26. doi:10.1016/j.ajhg.2017.11.002

88. Asselta R, Duga S, Buratti E, Velasco EA. RNA Splicing and Backsplicing: Disease and Therapy. Frontiers Media SA; 2020. Available: https://books.google.com/books/about/RNA_Splicing_and_Backsplicing_Disease_an.html?hl=&id=3sQQEAAAQBAJ

89. Consortium TG, The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020. pp. 1318–1330. doi:10.1126/science.aaz1776

90. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med. 2017;9. doi:10.1126/scitranslmed.aal5209

91. Koks S, Pfaff AL, Bubb VJ, Quinn JP. Transcript Variants of Genes Involved in Neurodegeneration Are Differentially Regulated by the APOE and MAPT Haplotypes. Genes . 2021;12. doi:10.3390/genes12030423

92. Sanchis-Juan A, Stephens J, French CE, Gleadall N, Mégy K, Penkett C, et al. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. Genome Med. 2018;10: 95. doi:10.1186/s13073-018-0606-6

93. Carvalho CMB, Ramocki MB, Pehlivan D, Franco LM, Gonzaga-Jauregui C, Fang P, et al. Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. Nat Genet. 2011;43: 1074–1081. doi:10.1038/ng.944

94. Beck CR, Carvalho CMB, Banser L, Gambin T, Stubbolo D, Yuan B, et al. Complex genomic rearrangements at the PLP1 locus include triplication and quadruplication. PLoS Genet. 2015;11: e1005050. doi:10.1371/journal.pgen.1005050

95. Zelenova MA, Yurov YB, Vorsanova SG, Iourov IY. Laundering CNV data for candidate process prioritization in brain disorders. Mol Cytogenet. 2019;12: 54. doi:10.1186/s13039-019-0468-7

96. Depienne C, Mandel J-L. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? Am J Hum Genet. 2021;108: 764–785. doi:10.1016/j.ajhg.2021.03.011

97. Matsuura T, Yamagata T, Burgess DL, Rasmussen A, Grewal RP, Watase K, et al. Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. Nat Genet. 2000;26: 191–194. doi:10.1038/79911

98. Schüle B, McFarland KN, Lee K, Tsai Y-C, Nguyen K-D, Sun C, et al. Parkinson's disease associated with pure repeat expansion. NPJ Parkinsons Dis. 2017;3: 27. doi:10.1038/s41531-017-0029-x

99. Mouro Pinto R, Arning L, Giordano JV, Razghandi P, Andrew MA, Gillis T, et al. Patterns of CAG repeat instability in the central nervous system and periphery in Huntington's disease and in spinocerebellar ataxia type 1. Hum Mol Genet. 2020;29: 2551–2567. doi:10.1093/hmg/ddaa139

100. Petrozziello T, Dios AM, Mueller KA, Vaine CA, Hendriks WT, Glajch KE, et al. SVA insertion in X-linked Dystonia Parkinsonism alters histone H3 acetylation associated with TAF1 gene. PLoS One. 2020;15: e0243655. doi:10.1371/journal.pone.0243655

101. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. Nature. 2020;581: 444–451. doi:10.1038/s41586-020-2287-8

102. Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, Dijkstra LJ, et al. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. Nat Commun. 2016;7: 12989. doi:10.1038/ncomms12989

103. Chiang C, GTEx Consortium, Scott AJ, Davis JR, Tsang EK, Li X, et al. The impact of structural variation on human gene expression. Nature Genetics. 2017. pp. 692–699. doi:10.1038/ng.3834

104. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet. 2010;86: 749–764. doi:10.1016/j.ajhg.2010.04.006

105. Dharmadhikari AV, Ghosh R, Yuan B, Liu P, Dai H, Al Masri S, et al. Copy number variant and runs of homozygosity detection by microarrays enabled more precise molecular diagnoses in 11,020 clinical exome cases. Genome Med. 2019;11: 30. doi:10.1186/s13073-019-0639-5

106. Weiss MM, Hermsen MA, Meijer GA, van Grieken NC, Baak JP, Kuipers EJ, et al. Comparative genomic hybridisation. Mol Pathol. 1999;52: 243–251. doi:10.1136/mp.52.5.243

107. Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S, et al. High-Resolution Analysis of DNA Copy Number Using Oligonucleotide Microarrays. Genome Research. 2004. pp. 287–295. doi:10.1101/gr.2012304

108. Oostlander AE, Meijer GA, Ylstra B. Microarray-based comparative genomic hybridization and its applications in human genetics. Clin Genet. 2004;66: 488–495. doi:10.1111/j.1399-0004.2004.00322.x

109. Bejjani BA, Theisen AP, Ballif BC, Shaffer LG. Array-based comparative genomic hybridization in clinical diagnosis. Expert Rev Mol Diagn. 2005;5: 421–429. doi:10.1586/14737159.5.3.421

110. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet. 2008;40: 1166–1174. doi:10.1038/ng.238

111. Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat Genet. 2008;40: 1199–1203. doi:10.1038/ng.236

112. Gordeeva V, Sharova E, Arapidi G. Progress in Methods for Copy Number Variation Profiling. Int J Mol Sci. 2022;23. doi:10.3390/ijms23042143

113. Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, et al. Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet. 2009;84: 148–161. doi:10.1016/j.ajhg.2008.12.014

114. Leslie Rubin I, Merrick J, Greydanus DE, Patel DR. Health Care for People with Intellectual and Developmental Disabilities across the Lifespan. Springer; 2016. Available: https://play.google.com/store/books/details?id=HUIWDAAAQBAJ

115. Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BWM, Willemsen MH, et al. Genome sequencing identifies major causes of severe intellectual disability. Nature. 2014;511: 344–347. doi:10.1038/nature13394

116. Thygesen JH, Wolfe K, McQuillin A, Viñas-Jornet M, Baena N, Brison N, et al. Neurodevelopmental risk copy number variants in adults with intellectual disabilities and comorbid psychiatric disorders. Br J Psychiatry. 2018;212: 287–294. doi:10.1192/bjp.2017.65

117. Li YR, Glessner JT, Coe BP, Li J, Mohebnasab M, Chang X, et al. Rare copy number variants in over 100,000 European ancestry subjects reveal multiple disease associations. Nat Commun. 2020;11: 255. doi:10.1038/s41467-019-13624-1

118. Dellinger AE, Saw S-M, Goh LK, Seielstad M, Young TL, Li Y-J. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res. 2010;38: e105. doi:10.1093/nar/gkq040

119. Roy S, Motsinger Reif A. Evaluation of calling algorithms for array-CGH. Front Genet. 2013;4: 217. doi:10.3389/fgene.2013.00217

120. Winchester L, Yau C, Ragoussis J. Comparing CNV detection methods for SNP arrays. Brief Funct Genomic Proteomic. 2009;8: 353–366. doi:10.1093/bfgp/elp017

121. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464: 704–712. doi:10.1038/nature08516

122. Ferreira EN, Quaio CRD. Copy Number Variation in the Human Genome. Human Genome Structure, Function and Clinical Considerations. 2021. pp. 275–300. doi:10.1007/978-3-030-73151-9_9

123. Locke DP, Segraves R, Nicholls RD, Schwartz S, Pinkel D, Albertson DG, et al. BAC microarray analysis of 15q11-q13 rearrangements and the impact of segmental duplications. J Med Genet. 2004;41: 175–182. doi:10.1136/jmg.2003.013813

124. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12: 363–376. doi:10.1038/nrg2958

125. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, et al. Paired-end mapping reveals extensive structural variation in the human genome. Science. 2007;318: 420–426. doi:10.1126/science.1149504

126. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, et al. Author Correction: A structural variation reference for medical and population genetics. Nature. 2021;590: E55. doi:10.1038/s41586-020-03176-6

127. Gross AM, Ajay SS, Rajan V, Brown C, Bluske K, Burns NJ, et al. Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease. Genet Med. 2019;21: 1121–1130. doi:10.1038/s41436-018-0295-y

128. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin Chem. 2009;55: 641–658. doi:10.1373/clinchem.2008.112789

129. Abul-Husn NS, Manickam K, Jones LK, Wright EA, Hartzel DN, Gonzaga-Jauregui C, et al. Genetic identification of familial hypercholesterolemia within a single U.S. health care system. Science. 2016. doi:10.1126/science.aaf7000

130. Schwartz MLB, McCormick CZ, Lazzeri AL, Lindbuchler DM, Hallquist MLG, Manickam K, et al. A Model for Genome-First Care: Returning Secondary Genomic Findings to Participants and Their Healthcare Providers in a Large Research Cohort. doi:10.1101/166975

131. Lin X, Yang Y, Melton PE, Singh V, Simpson-Yap S, Burdon KP, et al. Integrating genetic structural variations and whole-genome sequencing into clinical neurology. Neurol Genet. 2022;8: e200005. doi:10.1212/nxg.0000000000200005

132. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988;85: 2444–2448. doi:10.1073/pnas.85.8.2444

133. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25: 3389–3402. doi:10.1093/nar/25.17.3389

134. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21: 1859–1875. doi:10.1093/bioinformatics/bti310

135. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, et al. Human-mouse alignments with BLASTZ. Genome Res. 2003;13: 103–107. doi:10.1101/gr.809403

136. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–359. doi:10.1038/nmeth.1923

137. Marco-Sola S, Sammeth M, Guigó R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods. 2012;9: 1185–1188. doi:10.1038/nmeth.2221

138. Zhou Q, Lim J-Q, Sung W-K, Li G. An integrated package for bisulfite DNA methylation data analysis with Indel-sensitive mapping. BMC Bioinformatics. 2019;20: 47. doi:10.1186/s12859-018-2593-4

139. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26: 589–595. doi:10.1093/bioinformatics/btp698

140. Harris EY, Ponts N, Le Roch KG, Lonardi S. BRAT-BW: efficient and accurate mapping of bisulfite-treated reads. Bioinformatics. 2012;28: 1795–1796. doi:10.1093/bioinformatics/bts264

141. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009;10: 232. doi:10.1186/1471-2105-10-232

142. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6: 31. doi:10.1186/1471-2105-6-31

143. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410. doi:10.1016/S0022-2836(05)80360-2

144. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12: 656–664. doi:10.1101/gr.229202

145. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010;26: 873–881. doi:10.1093/bioinformatics/btq057

146. Huang S, Zhang J, Li R, Zhang W, He Z, Lam T-W, et al. SOAPsplice: Genome-Wide ab initio Detection of Splice Junctions from RNA-Seq Data. Front Genet. 2011;2: 46. doi:10.3389/fgene.2011.00046

147. Fan X, Abbott TE, Larson D, Chen K. BreakDancer: Identification of Genomic Structural Variation from Paired-End Read Mapping. Current Protocols in Bioinformatics. 2014. doi:10.1002/0471250953.bi1506s45

148. Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol. 2009;10: R23. doi:10.1186/gb-2009-10-2-r23

149. Hayes M, Pyon YS, Li J. A model-based clustering method for genomic structural variant prediction and genotyping using paired-end sequencing data. PLoS One. 2012;7: e52881. doi:10.1371/journal.pone.0052881

150. Zhang ZD, Du J, Lam H, Abyzov A, Urban AE, Snyder M, et al. Identification of genomic indels and structural variations using split reads. BMC Genomics. 2011;12: 375. doi:10.1186/1471-2164-12-375

151. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28: i333–i339. doi:10.1093/bioinformatics/bts378

152. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 2009;19: 1586–1592. doi:10.1101/gr.092981.109

153. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput Biol. 2016;12: e1004873. doi:10.1371/journal.pcbi.1004873

154. Michaelson JJ, Sebat J. forestSV: structural variant discovery through statistical learning. Nat Methods. 2012;9: 819–821. doi:10.1038/nmeth.2085

155. Parikh H, Mohiyuddin M, Lam HYK, Iyer H, Chen D, Pratt M, et al. svclassify: a method to establish benchmark structural variant calls. BMC Genomics. 2016;17: 64. doi:10.1186/s12864-016-2366-2

156. Cai L, Wu Y, Gao J. DeepSV: accurate calling of genomic deletions from high-throughput sequencing data using deep convolutional neural network. BMC Bioinformatics. 2019;20: 665. doi:10.1186/s12859-019-3299-y

157. Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A. 2001;98: 9748–9753. doi:10.1073/pnas.171285098

158. Nijkamp JF, van den Broek MA, Geertman J-MA, Reinders MJT, Daran J-MG, de Ridder D. De novo detection of copy number variation by co-assembly. Bioinformatics. 2012;28: 3195–3202. doi:10.1093/bioinformatics/bts601

159. Narzisi G, O'Rawe JA, Iossifov I, Fang H, Lee Y-H, Wang Z, et al. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. Nat Methods. 2014;11: 1033–1036. doi:10.1038/nmeth.3069

160. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;32: 1220–1222. doi:10.1093/bioinformatics/btv710

161. Wala JA, Bandopadhayay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. 2018;28: 581–591. doi:10.1101/gr.221028.117

162. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40: e72. doi:10.1093/nar/gks001

163. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet. 2018;19: 329–346. doi:10.1038/s41576-018-0003-4

164. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019;20: 246. doi:10.1186/s13059-019-1828-7

165. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323: 133–138. doi:10.1126/science.1162986

166. Chen P, Gu J, Brandin E, Kim Y-R, Wang Q, Branton D. PROBING SINGLE DNA MOLECULE TRANSPORT USING FABRICATED NANOPORES. Nano Lett. 2004;4: 2293–2298. doi:10.1021/nl048654j

167. Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. Nat Biotechnol. 2016;34: 518–524. doi:10.1038/nbt.3423

168. Hon T, Mars K, Young G, Tsai Y-C, Karalius JW, Landolin JM, et al. Highly accurate long-read HiFi sequencing data for five complex genomes. Sci Data. 2020;7: 399. doi:10.1038/s41597-020-00743-4

169. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021;39: 1348–1365. doi:10.1038/s41587-021-01108-x

170. Jiang T, Liu S, Cao S, Liu Y, Cui Z, Wang Y, et al. Long-read sequencing settings for efficient structural variation detection based on comprehensive evaluation. BMC Bioinformatics. 2021;22: 552. doi:10.1186/s12859-021-04422-y

171. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10: 1784. doi:10.1038/s41467-018-08148-z

172. Scott AJ, Chiang C, Hall IM. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. Genome Res. 2021. doi:10.1101/gr.275488.121

173. Quan C, Lu H, Lu Y, Zhou G. Population-scale genotyping of structural variation in the era of long-read sequencing. Comput Struct Biotechnol J. 2022;20: 2639–2647. doi:10.1016/j.csbj.2022.05.047

174. Miller DE, Sulovari A, Wang T, Loucks H, Hoekzema K, Munson KM, et al. Targeted long-read sequencing identifies missing disease-causing variation. Am J Hum Genet. 2021;108: 1436–1449. doi:10.1016/j.ajhg.2021.06.006

175. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. 2017;27: 677–685. doi:10.1101/gr.214007.116

176. Wagner J, Olson ND, Harris L, McDaniel J, Cheng H, Fungtammasan A, et al. Curated variation benchmarks for challenging medically relevant autosomal genes. Nat Biotechnol. 2022;40: 672–680. doi:10.1038/s41587-021-01158-1

177. Noyes MD, Harvey WT, Porubsky D, Sulovari A, Li R, Rose NR, et al. Familial long-read sequencing increases yield of de novo mutations. Am J Hum Genet. 2022;109: 631–646. doi:10.1016/j.ajhg.2022.02.014

178. Sone J, Mitsuhashi S, Fujita A, Mizuguchi T, Hamanaka K, Mori K, et al. Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. Nat Genet. 2019;51: 1215–1221. doi:10.1038/s41588-019-0459-y

179. Giesselmann P, Brändl B, Raimondeau E, Bowen R, Rohrandt C, Tandon R, et al. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. Nature Biotechnology. 2019. pp. 1478–1481. doi:10.1038/s41587-019-0293-x

180. Zeng S, Zhang M-Y, Wang X-J, Hu Z-M, Li J-C, Li N, et al. Long-read sequencing identified intronic repeat expansions in from Chinese pedigrees affected with familial cortical myoclonic tremor with epilepsy. J Med Genet. 2019;56: 265–270. doi:10.1136/jmedgenet-2018-105484

181. Stevanovski I, Chintalaphani SR, Gamaarachchi H, Ferguson JM, Pineda SS, Scriba CK, et al. Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. Sci Adv. 2022;8: eabm5386. doi:10.1126/sciadv.abm5386

182. Jenko Bizjan B, Katsila T, Tesovnik T, Šket R, Debeljak M, Matsoukas MT, et al. Challenges in identifying large germline structural variants for clinical use by long read sequencing. Comput Struct Biotechnol J. 2020;18: 83–92. doi:10.1016/j.csbj.2019.11.008

183. Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, et al. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nat Genet. 2021;53: 779–786. doi:10.1038/s41588-021-00865-4

184. Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, et al. Pangenome Graphs. Annu Rev Genomics Hum Genet. 2020;21: 139–162. doi:10.1146/annurev-genom-120219-080406

185. Miga KH, Wang T. The Need for a Human Pangenome Reference Sequence. Annu Rev Genomics Hum Genet. 2021;22: 81–102. doi:10.1146/annurev-genom-120120-081921

186. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36: 338–345. doi:10.1038/nbt.4060

187. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, et al. Telomere-to-telomere assembly of a complete human X chromosome. Nature. 2020;585: 79–84. doi:10.1038/s41586-020-2547-7

188. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nat Biotechnol. 2020;38: 1044–1053. doi:10.1038/s41587-020-0503-6

189. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013;29: 2669–2677. doi:10.1093/bioinformatics/btt476

190. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376: 44–53. doi:10.1126/science.abj6987

191. Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, et al. The Human Pangenome Project: a global resource to map genomic diversity. Nature. 2022;604: 437–446. doi:10.1038/s41586-022-04601-8

192. Yan SM, Sherman RM, Taylor DJ, Nair DR, Bortvin AN, Schatz MC, et al. Local adaptation and archaic introgression shape global diversity at human structural variant loci. Elife. 2021;10. doi:10.7554/eLife.67615

193. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science. 2021;372. doi:10.1126/science.abf7117

194. Sano Y, Koyanagi Y, Wong JH, Murakami Y, Fujiwara K, Endo M, et al. Likely pathogenic structural variants in genetically unsolved patients with retinitis pigmentosa revealed by long-read sequencing. J Med Genet. 2022;59: 1133–1138. doi:10.1136/jmedgenet-2022-108428

195. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17: 405–424. doi:10.1038/gim.2015.30

196. Kleinert P, Kircher M. A framework to score the effects of structural variants in health and disease. Genome Res. 2022;32: 766–777. doi:10.1101/gr.275995.121

197. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen--the Clinical Genome Resource. N Engl J Med. 2015;372: 2235–2242. doi:10.1056/NEJMsr1406261

198. Riggs ER, Andersen EF, Cherry AM, Kantarci S, Kearney H, Patel A, et al. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). Genet Med. 2020;22: 245–257. doi:10.1038/s41436-019-0686-8

199. Minoche AE, Lundie B, Peters GB, Ohnesorg T, Pinese M, Thomas DM, et al. ClinSV: clinical grade structural and copy number variant detection from whole genome sequencing data. Genome Med. 2021;13: 32. doi:10.1186/s13073-021-00841-x

200. Brandt T, Sack LM, Arjona D, Tan D, Mei H, Cui H, et al. Adapting ACMG/AMP sequence variant classification guidelines for single-gene copy number variants. Genet Med. 2020;22: 336–344. doi:10.1038/s41436-019-0655-2

201. Nicora G, Zucca S, Limongelli I, Bellazzi R, Magni P. A machine learning approach based on ACMG/AMP guidelines for genomic variant classification and prioritization. Sci Rep. 2022;12: 2517. doi:10.1038/s41598-022-06547-3

202. Danis D, Jacobsen JOB, Balachandran P, Zhu Q, Yilmaz F, Reese J, et al. SvAnna: efficient and accurate pathogenicity prediction of coding and regulatory structural variants in long-read genome sequencing. Genome Med. 2022;14: 44. doi:10.1186/s13073-022-01046-6

203. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44: D733–45. doi:10.1093/nar/gkv1189

204. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47: D506–D515. doi:10.1093/nar/gky1049

205. Kalvari I, Nawrocki EP, Argasinska J, Quinones-Olvera N, Finn RD, Bateman A, et al. Non-coding RNA analysis using the rfam database. Curr Protoc Bioinformatics. 2018;62: e51. doi:10.1002/cpbi.51

206. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47: D766–D773. doi:10.1093/nar/gky955

207. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74. doi:10.1038/nature11247

208. Alam T, Agrawal S, Severin J, Young RS, Andersson R, Arner E, et al. Comparative transcriptomics of primary cells in vertebrates. Genome Res. 2020;30: 951–961. doi:10.1101/gr.255679.119

209. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Rep. 2016;17: 2042–2059. doi:10.1016/j.celrep.2016.10.061

210. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions in the human genome. Nat Genet. 2019;51: 88–95. doi:10.1038/s41588-018-0294-6

211. Chen Z, Zhang D, Reynolds RH, Gustavsson EK, García-Ruiz S, D'Sa K, et al. Human-lineage-specific genomic elements are associated with neurodegenerative disease and APOE transcript usage. Nat Commun. 2021;12: 2076. doi:10.1038/s41467-021-22262-5

212. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15: 1034–1050. doi:10.1101/gr.3715005

213. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10: 57–63. doi:10.1038/nrg2484

214. Glinos DA, Garborcauskas G, Hoffman P, Ehsan N, Jiang L, Gokden A, et al. Transcriptome variation in human tissues revealed by long-read sequencing. Nature. 2022;608: 353–359. doi:10.1038/s41586-022-05035-y

215. Halperin RF, Hegde A, Lang JD, Raupach EA, C4RCD Research Group, Legendre C, et al. Improved methods for RNAseq-based alternative splicing analysis. Sci Rep. 2021;11: 10740. doi:10.1038/s41598-021-89938-2

216. Oluwadare O, Highsmith M, Cheng J. An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data. Biol Proced Online. 2019;21: 7. doi:10.1186/s12575-019-0094-0

217. Yan F, Powell DR, Curtis DJ, Wong NC. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome Biol. 2020;21: 22. doi:10.1186/s13059-020-1929-3

218. Fallah MS, Szarics D, Robson CM, Eubanks JH. Impaired Regulation of Histone Methylation and Acetylation Underlies Specific Neurodevelopmental Disorders. Front Genet. 2020;11: 613098. doi:10.3389/fgene.2020.613098

219. Martin BJE, Brind'Amour J, Kuzmin A, Jensen KN, Liu ZC, Lorincz M, et al. Transcription shapes genome-wide histone acetylation patterns. Nat Commun. 2021;12: 210. doi:10.1038/s41467-020-20543-z

220. Hannon E, Dempster E, Viana J, Burrage J, Smith AR, Macdonald R, et al. An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. Genome Biol. 2016;17: 176. doi:10.1186/s13059-016-1041-x

221. Alfaro JA, Bohländer P, Dai M, Filius M, Howard CJ, van Kooten XF, et al. The emerging landscape of single-molecule protein sequencing technologies. Nat Methods. 2021;18: 604–617. doi:10.1038/s41592-021-01143-1

222. Steen H, Mann M. The ABC's (and XYZ's) of peptide sequencing. Nat Rev Mol Cell Biol. 2004;5: 699–711. doi:10.1038/nrm1468

223. Lee J, Hyeon DY, Hwang D. Single-cell multiomics: technologies and data analysis methods. Exp Mol Med. 2020;52: 1428–1442. doi:10.1038/s12276-020-0420-2

224. Robinson P, Jtel TZ. Integrative genomics viewer (IGV): Visualizing alignments and variants. Computational Exome and Genome Analysis. 2017. pp. 233–245. doi:10.1201/9781315154770-17

225. Belyeu JR, Chowdhury M, Brown J, Pedersen BS, Cormier MJ, Quinlan AR, et al. Samplot: a platform for structural variant visual validation and automated filtering. Genome Biol. 2021;22: 161. doi:10.1186/s13059-021-02380-5

226. Nattestad M, Aboukhalil R, Chin C-S, Schatz MC. Ribbon: intuitive visualization for complex genomic variation. Bioinformatics. 2021;37: 413–415. doi:10.1093/bioinformatics/btaa680

227. Spies N, Zook JM, Salit M, Sidow A. svviz: a read viewer for validating structural variants. Bioinformatics. 2015;31: 3994–3996. doi:10.1093/bioinformatics/btv478

228. Liu Z, Roberts R, Mercer TR, Xu J, Sedlazeck FJ, Tong W. Author Correction: Towards accurate and reliable resolution of structural variants for clinical diagnosis. Genome Biol. 2022;23: 198. doi:10.1186/s13059-022-02773-0

229. Malamon JS, Farrell JJ, Xia LC, Dombroski BA, Lee W-P, Das RG, et al. A comparative study of structural variant calling strategies using the Alzheimer's Disease Sequencing Project's whole genome family data. doi:10.1101/2022.05.19.492472

230. Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, et al. A robust benchmark for detection of germline large deletions and insertions. Nat Biotechnol. 2020;38: 1347–1355. doi:10.1038/s41587-020-0538-8

231. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, et al. Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. Cell. 2020;182: 145–161.e23. doi:10.1016/j.cell.2020.05.021

232. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550. doi:10.1186/s13059-014-0550-8

233. Brechtmann F, Mertes C, Matusevičiūtė A, Yépez VA, Avsec Ž, Herzog M, et al. OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. Am J Hum Genet. 2018;103: 907–917. doi:10.1016/j.ajhg.2018.10.025

234. Li W, Freudenberg J. Mappability and read length. Front Genet. 2014;5: 381. doi:10.3389/fgene.2014.00381

235. Ren J, Chaisson MJP. lra: A long read aligner for sequences and contigs. PLOS Computational Biology. 2021. p. e1009078. doi:10.1371/journal.pcbi.1009078

236. Jain C, Rhie A, Zhang H, Chu C, Walenz BP, Koren S, et al. Weighted minimizer sampling improves long read mapping. Bioinformatics. 2020;36: i111–i118. doi:10.1093/bioinformatics/btaa435

237. Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, et al. Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. 2020;21: 189. doi:10.1186/s13059-020-02107-y

238. English AC, Menon VK, Gibbs R, Metcalf GA, Sedlazeck FJ. Truvari: Refined structural variant comparison preserves Allelic diversity. bioRxiv. 2022. doi:10.1101/2022.02.21.481353

239. Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genetics in Medicine. 2018. pp. 159–163. doi:10.1038/gim.2017.86

240. Chander V, Gibbs RA, Sedlazeck FJ. Evaluation of computational genotyping of structural variation for clinical diagnoses. Gigascience. 2019;8. doi:10.1093/gigascience/giz110

241. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15: 461–468. doi:10.1038/s41592-018-0001-7

242. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34: 3094–3100. doi:10.1093/bioinformatics/bty191

243. Bolognini D, Magi A. Evaluation of Germline Structural Variant Calling Methods for Nanopore Sequencing Data. Front Genet. 2021;12: 761791. doi:10.3389/fgene.2021.761791

244. Biémont C. A brief history of the status of transposable elements: from junk DNA to major players in evolution. Genetics. 2010;186: 1085–1093. doi:10.1534/genetics.110.124180

245. Lu JY, Shao W, Chang L, Yin Y, Li T, Zhang H, et al. Genomic Repeats Categorize Genes with Distinct Functions for Orchestrated Regulation. Cell Rep. 2020;30: 3296–3311.e5. doi:10.1016/j.celrep.2020.02.048

246. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. 2013;14: 125–138. doi:10.1038/nrg3373

247. González F, Georgieva D, Vanoli F, Shi Z-D, Stadtfeld M, Ludwig T, et al. Homologous recombination DNA repair genes play a critical role in reprogramming to a pluripotent state. Cell Rep. 2013;3: 651–660. doi:10.1016/j.celrep.2013.02.005

248. Tilgner K, Neganova I, Moreno-Gimeno I, Al-Aama JY, Burks D, Yung S, et al. A human iPSC model of Ligase IV deficiency reveals an important role for NHEJ-mediated-DSB repair in the survival and genomic stability of induced pluripotent stem cells and emerging haematopoietic progenitors. Cell Death Differ. 2013;20: 1089–1100. doi:10.1038/cdd.2013.44

249. Bressan E, Reed X, Bansal V, Hutchins E, Cobb MM, Webb MG, et al. The Foundational data initiative for Parkinson's disease (FOUNDIN-PD): enabling efficient translation from genetic maps to mechanism. bioRxiv. bioRxiv; 2021. doi:10.1101/2021.06.03.446785

250. Assou S, Girault N, Plinet M, Bouckenheimer J, Sansac C, Combe M, et al. Recurrent Genetic Abnormalities in Human Pluripotent Stem Cells: Definition and Routine Detection in Culture Supernatant by Targeted Droplet Digital PCR. Stem Cell Reports. 2020;14: 1–8. doi:10.1016/j.stemcr.2019.12.004

251. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581: 434–443. doi:10.1038/s41586-020-2308-7

252. Han L, Zhao X, Benton ML, Perumal T, Collins RL, Hoffman GE, et al. Functional annotation of rare structural variation in the human brain. Nat Commun. 2020;11: 2990. doi:10.1038/s41467-020-16736-1

253. Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature. 2016;538: 265–269. doi:10.1038/nature19800

254. Brnich SE, Abou Tayoun AN, Couch FJ, Cutting GR, Greenblatt MS, Heinen CD, et al. Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. Genome Med. 2019;12: 3. doi:10.1186/s13073-019-0690-2

255. ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583: 699–710. doi:10.1038/s41586-020-2493-4

256. Ramamurthy E, Welch G, Cheng J, Yuan Y, Gunsalus L, Bennett DA, et al. Cell type-specific histone acetylation profiling of Alzheimer's Disease subjects and integration with genetics. bioRxiv. bioRxiv; 2020. doi:10.1101/2020.03.26.010330

257. Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, et al. The impact of structural variation on human gene expression. doi:10.1101/055962

258. Hashemabadi M, Sasan H, Amandadi M, Esmaeilzadeh-Salestani K, Esmaeili-Mahani S, Ravan H. CRISPR/Cas9-Mediated Disruption of ZNF543 Gene: An Approach Toward Discovering Its Relation to TRIM28 Gene in Parkinson's Disease. Mol Biotechnol. 2022. doi:10.1007/s12033-022-00494-0

259. Fujikawa R, Higuchi S, Nakatsuji M, Yasui M, Ikedo T, Nagata M, et al. Deficiency in EP4 Receptor-Associated Protein Ameliorates Abnormal Anxiety-Like Behavior and Brain Inflammation in a Mouse Model of Alzheimer Disease. Am J Pathol. 2017;187: 1848–1854. doi:10.1016/j.ajpath.2017.04.010

260. Jacobs FMJ, van der Linden AJA, Wang Y, von Oerthel L, Sul HS, Burbach JPH, et al. Identification of Dlk1, Ptpru and Klhl1 as novel Nurr1 target genes in meso-diencephalic dopamine neurons. Development. 2009;136: 2363–2373. doi:10.1242/dev.037556

261. Varga B, Barabás O, Kovári J, Tóth J, Hunyadi-Gulyás E, Klement E, et al. Active site closure facilitates juxtaposition of reactant atoms for initiation of catalysis by human dUTPase. FEBS Lett. 2007;581: 4783–4788. doi:10.1016/j.febslet.2007.09.005

262. Ashouri S, Wong JH, Nakagawa H, Shimada M, Tokunaga K, Fujimoto A. Characterization of intermediate-sized insertions using whole-genome sequencing data and analysis of their functional impact on gene expression. Hum Genet. 2021;140: 1201–1216. doi:10.1007/s00439-021-02291-2

263. Hurles ME, Dermitzakis ET, Tyler-Smith C. The functional impact of structural variation in humans. Trends Genet. 2008;24: 238–245. doi:10.1016/j.tig.2008.03.001

264. Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat Methods. 2019;16: 1297–1305. doi:10.1038/s41592-019-0617-2

265. Tilgner H, Grubert F, Sharon D, Snyder MP. Defining a personal, allele-specific, and single-molecule long-read transcriptome. Proc Natl Acad Sci U S A. 2014;111: 9869–9874. doi:10.1073/pnas.1400447111

266. Oppermann U. Carbonyl reductases: the complex relationships of mammalian carbonyl- and quinone-reducing enzymes and their role in physiology. Annu Rev Pharmacol Toxicol. 2007;47: 293–322. doi:10.1146/annurev.pharmtox.47.120505.105316

267. Botella JA, Ulschmid JK, Gruenewald C, Moehle C, Kretzschmar D, Becker K, et al. The Drosophila carbonyl reductase sniffer prevents oxidative stress-induced neurodegeneration. Curr Biol. 2004;14: 782–786. doi:10.1016/j.cub.2004.04.036

268. Jang M, Kim Y, Won H, Lim S, K R J, Dashdorj A, et al. Carbonyl reductase 1 offers a novel therapeutic target to enhance leukemia treatment by arsenic trioxide. Cancer Res. 2012;72: 4214–4224. doi:10.1158/0008-5472.CAN-12-1110

269. Solimena M, Dirkx R Jr, Hermel JM, Pleasic-Williams S, Shapiro JA, Caron L, et al. ICA 512, an autoantigen of type I diabetes, is an intrinsic membrane protein of neurosecretory granules. EMBO J. 1996;15: 2102–2114. Available: https://www.ncbi.nlm.nih.gov/pubmed/8641276

270. Labandeira CM, Fraga-Bau A, Arias Ron D, Alvarez-Rodriguez E, Vicente-Alba P, Lago-Garma J, et al. Parkinson's disease and diabetes mellitus: common mechanisms and treatment repurposing. Neural Regeneration Res. 2022;17: 1652–1658. doi:10.4103/1673-5374.332122

271. Lee S. The association of genetically controlled CpG methylation (cg158269415) of protein tyrosine phosphatase, receptor type N2 (PTPRN2) with childhood obesity. Sci Rep. 2019;9: 4855. doi:10.1038/s41598-019-40486-w

272. Chuang Y-H, Lu AT, Paul KC, Folle AD, Bronstein JM, Bordelon Y, et al. Longitudinal Epigenome-Wide Methylation Study of Cognitive Decline and Motor Progression in Parkinson's Disease. J Parkinsons Dis. 2019;9: 389–400. doi:10.3233/JPD-181549

273. Herold C, Hooli BV, Mullin K, Liu T, Roehr JT, Mattheisen M, et al. Family-based association analyses of imputed genotypes reveal genome-wide significant association of Alzheimer's disease with OSBPL6, PTPRG, and PDCL3. Mol Psychiatry. 2016;21: 1608–1612. doi:10.1038/mp.2015.218

274. Luo J, Huang X, Li R, Xie J, Chen L, Zou C, et al. PTPRG activates m6A methyltransferase VIRMA to block mitochondrial autophagy mediated neuronal death in Alzheimer's disease. bioRxiv. 2022. doi:10.1101/2022.03.11.22272061

275. Boni C, Laudanna C, Sorio C. A Comprehensive Review of Receptor-Type Tyrosine-Protein Phosphatase Gamma (PTPRG) Role in Health and Non-Neoplastic Disease. Biomolecules. 2022;12. doi:10.3390/biom12010084

276. Petrovic D, Kouroussis E, Vignane T, Filipovic MR. The Role of Protein Persulfidation in Brain Aging and Neurodegeneration. Front Aging Neurosci. 2021;13: 674135. doi:10.3389/fnagi.2021.674135

277. Katsouda A, Valakos D, Dionellis VS, Bibli S-I, Akoumianakis I, Karaliota S, et al. MPST sulfurtransferase maintains mitochondrial protein import and cellular bioenergetics to attenuate obesity. J Exp Med. 2022;219. doi:10.1084/jem.20211894

278. Scheijen EEM, Wilson DM 3rd. Genome Integrity and Neurological Disease. Int J Mol Sci. 2022;23. doi:10.3390/ijms23084142

9 Statement of contributions

SV Calling: Long read aligner benchmarking

Anastasia Illarionova, Zih-Hua Fang, Vikas Bansal, Peter Heutink

Personal contribution: experiment design of the study (together with ZF, VB and PH), ONT DNA sequencing data analysis, design and implementation of the SV calling pipeline, statistical analysis, and figure preparation.

Others: ZF, VB and PH helped with the experiment design.

SV detection and annotation on FOUNDIN PD cases: database construction

Anastasia Illarionova, Zih-Hua Fang, Elisangela Bressan, Noemia Fernandes, Patrizia Rizzu, Peter Heutink

Personal contribution: experiment design of the study (together with ZF and PH), ONT DNA sequencing data analysis, implementation of the SV calling pipeline, statistical analysis, and figure preparation.

Others: EB differentiated and collected the PSC lines. NF and PR performed the ONT DNA sequencing. ZF and PH helped with the experiment design.

SV functional annotation and PD risk factor prioritization

Anastasia Illarionova, Zih-Hua Fang, Vikas Bansal, Natalia Savytska, Elisangela Bressan, Noemia Fernandes, Patrizia Rizzu, Peter Heutink

Personal contribution: experiment design of the study (together with ZF, VB, NS, and PH), ONT RNA sequencing data analysis, differential gene expression analysis, expression outlier analysis, transcript identification, annotation and differential transcript usage analysis, statistical analysis, and figure preparation.

Others: EB differentiated and collected the iPSC lines. NF and PR performed the ONT RNA and scRNA sequencing, VB performed scRNA sequencing data analysis. ZF, VB, NS, and PH helped with the experiment design.

10 Acknowledgements

First, I would like to express my deepest appreciation to Prof. Dr. Peter Heutink for giving me the opportunity to be part of his department at the German Center for Neurodegenerative Diseases. He was a profound mentor for me. I could not have undertaken the journey through these interesting and challenging projects without his scientific guidance and constant encouragement. I would like to gratefully thank my Advisory Board members, Prof. Thomas Gasser and Dr. Stefan Bonn, for their insightful comments and useful suggestions. Additionally, this endeavor would not have been possible without the generous support from Michael J. Fox Foundation and Global Parkinson's Genetics Program, who financed my research.

I am also grateful to my colleagues and friends for their support and meaningful advice. I would like to thank Dr. Zih-Hua Fang for her time and effort she spent helping me with data analysis and methodology, support in technical issue troubleshooting, and project feedback sessions. I am very grateful to Dr. Vikas Bansal and Dr. Elisangela Bressan for their excellent support in regarding the projects and the great time we spent having captivating scientific discussions. Additionally, I am very grateful to Dr. Elisangela Bressan for her outstanding work in the iPSC lines differentiation and I very much appreciate Dr. Patrizia Rizzu and Noemia-Rita Alves-Fernandes for their excellent work in the generation of the long-read sequencing datasets, without which this study would not have been possible. Last but not least, I would like to gratefully thank my friend and colleague Natalia Savytska for the never-ending moral support, inspiration, and compelling discussions about the challenging and fascinating world of bioinformatics.

Furthermore, I would like to sincerely thank Dr. Kimberley Billingsley, Dr. Cornelis Blauwendraat and Prof. Fritz Sedlazeck for exceptional introduction to the topic and fruitful discussions, and all partners from Europe and the United States from the IPDGC and FOUNDIN-PD for the excellent and successful collaboration.

Lastly, I would be remiss in not mentioning my family for keeping my spirits and motivation high during this important part of my life.

11 Appendix

11.1 Abbreviations

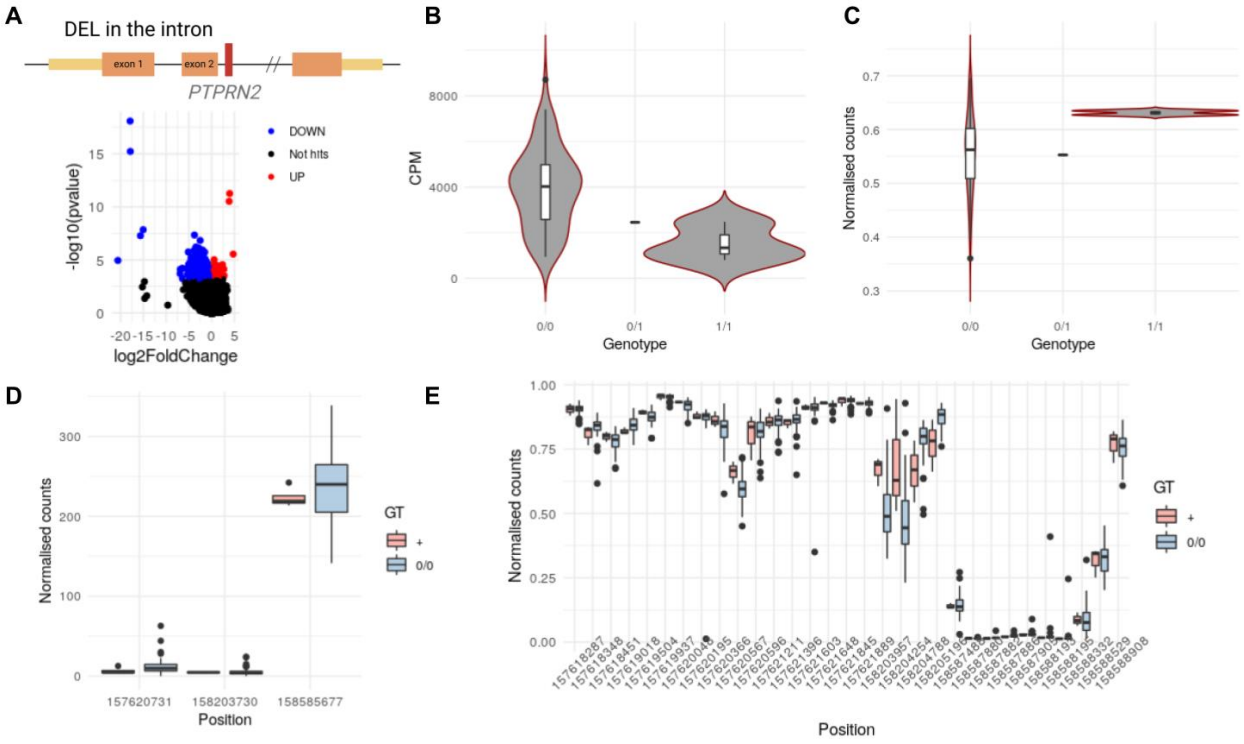
3GS	Third generation sequencing (LR sequencing)
CG	Copy gain
CNV	Copy number variant
DE(A)	Differential expression (analysis)
DEL	Deletion
DT(U)	Differential transcript (usage)
DUP	Duplication
eSV	SV associated with a DE gene
eoSV	SV associated with an expression outlier gene
GWAS	Genome-wide association study
IEDUP	Intra (whole) exon duplication
INS	Insertion
INV	Inversion
LR	Long read
NGS	Next generation sequencing (SR sequencing)
ONT	Oxford Nanopore Technologies
PD	Parkinson's disease
pLoF	Putative loss-of-function
PPMI	Parkinson's Progression Markers Initiative
ROS	Reactive oxygen species
SMRT	Single Molecule, Real-Time (SMRT) sequencing
SNP	Single nucleotide polymorphism
SR	Short read
SV	Structural variant
TAD	Topologically associated domain
WGS	Whole genome sequencing

11. 2 Supplementary Figures

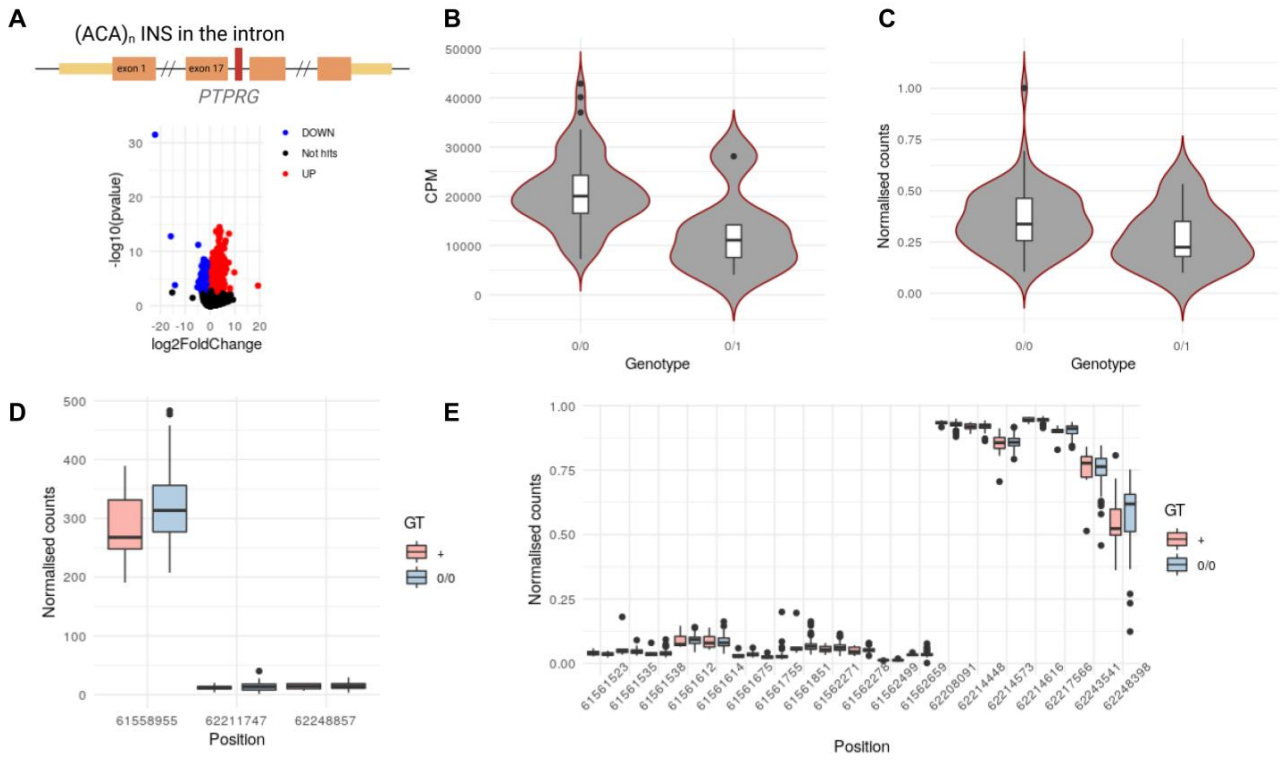
GO:BP		stats
Term name	Term ID	p_{adj}
multicellular organism development	GO:0007275	6.692×10^{-87}
system development	GO:0048731	2.030×10^{-85}
regulation of cellular process	GO:0050794	6.433×10^{-83}
nervous system development	GO:0007399	1.953×10^{-80}
positive regulation of cellular process	GO:0048522	1.884×10^{-79}
anatomical structure development	GO:0048856	1.888×10^{-78}
positive regulation of biological process	GO:0048518	1.115×10^{-74}
developmental process	GO:0032502	7.486×10^{-73}
anatomical structure morphogenesis	GO:0009653	2.897×10^{-71}
neurogenesis	GO:0022008	2.477×10^{-69}
regulation of developmental process	GO:0050793	2.477×10^{-69}
generation of neurons	GO:0048699	3.224×10^{-66}
positive regulation of nitrogen compound metabolic process	GO:0051173	1.340×10^{-65}
positive regulation of cellular metabolic process	GO:0031325	5.526×10^{-64}
regulation of biological process	GO:0050789	1.079×10^{-63}
multicellular organismal process	GO:0032501	1.160×10^{-62}
regulation of signaling	GO:0023051	1.481×10^{-62}
regulation of nitrogen compound metabolic process	GO:0051171	8.247×10^{-62}
regulation of cell communication	GO:0010646	8.247×10^{-62}
negative regulation of cellular process	GO:0048523	8.286×10^{-62}

GO:CC		stats
Term name	Term ID	p_{adj}
cell junction	GO:0030054	1.388×10^{-68}
synapse	GO:0045202	8.899×10^{-65}
cytosol	GO:0005829	2.740×10^{-53}
neuron projection	GO:0043005	3.542×10^{-52}
nucleoplasm	GO:0005654	6.873×10^{-52}
cell projection	GO:0042995	6.446×10^{-50}
plasma membrane bounded cell projection	GO:0120025	5.182×10^{-48}
postsynapse	GO:0098794	7.317×10^{-48}
cytoplasm	GO:0005737	5.632×10^{-46}
glutamatergic synapse	GO:0098978	1.800×10^{-43}
axon	GO:0030424	6.789×10^{-43}
somatodendritic compartment	GO:0036477	2.512×10^{-36}
dendrite	GO:0030425	3.480×10^{-34}
dendritic tree	GO:0097447	4.811×10^{-34}
plasma membrane region	GO:0098590	1.537×10^{-32}
synaptic membrane	GO:0097060	1.915×10^{-32}
nuclear lumen	GO:0031981	2.298×10^{-32}
neuron to neuron synapse	GO:0098984	3.982×10^{-32}
asymmetric synapse	GO:0032279	1.007×10^{-28}
postsynaptic density	GO:0014069	4.867×10^{-28}

Supplemental Figure S1. Gene ontology (GO) enrichment analysis results of haploinsufficient genes affected by PD prevalent pLoF SVs. BP - biological process, CC - cellular component. The figure was generated in https://biit.cs.ut.ee/gprofiler.



Supplemental Figure S2. DEL (DEL-PTPRN2) in the intronic region of PTPRN2 and PTPRN2 expression in multi omics datasets by DEL-PTPRN2 GT groups. **A.** Schematic representation of DEL localization and volcano plot of DEA results for DEL-PTPRN2 GT split. **B.** CBR1 expression in bRNA-seq (day 65). **C.** CBR1 expression in DA clusters of scRN-seq (day 65). **D.** Chromatin availability detected by bATAC-seq upstream and within the PTPRN2 gene body (day 65). **E.** Methylation profile of PTPRN2 gene body and upstream region (day 65).



Supplemental Figure S3. INS (INS-PTPRG) in the intronic region of PTPRG and PTPRG expression in multi omics datasets by INS-PTPRG GT groups. **A.** Schematic representation of PTPRG localization and volcano plot of DEA results for INS-PTPRG GT split. **B.** PTPRG expression in bRNA-seq (day 65). **C.** PTPRG expression in DA clusters of scRN-seq (day 65). **D**. Chromatin availability detected by bATAC-seq in the region of PTPRG gene body (day 65). **E**. Methylation profile of PTPRG gene body and upstream region (day 65).

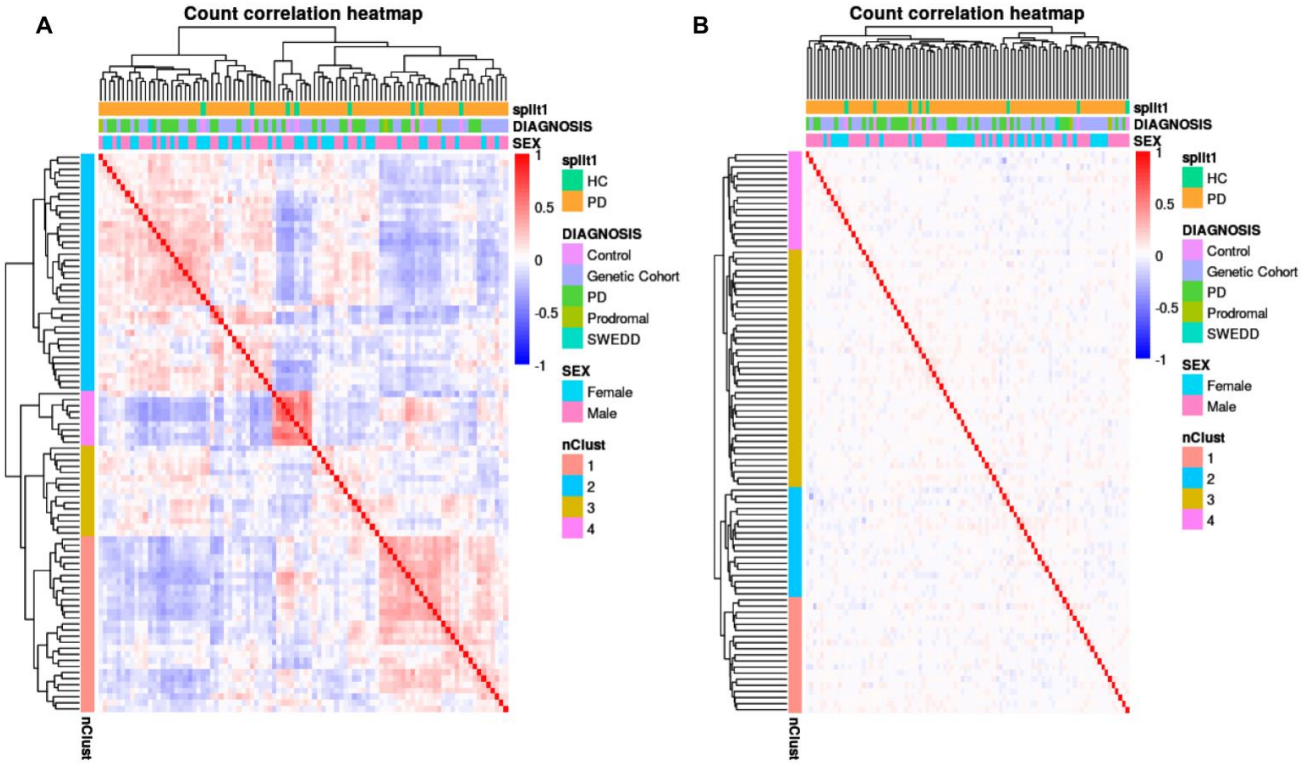
A

GO:BP		stats
Term name	Term ID	p_{adj}
anterograde trans-synaptic signaling	GO:0098916	7.645×10^{-67}
chemical synaptic transmission	GO:0007268	7.645×10^{-67}
trans-synaptic signaling	GO:0099537	2.337×10^{-66}
synaptic signaling	GO:0099536	1.787×10^{-65}
nervous system development	GO:0007399	1.195×10^{-55}
cell-cell signaling	GO:0007267	4.601×10^{-48}
generation of neurons	GO:0048699	4.044×10^{-42}
modulation of chemical synaptic transmission	GO:0050804	1.181×10^{-40}
regulation of trans-synaptic signaling	GO:0099177	1.269×10^{-40}
neuron development	GO:0048666	1.355×10^{-40}
neuron differentiation	GO:0030182	1.355×10^{-40}
neurogenesis	GO:0022008	1.848×10^{-39}
signaling	GO:0023052	1.798×10^{-38}
cell communication	GO:0007154	3.566×10^{-36}
neuron projection development	GO:0031175	6.568×10^{-36}
regulation of transport	GO:0051049	1.798×10^{-35}
regulation of localization	GO:0032879	7.493×10^{-35}
vesicle-mediated transport in synapse	GO:0099003	8.758×10^{-35}
synaptic vesicle cycle	GO:0099504	4.801×10^{-32}

B

GO:BP		stats
Term name	Term ID	p_{adj}
spliceosomal complex assembly	GO:0000245	5.438×10^{-5}
mRNA 5'-splice site recognition	GO:0000395	1.038×10^{-2}
mRNA cis splicing, via spliceosome	GO:0045292	1.234×10^{-2}
mRNA splice site selection	GO:0006376	1.271×10^{-2}
mRNA branch site recognition	GO:0000348	2.244×10^{-2}

Supplemental Figure S4. Gene ontology (GO) enrichment analysis results of DE genes obtained from the PTPRN2-DEL (A) and PTPRG-INS (B) GT splits. The figure was generated in https://biit.cs.ut.ee/gprofiler.



Supplemental Figure S5. Control for the confounding factors in the expression outlier analysis. The heatmap shows correlation for the samples before (**A**) and after (**B**) regulation for the batch effects and unknown covariation with autoencoder-based algorithm. Split1: HC – healthy controls, PD – PD cases and genetic cohort, Diagnosis (PPMI): Control – healthy controls, genetic cohort – affected and unaffected individuals with high risk PD mutation, prodromal – prodromal PD cases, SWEDD - scans without evidence for dopaminergic deficit.

GO:BP		stats
Term name	Term ID	p_{adj}
cellular protein modification process	GO:0006464	4.701×10^{-2}
organonitrogen compound metabolic process	GO:1901564	4.701×10^{-2}
regulation of chromatin organization	GO:1902275	4.701×10^{-2}
iron-sulfur cluster transmembrane transport	GO:1902497	4.701×10^{-2}
regulation of iron-sulfur cluster assembly	GO:1903329	4.701×10^{-2}
positive regulation of iron-sulfur cluster assembly	GO:1903331	4.701×10^{-2}
positive regulation of smooth muscle cell-matrix adhesion	GO:1905609	4.701×10^{-2}
response to epidermal growth factor	GO:0070849	4.701×10^{-2}
response to growth factor	GO:0070848	4.701×10^{-2}
cellular protein metabolic process	GO:0044267	4.701×10^{-2}
regulation of chromatin assembly	GO:0010847	4.701×10^{-2}
iron-sulfur cluster export from the mitochondrion	GO:0140466	4.701×10^{-2}
protein modification process	GO:0036211	4.701×10^{-2}
cellular response to platelet-derived growth factor stimulus	GO:0036120	4.701×10^{-2}
response to platelet-derived growth factor	GO:0036119	4.701×10^{-2}
negative regulation of dephosphorylation	GO:0035305	4.701×10^{-2}
regulation of dephosphorylation	GO:0035303	4.701×10^{-2}
regulation of heterochromatin organization	GO:0120261	4.701×10^{-2}
positive regulation of intracellular transport	GO:0032388	4.701×10^{-2}
regulation of cellular protein metabolic process	GO:0032268	4.701×10^{-2}
regulation of heterochromatin assembly	GO:0031445	4.701×10^{-2}
cellular macromolecule metabolic process	GO:0044260	4.701×10^{-2}

Supplemental Figure S6. Gene ontology (GO) enrichment analysis results for differentially expressed genes between MPST-DEL carriers and non-carriers. BP - biological process. The figure was generated in https://biit.cs.ut.ee/gprofiler.

Supplementary Table 1. iPSC specific CNV regions GRCh19-GRCh38 assembly build coordinates conversion.

Rank	Chr	Start	End	No. of CNVs	CNV No., %	Cumulated %
1	chr20	31260580	32166810	169	22.9	22.9
2	chr12	11784484	25403186	116	15.7	38.6
3	chr17	53204134	54216532	75	10.2	48.8
5	chr1	172930860	185830868	34	4.6	53.4
6	chr5	105164299	118068507	26	3.5	56.9
7	chr18	58532768	63932766	24	3.3	60.2
8	chr17	7307685	8140856	21	2.8	63
9	chr7	132915240	134101006	18	2.4	65.4
10	chr9	40965786	112137720	17	2.3	67.7
11	chr11	2778770	10678453	17	2.3	70
12	chr13	87047745	101047648	16	2.2	72.2
14	chr1	16200000	17074942	14	1.9	74.1
15	chr8	92287772	126218574	13	1.8	75.9
17	chr6	129978855	138678863	12	1.6	77.5
18	chr15	66907662	67007662	11	1.5	79
19	chr3	19121	26358509	9	1.2	80.2
20	chr22	23925838	49174577	9	1.2	81.4

Supplementary Table 2. Expression outlier gene-sample pairs and associated SVs. Patient Study Arm: GENUN - unaffected individual with PD mendelian mutation. GENPD - unaffected individual with PD mendelian mutation. PD - idiopathic individual. Brain cis-eQTL: eQTL DEL - SV overlaps cis-eQTL active in the brain.

SV ID	PPMI Patient Study Arm	SV length, bp	SV type	Gene	SV consequence	Brain cis-eQTL
1_cuteSV.DEL.21004	GENUN	31	DEL	CBR1	Non-coding transcript exon variants	
10_cuteSV.DEL.28586	GENPD	6319	DEL	ZNF543	Coding variants	eQTL DEL
11_cuteSV.DEL.48244	PD	261	DEL	GTPBP6	Upstream variants	
21_cuteSV.DEL.26947	PD	15870	DEL	FEM1A	Transcript ablation	eQTL DEL
22_cuteSV.DEL.36656	GENUN	41327	DEL	UBE2B	Transcript ablation	eQTL DEL
23_cuteSV.DUP.4527	PD	553	DUP	EHMT1	Upstream variants	
26_cuteSV.DEL.30202	GENUN	3954	DEL	PLA2G12A	Coding variants	eQTL DEL
26_cuteSV.DEL.35148	GENUN	29630	DEL	FAM120B	Transcript ablation	eQTL DEL
30_cuteSV.DEL.21443	GENPD	12867	DEL	ZNF844	Coding variants	eQTL DEL
40_cuteSV.DEL.7746	GENPD	15571	DEL	UEVLD	Coding variants	eQTL DEL
40_cuteSV.INS.26554	GENPD	75	INS	RAPGEF6	Upstream variants	
				FNIP1	Downstream variants	
41_cuteSV.INS.9204	GENPD	244	INS	LIG4	Intron variants	
51_cuteSV.INS.1834	GENUN	72	INS	LOC101928626	Intron variants	
7_cuteSV.INS.32171	PD	158	INS	EHMT1	Intron variants	
70_cuteSV.INS.38376	PD	168	INS	EHMT1	Downstream variants	
78_cuteSV.DEL.25137	PD	5368	DEL	MPST	Coding variants	eQTL DEL
89_cuteSV.DEL.47022	GENUN	2941	DEL	MAP7D3	Splicing variants	eQTL DEL
9_cuteSV.INS.35791	PD	32	INS	AFF2	Intron variants	