Comprehensive analysis of *GBA* using a novel algorithm for Illumina whole-genome sequence data or targeted Nanopore sequencing

Marco Toffoli*[1], Xiao Chen*[2], Fritz J Sedlazeck[3], Chiao-Yin Lee[1], Stephen Mullin[1,4], Abigail Higgins[1], Sofia Koletsi[1], Monica Emili Garcia-Segura[1], Esther Sammler[5,6], Sonja W. Scholz[7,8], Anthony HV Schapira[1], Michael A. Eberle**[2], Christos Proukakis**[1]

1. Department of Clinical and Movement Neurosciences, Queen Square Institute of Neurology, University College London, WC1N 3BG, United Kingdom
2. Illumina Inc., San Diego, CA, USA
3. Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA.
4. Institute of Translational and Stratified Medicine, University of Plymouth School of Medicine, Plymouth, United Kingdom.
5. MRC Protein Phosphorylation and Ubiquitylation Unit, School of Life Sciences, University of Dundee
6. Molecular and Clinical Medicine, School of Medicine, University of Dundee
7. Neurodegenerative Diseases Research Unit, National Institute of Neurological Disorders and Stroke, Bethesda, MD 20892, USA
8. Department of Neurology, Johns Hopkins University Medical Center, Baltimore, MD 21287, USA

correspondence email(s)

c.proukakis@ucl.ac.uk

meberle@illumina.com

* equal contributions

** corresponding authors, equal contributions

Abstract

GBA variants cause the autosomal recessive Gaucher disease, and carriers are at increased risk of Parkinson's disease (PD) and Lewy body dementia (LBD). The presence of a highly homologous nearby pseudogene (*GBAP1*) predisposes to a range of structural variants arising from either gene conversion or reciprocal recombination, the latter resulting in copy number gains or losses, complicating genetic testing and analysis. To date, short-read sequencing has not been able to fully resolve these or other variants in the key homology region, and targeted long-read sequencing has not previously resolved reciprocal recombinants. We present and validate two independent methods to resolve recombinant alleles and other variants in *GBA*: Gauchian, a novel bioinformatics tool for short-read, whole-genome sequencing data analysis, and Oxford Nanopore long-read sequencing after enrichment with appropriate PCR. The methods were concordant for 42 samples including 30 with a range of recombinants and *GBAP1*-related mutations, and Gauchian outperforms the GATK Best Practices pipeline. Applying Gauchian to Illumina sequencing of over 10,000 individuals from publicly available cohorts shows that copy number variants (CNVs) spanning *GBAP1* are relatively common in Africans. CNV frequencies in PD and LBD are similar to controls, but gains may coexist with other mutations in patients, and a modifying effect cannot be excluded. Gauchian detects a higher frequency of *GBA* variants in LBD than PD, especially severe ones. These findings highlight the importance of accurate *GBA* mutation detection in these patients, which is possible by either Gauchian analysis of short-read whole genome sequencing, or targeted long-read sequencing.

Introduction

The *GBA* gene encodes the lysosomal enzyme glucocerebrosidase, and biallelic mutations in *GBA* cause the autosomal recessive disorder Gaucher disease (GD [MIM: #230800, #230900 and #231000])[1]. Around 500 disease-causing mutations have been reported, mostly missense changes arising from single nucleotide variants (SNVs)[2]. Heterozygous variants in *GBA* [MIM: *606463] are associated with an increased risk of developing Parkinson disease (PD)[3], the second most common neurodegenerative disease, and the closely related Lewy body dementia (LBD)[4]. Identifying *GBA* mutations is difficult due to a pseudogene (*GBAP1*) located 6.9 kb downstream[5] which has an overall homology of 96% with *GBA*. This rises to 98% in the region from intron 8 to the 3'-UTR, where there are five identical segments >200 bp each[6]. The high homology predisposes to non-allelic homologous recombination between *GBA* and *GBAP1*, leading to a wide range of structural variants (SV)[7]. These can be non-reciprocal, also termed gene conversion, or reciprocal, resulting in copy number variants (CNV). Throughout this paper, we use the term copy number gain (CNG) for reciprocal duplication alleles where a 20.6 kb long region of DNA between the homology segments of *GBA* and *GBAP1* is multiplied, and copy number loss (CNL) for reciprocal fusion alleles where the same region is deleted, creating *GBA-GBAP1* fusions[7] (see Figure 1). SVs that disrupt the coding sequence by gene conversion or reciprocal recombination are expected to be pathogenic for GD and risk factors for PD. Conversely, SVs not affecting the coding sequence are not pathogenic, although a modifier effect cannot be excluded[7]. These include CNLs outside the coding region, and all CNGs, which consist of a partial duplication of pseudogene sequence merged with a variable part of the gene, often only the 3' UTR, with the resulting allele still containing a normal copy of the *GBA* coding region (Figure 1D). The SV variability and population prevalence remain largely unknown. Pathogenic missense changes in the high homology exon 9-11 region such as the common p.L483P (NC_000001.11:g.155235252A>G, also known as p.L444P)

may arise by gene conversion, rather than simple base substitutions, with pseudogene sequence incorporated into the gene[7]. We refer to variants corresponding to pseudogene bases in this region as "*GBAP1*-like".

Current sequencing approaches to characterize *GBA* have major pitfalls, and to date no single approach has fully resolved recombinants[6]. The correct alignment of short reads when there is a highly similar pseudogene is intrinsically problematic, and *GBA* is challenging in exome and whole-genome sequencing (WGS) [8–10]. Moreover, the reliability of the standard WGS secondary analysis pipelines such as the Genome Analysis Toolkit best practice workflow[11] has not been formally assessed. Targeted short-read sequencing approaches are also possible but may require forced alignment to *GBA* and visual inspection and Sanger validation to detect recombinant variants, and are not likely to provide copy number information[6,12]. We have already performed refinement of the Illumina WGS analysis for other difficult regions due to sequence homology, demonstrating reliable resolution of SVs in such regions on Illumina WGS data in the *SMN1*[MIM: *600354]/*SMN2*[MIM: *601627] genes in spinal muscular atrophy[13] and the pharmacogene *CYP2D6*[MIM: *124030][14]. We also previously reported a method for *GBA* analysis using enrichment by long-range PCR, followed by sequencing on the Oxford Nanopore Technologies (ONT) MinION[15], which reliably detected SNVs, including *GBAP1*-like variants, and could also detect non-reciprocal recombinants, but not reciprocal recombinants (Figure 1)[15].

To overcome these limitations and improve the characterisation of *GBA* at scale, we have developed refined pipelines based on either targeted analysis of short-read (Illumina) WGS data or targeted long-read (ONT) single molecule sequencing. For Illumina data, we present and validate 'Gauchian', a novel algorithm for *GBA* locus analysis which can reliably resolve SVs and *GBAP1*-like variants. For ONT data, we have addressed the problem of reciprocal recombinants by using PCR primers designed to amplify CNGs and CNLs when they exist. We validated these methods and

then applied them to large PD, LBD, and population control samples. We demonstrate that complete resolution of all variant types in *GBA* is possible using either Gauchian analysis of Illumina WGS data or targeted ONT sequencing. Finally, we confirm that *GBA* variants are more common in LBD than PD, we report the frequency of CNVs in different populations, and suggest that a possible modifier role of CNG in PD and LBD merits further study. Both methods finally enable a precise characterisation of *GBA* at scale, thus driving the identification of causative variants forward.

Subjects and methods

Population cohorts and samples used

We downloaded WGS CRAM files from the 1000 Genomes Project (1kGP). These were generated by 2x150bp reads on Illumina NovaSeq 6000 instruments from PCR-free libraries sequenced to an average depth of at least 30x and aligned to the human reference, hs38DH, using BWA-MEM v0.7.15. We downloaded WGS CRAM files from PD[16] and DLB[17] cohorts and their controls from the AMP-PD knowledge portal. These were generated by sequencing 2x150bp reads to >25x coverage and processing against hs38DH using the Broad Institute's implementation of the Functional Equivalence Pipeline[18]. We also downloaded AMP-PD variant calls, generated using the Broad Institute's Joint Genotyping pipeline (referred to as BWA-GATK in this paper). Where samples had been recorded as European / Caucasian / white, or African / black in the original database, we refer to them as "European" or "African" for consistency and simplicity, despite the lack of scientific validity, in quotation marks as suggested[19].

Selected DNA samples were obtained from the Parkinson's Progression Markers Initiative (PPMI)[20] and the NHGRI Sample Repository for Human Genetic Research at the Coriell Institute for Medical Research. DNA samples from living individuals for ONT analysis were obtained from a clinical cohort, RAPSODI [21], which aims to define the risk of PD in *GBA* mutation carriers. Recruitment and analysis are ongoing, and the clinical results will be reported separately. DNA from saliva (Oragene DNA OG-500 kit, DNA Genotek) was extracted according to the manufacturer protocol. Brain samples from 16 PD patients were obtained from the Queen Square Brain Bank, and DNA extracted with phenol-chloroform[22] or MagAttract HMW DNA kit (Qiagen) from the frontal cortex, cerebellum, or midbrain. Ethics approval was provided by the National Research Ethics Service London—Hampstead Ethics Committee for RAPSODI, NRES Committee central—London for QSBB samples, and UCL Ethics Committee for PPMI samples. All participants provided informed consent.

Gauchian - a WGS-based *GBA* caller

Gauchian builds upon the strategies to solve closely related paralogs, as described in our previously developed *SMN1/2* and *CYP2D6* callers[13,14]. Gauchian calculates the total number of copies of *GBA*, *GBAP1,* and *GBA*/*GBAP1* gene hybrids. Reciprocal recombinations across homologous regions lead to CNG and CNL of the 20.6 kb region between the homologous parts of the two genes. Since the breakpoint may vary in position, to detect CNVs, Gauchian uses the sequencing depth in the 10kb unique region between *GBA* and *GBAP1* (chr1:155220429-155230539; hg38) (Figure 2A). The number of reads aligned to this region is normalized and corrected for GC content, and the copy number is called from a Gaussian mixture model (Figure 2B). A deviation of this copy number (CN) from the diploid expectation indicates the presence of a CNV, e.g. one copy indicates a CNL, and three or more copies indicate a CNG. Thus, this number plus two gives the total copies of *GBA* and *GBAP1* combined, i.e., CN (*GBA*+*GBAP1*). Included in this CN calculation, in addition to *GBA* and *GBAP1* genes, are gene hybrids where part of *GBA* and

GBAP1 are fused.

CNG always leaves an intact copy of *GBA*, while CNL can create pathogenic *GBA-GBAP1* fusions if the deletion breakpoint falls within the *GBA* gene coding region. Next, Gauchian identifies the breakpoint of the CNV, following a similar approach as previously described[14]. To do this, we identified 82 reliable sites (Table S1) that differ between *GBA* and *GBAP1*. Gauchian estimates the *GBA* CN at each *GBA*/*GBAP1* differentiating site based on CN (*GBA*+*GBAP1*) and the numbers of reads supporting *GBA*- and *GBAP1*-specific bases. CNV breakpoints are identified when the CN of *GBA* changes. For example, a transition between CN 1 and CN 2 indicates the breakpoint of a CNL, and a transition between CN 3 and CN 2 indicates the breakpoint of a CNG. The exact breakpoint is further refined by haplotype phasing as described in the next paragraph. To identify recombinant variants, Gauchian analyses the 1.1 kb homology region in exons 9-11 (Figure 2C) containing 10 *GBA/GBAP1* differentiating sites that are 14-315 bp away from each other, several of which are critical *GBAP1-like* variants. These include p.L483P, p.D448H (NC_000001.11:g.155235727C>G), c.1263del55 (NC_000001.11:g.155235752_155235806del), RecNciI (which comprises three SNVs: p.L483P, p.A495P, and p.Val499=), RecTL (RecNciI and p.D448H) and c.1263del+RecTL (RecNciI, p.D448H, and c.1263del55) (Figure 2C). The high homology and the frequent gene conversion between *GBA* and *GBAP1* make exons 9-11 a challenging region for standard secondary analysis pipelines, which often miscall variants due to misalignments of recombinant variant reads. Additionally, three positions in the *GBAP1* reference sequence in hg38 erroneously contain the *GBA* bases (Figure 2C, yellow shading), so *GBA* p.L483P reads would likely align to *GBAP1*, causing false-negative calls (*GBA*/*GBAP1* is among the regions enriched for discordant variant calls between hg19/hg38[23]). In addition, we found *GBAP1* haplotypes that have been partially converted to *GBA*. Those converted bases would direct *GBAP1* reads to align to *GBA*, causing false-positive *GBA* variant calls at nearby positions (Figure 2C, purple shading). Gauchian takes a novel

approach that does not rely on accurate alignments between *GBA* and *GBAP1*. Based on the linking information of reads and read pairs covering the ten differentiating sites in either *GBA* or *GBAP1*, Gauchian phases all the haplotypes at these sites originating from either *GBA* or *GBAP1* and identifies hybrid haplotypes (i.e., a mixture of *GBA* and *GBAP1* bases on the same haplotype). This allows us to identify CNL breakpoints as well as small and big gene conversion events.

To assess the relative abundance of the different haplotypes, Gauchian uses CN (*GBA*+*GBAP1*) and haplotype-supporting read counts at the differentiating bases to call CN of each haplotype. Gauchian compares two scenarios: one copy of the wildtype *GBA* haplotype vs. two copies of the wildtype *GBA* haplotype. Gauchian determines which scenario is more likely given the number of supporting reads in the data. If we call only one copy of the wildtype *GBA* haplotype, this indicates that the individual is a carrier of the disease-causing variant. If an individual is a carrier of more than one variant haplotype and there is no haplotype that carries the *GBA* base at all variant sites of interest, Gauchian calls this sample compound heterozygous. Homozygous variants are called when the CN of the *GBA* base is called 0.

In addition to *GBAP1*-like variants in exons 9-11 homology region, Gauchian targets all known *GBA* pathogenic or likely pathogenic variants as classified by ClinVar (Table S2), including non-*GBAP1*-like variants, and *GBAP1*-like variants outside the exons 9-11 homology region. For these, since variants don't correspond to *GBAP1*, or, if they do, the region between *GBA* and *GBAP1* is not highly similar and alignments are accurate, Gauchian parses read alignments and calls the CN of variants based on the number of variant supporting reads as described for *SMN/CYP2D6* callers. Gauchian is a targeted caller for known variants and thus does not call novel variants.

ONT long-read sequencing with PCR enrichment

GBA enrichment was obtained via PCR (Table S3), with primers previously described[24] (henceforth primer pair 1), modified to carry the ONT barcode adapter sequence. The product was a 8.9 kb

amplicon containing the entire *GBA* coding region and introns (chr1:155232524-155241392; hg38). Samples were barcoded using the 96-sample barcoding kit (EXP- PBC096). Amplicons were purified with Agencourt AMPure XP magnetic beads at a ratio of 0.4x. Library preparation was carried out according to the ONT protocol (version: PBAC96_9069_v109_revO_14Aug2019 – long fragment selection) and sequencing with a MinION device on R9.4 flow-cells.

Primary acquisition of sequencing data was carried out with MinION (version 20.10.3) [25], and base-calling and demultiplexing with Guppy (version 4.2.2). The resulting reads were aligned to GRCh38.p13, without the alternative reference contigs, using NGMLR (version 0.2.7)[26] unless otherwise stated. Clair (version 2.1.1)[27] was used for SNV calling. Since the most recent Clair ONT models were trained with up to 578-fold coverage, each sample was down-sampled to 550-fold. SNV calls were filtered with the Clair genome quality (GQ) score with a threshold set at 650. We only called SNV in *GBA* coding exons and ten flanking bases. Intronic haplotypes for some of these samples with no coding mutations were recently reported separately[28]. Phasing of SNV was carried out with Whatshap (version 1.0)[29] and data manipulation with Samtools (version 1.10)[30], Bedtools (version 2.29.1)[31], and Tabix (version 1.7-2). Optimisation from our previous method[15] comprised Guppy instead of Albacore for base-calling and Clair instead of Nanopolish for SNV calling. The pipeline used here identified all previously reported coding SNVs in 95 samples which were re-analysed.

As homopolymer regions are challenging for ONT[32], and *GBA* has two coding poly-G stretches, we devised a method to detect variants within these (Figure S1). This involves analysis of.bam files with S*amtools depth* to obtain depth of coverage across these *A* (chr1:155239990-155239995 and chr1:155239657-155239661; hg38). The depth of coverage at each position was then adjusted for the depth of coverage at the 100 flanking positions, and the result was compared with the mean of all other samples in the run. If the adjusted depth of coverage at one position was more than

five median absolute deviations from the median adjusted depth of coverage of the other samples in the run at that position, this was considered as evidence of a deletion (coverage lower than the mean for that position) or a SNV (coverage higher than the mean for that position). Variants detected with this method were validated with Sanger sequencing [33].

To detect and amplify reciprocal recombination events, two additional sets of primers were used[24]. These primers were specifically designed to amplify the recombinant alleles: the set *MTX1-r/GBA-nf* (henceforth referred to as primer pair 2) only amplifies recombinants with *GBA* gene sequence at the 5'UTR end and *GBAP1* sequence at the 3'-UTR end (CNL), while the set of primers *ΨMTX1-r/ΨGBA-nf* (primer pair 3) only amplifies recombinants with the 3'-UTR end and *GBAP1* sequence at the 5'-UTR end (CNG, see Figure 1). Samples underwent PCR using these pairs of primers. If a product was detected on agarose electrophoresis, the sample was re-amplified with the same primer pair modified to carry the ONT barcode adapters, and the amplicons were barcoded and sequenced as described. The primer sequences and PCR conditions are given in Table S4. To define the breakpoint of CNL, the products of primer pair 2 were aligned to *GBA* (to avoid alignment to *GBAP1;* chr1:155222384-155241249; hg38) using LAST (Version 1243)[34]. The resulting alignment was analysed with Clair to look for variants at positions where *GBA* and *GBAP1* differ (Table S1). If a sample displayed an SNV in a certain position, it meant that the breakpoint must be upstream of that but downstream of the next sentinel position where no variant is detected (Figure S2).

GBA enrichment with UNCALLED

To validate *GBA* SV without PCR enrichment, we used UNCALLED, which uses adaptive sampling to enable real-time enrichment or depletion on MinION runs via the MinKNOW API ReadUntil[35]. UNCALLED analyses the signal generated by the DNA molecules passing through each pore of the device in real time and decides whether they align to a reference sequence provided. It can then

prematurely eject the molecule from the pore if not of interest, freeing up sequencing capacity for new reads and ultimately achieving purely computational enrichment. The target region for enrichment was chr1:155193567-155264811, with repetitive regions masked with the UCSC RepeatMasker.

Digital PCR

Digital PCR (dPCR) was performed by QIAGEN (Hilden, Germany) on the QIAcuity instrument. Three probes were selected: DCH101-0776005A (chr1:155231010-155231209; hg38) and DCH101-0776012A (chr1:155232410-155232609; hg38) target the region affected by the recombination event, while DCH101-1260927A (chr1:155208699-155208804; hg38) is outside of this region and was used as a reference for analysis. Each sample was tested three times, and the result is the average of the three assays.

Illumina sequencing

WGS was performed on the Illumina NovaSeq instruments using Illumina TruSeq Nano DNA Library Prep[9].

Statistical analysis

Analysis was carried out on R (version 4.0.5). Odds Ratios (OR) were calculated with logistic regression. To check for an additive effect on risk of CNGs on *GBA* variants carriers, a multivariate logistic regression was used, with disease status as the outcome variable, and *GBA*-carrier status and CNG-carrier status as independent variables.

Results

Cross-validation confirms both Gauchian and ONT methods

To select appropriate samples for validation of Gauchian with a broad range of mutations, we first obtained Gauchian results on 1kGP samples and the AMP-PD PPMI PD and control cohorts. We selected 37 of these for validation by ONT targeted sequencing. These included 15 samples from 1kGP with CNVs or gene conversions, and 22 samples from PPMI, where Gauchian showed CNVs or *GBAP1*-related variants (n=7), or was discordant with available BWA-GATK results (n=4), or no mutation was reported by Gauchian or BWA-GATK (n=11). Additionally, for 5 brain DNA samples analysed first by ONT with recombinations or *GBAP1*-related mutations, we performed Illumina WGS and Gauchian analysis. All 42 Gauchian results were consistent with ONT. Within these validation samples, Gauchian reported 5 CNL, which included one in which the p.L483P was also found, and one resulting in the pathogenic RecNciI, and 14 CNG's, including two samples which also carried a gene conversion, and one which carried p.L483P. Additionally, in 11 samples Gauchian called *GBAP1*-like variant calls within *GBA*, including two gene conversions. The remaining 12 samples were wild type calls. Notably, Gauchian and ONT gave concordant results in two samples where previous BWA-GATK analysis had missed p.L483P, and two where BWA-GATK had wrongly called p.A495P (NC_000001.11:155235216:C:G). Results of cross-validation of the two methods are reported in Table 1.

To obtain further orthogonal validation, we used dPCR for copy number estimation of the 20.6 kb region involved in recombination in six samples with a range of copy numbers. These included four samples where Gauchian and ONT both detected a CNG (additional copy numbers 1, 3, 5, and 6), and two where we only had ONT data, one with a CNL, and one with no CNV. The results were fully concordant (Table S5). Finally, we applied ONT sequencing with PCR-free enrichment by

adaptive sampling (UNCALLED)[35] to four reciprocal recombinants, two CNG and two CNL (one pathogenic and one non-pathogenic). Inspection of the resulting alignments confirmed the presence of the SV and the breakpoints of CNL alleles (Figure S3).

Detection of all classes of *GBA* variants with targeted ONT long-read sequencing

We analysed 397 samples from PD or GD patients, their relatives, and controls (Table S6) using all three pairs of PCR primers, followed by ONT amplicon sequencing. These included 95 individuals previously sequenced with ONT using only primer pair 1[15]. All results are shown in Table S7. We detected two c.1263del+RecTL alleles and one RecNciI allele arising from gene conversion. Additionally, we also detected 94 coding or splice site SNVs, including the pathogenic mutations p.N409S (NC_000001.11:g.155235843T>G, also known as p.N370S; 38% of all SNV detected) and the *GBAP1*-like p.L483P (20%), and the PD risk alleles p.E365K (NC_000001.11:g.155236376C>T, also known as p.E326K, 16%) and p.T408M (NC_000001.11:g.155236246G>A, also known as p.T369M, 4%). Notably, we also detected c.84dupG (NC_000001.11:g.155240661dup), the most common pathogenic indel in the *GBA* gene. Additional homopolymer analysis identified one single base deletion and one SNV within homopolymers that would have been missed by our old ONT pipeline (c.413delC and p.P68=, NC_000001.11:g.155239661del and NC_000001.11:g.155239989C>T, Figure S1). We detected CNLs using primer pair 2 in nine samples. According to the position of the breakpoints (Figure S2), five of them were pathogenic, and four were not. Two non-pathogenic CNLs were *in cis* with p.L483P, a pattern already described[7]. We also detected a CNG using primer pair 3 in seven samples, four of which also carried a *GBAP1*-like variant (two c.1263del+RecTL, two p.L483P).

Comprehensive *GBA* analysis by Gauchian in short-read WGS population data

A total of 10623 samples were analysed with Gauchian, including 2504 samples from the 1kGP cohort, 2325 PD and 2598 LBD samples from the AMP-PD knowledge portal, and their respective controls. We identified 55 non-pathogenic CNLs and 146 CNGs (roughly correspond to DGV variant accessions dgv55e214 and esv3587619; Table 2). Additionally, we detected 97 *GBAP1*-like variants (including those generated by pathogenic CNL or gene conversion) in the exons 9-11 homology region of *GBA* in all three cohorts (Table 3 and Table S8).

BWA-GATK variant calls were available for all from AMP-PD samples analysed by Gauchian. For all PD and LBD case/control populations, BWA-GATK called 44 *GBAP1*-like variants, and Gauchian called 86, almost doubling the variant calls. Due to the sequence homology and misalignment of reads in exons 9-11, the BWA-GATK pipeline under-called all *GBAP1*-like variants except p.A495P and p.D448H. For p.A495P, GATK called 11 false positives (including two confirmed as false positives by ONT amplicon sequencing- see earlier) and for p.D448H BWA-GATK called two false positives. The false-positive calls by BWA-GATK are due to alignment errors caused by *GBAP1* haplotypes containing *GBA* bases (see Figure 2C). Gauchian also detected other coding SNVs and indels that are not *GBAP1*-like in the three cohorts (see Table S9 for all variants). All these calls were concordant with BWA-GATK except in one sample where Gauchian called p.L483R, a rare pathogenic variant in the same codon as the common *GBAP1*-like p.L483P, but BWA-GATK did not. This variant is in the exon 9-11 homology region, and variant reads misaligned to *GBAP1*, causing the false-negative by BWA-GATK (Figure S4).

Prevalence of *GBA* recombinant and non-recombinant variants in healthy, PD, and LBD populations

Gauchian allowed us to provide the first large-scale analysis of all classes of *GBA* mutations in healthy, PD, and LBD populations. Non-pathogenic CNVs, where the breakpoints do not alter the *GBA* coding region, were ten times more frequent in "Africans" than "Europeans" (11.3% vs 1.1%, Table 2). This was primarily driven by a striking difference in the prevalence of CNGs (10.8% v 0.6% for controls from both cohorts; p<2.2e-16). Additionally, "Africans" also had more copies gained, with a median gain of three copies compared to one for "Europeans".

As non-pathogenic CNVs in *GBA* have not previously been considered as possible PD or LBD risk factors, we compared these across the combined disease cohorts to their controls. We detected no difference in the prevalence of all non-pathogenic CNVs (1.10% vs 1.25%), CNGs (0.67% vs 0.63%) or CNLs (0.43% vs 0.63%, Table 2). Addressing SNVs next, we noticed that p.N409S was found at a very high frequency in the PD cohort of AMP-PD (in both cases and controls, 5.5% and 12.6% respectively), because of the recruitment of a large number of individuals with Ashkenazi Jewish ancestry[16], where it is very common. After excluding individuals carrying it from both cohorts for consistency, *GBA* variants were more common in each disease cohort than the respective controls as expected (Table 4) (PD 7.8% v 3.9%; LBD 11.7% v 3.5%). This was also true for severe *GBA* variants[36] (PD 1.7% vs 0.8%; LBD 3.1% vs 0.1%). The overall OR for mutations in each disease against its controls was higher in LBD than PD (3.68 v 2.07; p=0.0098), and this was even more striking for the severe mutations (30.83 v 2.12; p=0.0009).

We also noted that some individuals carried both a CNG and another *GBA* variant, mostly a *GBAP1*-like variant in the exon 9-11 homology region (Table 2), as also seen in the cohort analysed by ONT. There were no individuals with a CNG and another *GBA* variant in the 1kGP cohort.

Considering all samples analysed by Gauchian, 7 out of 146 with a CNG also had a *GBAP1*-like variant, against 71 of 10,407 without a CNG (4.8% v 0.68%; p-value=9.77e-5). Three additional individuals carried a non-*GBAP1*-like variant and a CNG. In the PD and LBD cohorts and their controls, 9 of 10 individuals carrying a CNG as well as a coding variant in *GBA* were patients (four PD, five LBD). One healthy control carried a CNG and p.T408M, which is a mild PD risk allele but does not cause GD. As we did not find any healthy controls with a CNG and a pathogenic variant, we considered whether the combination of both is more detrimental than a coding variant alone (excluding again p.N409S carriers, one of whom who also had a CNG). We did not detect a significant added risk for disease in the combined PD and LBD AMP-PD data against their controls (OR for CNG and other variant *vs* other variant alone 2.31, 95% CI 0.37-45.01).

Discussion

The recent dramatic improvements in sequencing techniques have allowed a much better understanding of human genetic variation, but several regions, including some key disease-related genes, have remained challenging. One example is *GBA*, responsible for the autosomal recessive lysosomal storage disorder GD[1], and one of the most important genetic determinants of risk for PD and the closely related LBD[17]. Here we present and validate Gauchian, a novel *GBA* caller for Illumina WGS data, capable of detecting SVs and SNVs within *GBA*. Using ONT targeted sequencing, we demonstrate that in the cases of discrepant calls between Gauchian and BWA-GATK analysis, the Gauchian calls are correct. We also demonstrate that a refined ONT amplicon-sequencing pipeline can detect reciprocal recombinants, and indels as well as mutations within homopolymers in coding exons. Importantly, both methods detect CNGs and CNLs arising from reciprocal recombination, and allow straightforward classification of CNLs into those that do and do not affect the coding region, previously a complex task[8,37]. We thus provide two complementary new tools for fully resolving the *GBA* gene, which will be helpful to the community. Illumina WGS data can now be analysed robustly, and ONT targeted sequencing can be applied in a cost-effective way where an analysis of *GBA* is sufficient.

To explore the potential of Gauchian in the population and in disease contexts, we applied it to a total of 10,623 samples from the 1000 Genomes Project and PD and LBD cohorts with their controls from the AMP-PD initiative. This allowed us to provide the first large scale data on CNVs , and to evaluate the frequency of all classes of *GBA* variants in PD and LBD with greater accuracy than before. Reciprocal recombinants in particular are likely missed in PD studies[15], including a recent targeted short-read study which detected none in 3402 patients[8] , although one study using exome data with qPCR validation reported CNGs in 1.2% of PD and 0.7% of controls[37]. In non-diseased individuals, we noted that CNGs were more common in those with “African”

ancestry, with greater copy number variability. These results are consistent with the greater African genetic diversity, with recent evidence of "African" genomes demonstrating unexplored structural variation[38] and more variability in copy numbers of *SMN1*, *SMN2* and *CYP2D6*[13,14]. They further highlight the need to study non-European genomes , which has yielded additional insights into Alzheimer's disease[39], and is being expanded in PD[40].

In the PD and LBD cohorts, Gauchian analysis almost doubled the pathogenic variants detected in the homology region compared to BWA-GATK (86 vs 44) and eliminated false positives. We also performed a direct comparison of *GBA* variant frequencies between PD and LBD, after excluding the common p.N409S variant due to selection bias in the PD cohort[16]. The prevalence of *GBA* pathogenic or PD-risk variants was significantly higher in LBD than PD (11.7% vs 7.8%), and this difference was even larger for severe pathogenic variants (3.1% v 1.7%). The OR for *GBA* mutations in LBD compared to controls was higher in our analysis than in the original report in this cohort[17] (3.68 v 2.90), and a previous study (2.55)[4], due to the detection of additional mutations and the filtering of p.N409S. *GBA* mutations increase the risk of cognitive decline in PD[41], and the odds ratio for *GBA* variants is higher in LBD than PD with dementia[42]. Severe *GBA* variants in particular, which cause the neuronopathic form of GD[1], have a higher risk of PD [43] and a faster cognitive decline in PD than mild variants[44]. If PD and LBD are considered as a spectrum of phenotypes with variable cognitive involvement, our findings further suggest that *GBA* variants, especially severe ones, tend to predispose to a phenotype on the LBD end of the spectrum. The main limitation of this analysis is the use of LBD and PD cohorts recruited separately, with the selection process necessitating the exclusion of p.N409S, and further comparisons in unselected matched cohorts are needed.

The variable penetrance and phenotypic heterogeneity of PD and LBD patients with *GBA*

mutations is attracting a lot of attention, with lysosomal gene variants acting as genetic modifiers[45]. An effect of common intronic haplotypes was suggested[46], but not seen by us in AMP-PD cohort and part of the RAPSODI cohort used here[28]. A possible influence of CNGs has not yet been investigated. Although these do not alter the *GBA* coding region, they could affect expression and function, for example by acting as a competing endogenous microRNA sponge[47]. CNGs were not enriched in PD or LBD. There were, however, rare carriers of both a CNG and a pathogenic or PD-associated *GBA* variant. In the PD and LBD cohorts, nine out of ten of these were patients, and in the only control the coding variant was the mild PD risk allele p.T408M. This raises the possibility that CNGs are modifiers, increasing the penetrance of other *GBA* variants. We have not, however, phased the CNGs and other variants, and did not show a statistically significant increased risk for carriers of a CNG and mutation compared to mutation alone. Therefore, further population and mechanistic work is required.

In conclusion, we have demonstrated that SNV detection and complete resolution of all classes of SVs is possible using the novel Gauchian caller with Illumina WGS, which outperforms BWA-GATK analysis, or with targeted ONT sequencing. We also demonstrate that CNVs are relatively common, and suggest that these merit investigation as possible modifiers of PD or LBD risk. Given the importance of this gene, and the rapid progress to targeted clinical trials in PD[48], we propose that the adoption of either workflow should be considered by research and diagnostic labs, based on local resource and data availability.

Description of supplemental data

Supplemental data include 4 figures and 9 tables.

Declaration of interests

XC and MAE are employees of Illumina Inc.

S.W.S. serves on the Scientific Advisory Council of the Lewy Body Dementia Association. S.W.S. is an editorial board member for the Journal of Parkinson's Disease and JAMA Neurology.

AHVS is supported by the UCLH NIHR BRC.

This study was supported in part by the Intramural Research Program of the National Institutes of Health (National Institute of Neurological Disorders and Stroke; project numbers: 1ZIANS003154) and the JPND through the MRC grant code MR/T046007/1.

Acknowledgments

We thank the New York Genome Center and the Coriell Institute for Medical Research for generating and releasing the 1kGP WGS data. We thank the AMP PD Knowledge Platform for hosting WGS data for patient and control cohorts.

We thank Salomé Bagayan, Diem Cao, Angela Henry, Wen Liang, Leland Mencik, Lisa Robison, Christina Toma and Sasha Treadup at Illumina for helping with the sequencing of validation samples, and Iska Steffens at QIAGEN for performing digital PCR.

Data and biospecimens used in the analyses presented in this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) (www.ppmi-info.org/specimens). As such, the investigators within Revised April 2019 PPMI contributed to the design and implementation of PPMI and/or provided data and collected biospecimens, but did not participate in the analysis or writing of this report. For up-to- date information on the study, visit www.ppmi-info.org.

PPMI – a public-private partnership – is funded by The Michael J. Fox Foundation for Parkinson's Research and funding partners, including [list the full names of all PPMI funding partners found at www.ppmi-info.org/fundingpartners]

Data used in the preparation of this article were obtained from the AMP PD Knowledge Platform. For up-to-date information on the study, https://www.amp-pd.org.
AMP PD – a public-private partnership – is managed by the FNIH and funded by Celgene, GSK, the Michael J. Fox Foundation for Parkinson's Research, the National Institute of Neurological Disorders and Stroke, Pfizer, Sanofi, and Verily.
Clinical data and biosamples used in preparation of this article were obtained from the Fox Investigation for New Discovery of Biomarkers (BioFIND), the Harvard Biomarker Study (HBS), the Parkinson's Progression Markers Initiative (PPMI), and the Parkinson's Disease Biomarkers Program (PDBP).
BioFIND is sponsored by The Michael J. Fox Foundation for Parkinson's Research (MJFF) with support from the National Institute for Neurological Disorders and Stroke (NINDS). The BioFIND Investigators have not participated in reviewing the data analysis or content of the manuscript. For up-to-date information on the study, visit michaeljfox.org/biofind.
Harvard NeuroDiscovery Biomarker Study (HBS) is a collaboration of HBS investigators [full list of HBS investigator found at https://www.bwhparkinsoncenter.org/biobank]and funded through philanthropy and NIH and Non-NIH funding sources. The HBS Investigators have not participated in reviewing the data analysis or content of the manuscript.
Parkinson's Disease Biomarker Program (PDBP) consortium is supported by the National Institute of Neurological Disorders and Stroke (NINDS) at the National Institutes of Health. A full list of

PDBP investigators can be found at https://pdbp.ninds.nih.gov/policy. The PDBP Investigators have not participated in reviewing the data analysis or content of the manuscript.

We are grateful to the Queen Square Brain Bank, and to individuals who donated their brains. The Queen Square Brain Bank is supported by the Reta Lila Weston Institute for Neurological Studies and the Medical Research Council UK.

Web resources

1k GP project https://www.ncbi.nlm.nih.gov/bioproject/PRJEB31736

RAPSODI, https://rapsodistudy.com

OMIM, http://www.omim.org/

The NCBI reference sequence for *GBA* on which the numbering of exons is based is NM_000157.4.

Data and code availability

Gauchian will be a part of Version 3.10 of the Illumina DRAGEN (Dynamic Read Analysis for GENomics) Bio-IT platform.

ONT and UNCALLED scripts used will be downloadable at https://github.com/marcotoffoli.

Individual-level genome sequence data for the PD patients, LBD patients, and neurologically healthy controls are available at AMP-PD (https://amp-pd.org).

The datasets of DNA from QSBB brain samples and NHGRI samples generated during this study (Illumina WGS and targeted ONT sequencing) will be made available on the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home), ascession number PRJEB48317. The datasets only include read alignments to *GBA/GBAP1* regions (other regions of the genome have been

removed or masked) to minimize the amount of genetic information made available for public access.

The datasets of DNA from PPMI samples generated during this study (targeted ONT sequencing) will be made available on the PPMI repository (https://www.ppmi-info.org/).

ONT sequencing data on living individuals are not available due to consent / IRB restrictions.

References

1. Hruska, K.S., LaMarca, M.E., Scott, C.R., and Sidransky, E. (2008). Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA). Human Mutation *29*, 567–583.
2. Do, J., McKinney, C., Sharma, P., and Sidransky, E. (2019). Glucocerebrosidase and its relevance to Parkinson disease. Molecular Neurodegeneration *14*, 36.
3. Sidransky, E., Nalls, M.A.A., Aasly, J.O.O., Aharon-Peretz, J., Annesi, G., Barbosa, E.R.R., Bar-Shira, A., Berg, D., Bras, J., Brice, A., et al. (2009). Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. The New England Journal of Medicine *361*, 1651–1661.
4. Guerreiro, R., Ross, O.A., Kun-Rodrigues, C., Hernandez, D.G., Orme, T., Eicher, J.D., Shepherd, C.E., Parkkinen, L., Darwent, L., Heckman, M.G., et al. (2018). Investigating the genetic architecture of dementia with Lewy bodies: a two-stage genome-wide association study. Lancet Neurol *17*, 64–74.
5. Horowitz, M., Wilder, S., Horowitz, Z., Reiner, O., Gelbart, T., and Beutler, E. (1989). The human glucocerebrosidase gene and pseudogene: structure and evolution. Genomics *4*, 87–96.
6. Zampieri, S., Cattarossi, S., Bembi, B., and Dardis, A. (2017). GBA Analysis in Next-Generation Era: Pitfalls, Challenges, and Possible Solutions. Journal of Molecular Diagnostics *19*, 733–741.
7. Tayebi, N., Stubblefield, B.K., Park, J.K., Orvisky, E., Walker, J.M., LaMarca, M.E., and Sidransky, E. (2003). Reciprocal and Nonreciprocal Recombination at the Glucocerebrosidase Gene Region: Implications for Complexity in Gaucher Disease. The American Journal of Human Genetics *72*, 519–534.
8. Woo, E.G., Tayebi, N., and Sidransky, E. (2021). Next-Generation Sequencing Analysis of GBA1: The Challenge of Detecting Complex Recombinant Alleles. Front Genet *12*, 684067.
9. Auwera, G.A.V. der, Carneiro, M.O., Hartl, C., Poplin, R., Angel, G. del, Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics *43*, 11.10.1-11.10.33.
10. Bodian, D.L., Klein, E., Iyer, R.K., Wong, W.S.W., Kothiyal, P., Stauffer, D., Huddleston, K.C., Gaither, A.D., Remsburg, I., Khromykh, A., et al. (2016). Utility of whole-genome sequencing for detection of newborn screening disorders in a population cohort of 1,696 neonates. Genet Med *18*, 221–230.
11. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv:1303.3997 [q-Bio].
12. Heijer, J.M. den, Cullen, V.C., Quadri, M., Schmitz, A., Hilt, D.C., Lansbury, P., Berendse, H.W., Berg, W.D.J. van de, Bie, R.M.A. de, Boertien, J.M., et al. (2020). A Large-Scale Full GBA1 Gene Screening in Parkinson's Disease in the Netherlands. Movement Disorders *35*, 1667–1674.
13. Chen, X., Sanchis-Juan, A., French, C.E., Connell, A.J., Delon, I., Kingsbury, Z., Chawla, A., Halpern, A.L., Taft, R.J., Bentley, D.R., et al. (2020). Spinal muscular atrophy diagnosis and carrier

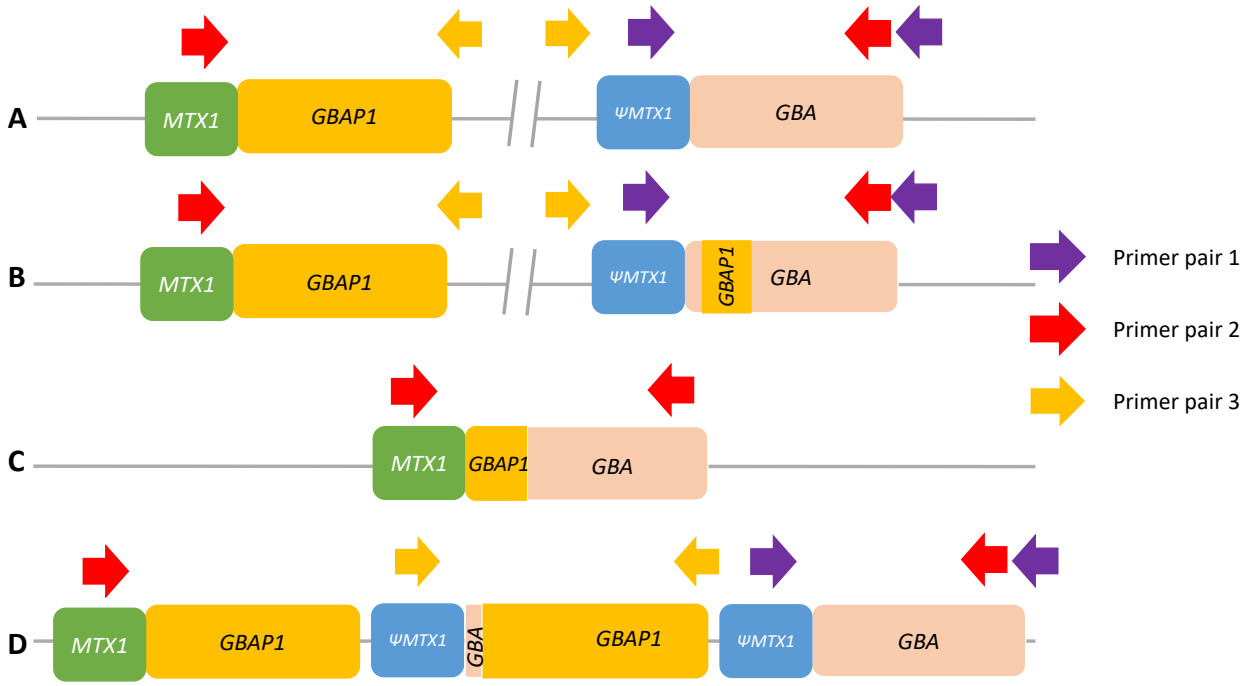
screening from genome sequencing data. Genet Med *22*, 945–953.
14. Chen, X., Shen, F., Gonzaludo, N., Malhotra, A., Rogert, C., Taft, R.J., Bentley, D.R., and Eberle, M.A. (2021). Cyrius: accurate CYP2D6 genotyping using whole-genome sequencing data. Pharmacogenomics J *21*, 251–261.
15. Leija-Salazar, M., Sedlazeck, F.J.F.J., Toffoli, M., Mullin, S., Mokretar, K., Athanasopoulou, M., Donald, A., Sharma, R., Hughes, D., Schapira, A.H.V.A.H.V.V., et al. (2019). Evaluation of the detection of GBA missense mutations and other variants using the Oxford Nanopore MinION. Molecular Genetics & Genomic Medicine *7*, e564.
16. Iwaki, H., Leonard, H.L., Makarious, M.B., Bookman, M., Landin, B., Vismer, D., Casey, B., Gibbs, J.R., Hernandez, D.G., Blauwendraat, C., et al. (2021). Accelerating Medicines Partnership: Parkinson's Disease. Genetic Resource. Movement Disorders *36*, 1795–1804.
17. Chia, R., Sabir, M.S., Bandres-Ciga, S., Saez-Atienzar, S., Reynolds, R.H., Gustavsson, E., Walton, R.L., Ahmed, S., Viollet, C., Ding, J., et al. (2021). Genome sequencing analysis identifies new loci associated with Lewy body dementia and provides insights into its genetic architecture. Nat Genet *53*, 294–303.
18. Regier, A.A., Farjoun, Y., Larson, D.E., Krasheninina, O., Kang, H.M., Howrigan, D.P., Chen, B.-J., Kher, M., Banks, E., Ames, D.C., et al. (2018). Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat Commun *9*, 4038.
19. Birney, E., Inouye, M., Raff, J., Rutherford, A., and Scally, A. (2021). The language of race, ethnicity, and ancestry in human genetic research. ArXiv:2106.10041 [q-Bio].
20. Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., Coffey, C., Kieburtz, K., Flagg, E., Chowdhury, S., et al. (2011). The Parkinson Progression Marker Initiative (PPMI). Progress in Neurobiology *95*, 629–635.
21. Higgins, A.L., Toffoli, M., Mullin, S., Lee, C.-Y., Koletsi, S., Avenali, M., Blandini, F., and Schapira, A.H.V. (2021). The Remote Assessment of Parkinsonism Supporting Ongoing Development of Interventions in Gaucher Disease – Study Protocol. MedRxiv 2021.07.21.21260533.
22. Nacheva, E., Mokretar, K., Soenmez, A., Pittman, A.M., Grace, C., Valli, R., Ejaz, A., Vattathil, S., Maserati, E., Houlden, H., et al. (2017). DNA isolation protocol effects on nuclear DNA analysis by microarrays, droplet digital PCR, and whole genome sequencing, and on mitochondrial DNA copy number estimation. PLOS ONE *12*, e0180467.
23. Li, H., Dawood, M., Khayat, M.M., Farek, J.R., Jhangiani, S.N., Khan, Z.M., Mitani, T., Coban-Akdemir, Z., Lupski, J.R., Venner, E., et al. (2021). Exome variant discrepancies due to reference-genome differences. Am J Hum Genet *108*, 1239–1250.
24. Jeong, S.-Y., Kim, S.-J., Yang, J.-A., Hong, J.-H., Lee, S.-J., and Kim, H.J. (2011). Identification of a novel recombinant mutation in Korean patients with Gaucher disease using a long-range PCR approach. Journal of Human Genetics *56*, 469–471.
25. Jain, M., Olsen, H.E., Paten, B., and Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biology *17*, 239.
26. Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., and Schatz, M.C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. Nature Methods *15*, 461–468.
27. Luo, R., Wong, C.-L., Wong, Y.-S., Tang, C.-I., Liu, C.-M., Leung, C.-M., and Lam, T.-W. (2020). Exploring the limit of using a deep neural network on pileup data for germline variant calling. Nature Machine Intelligence *2*, 220–227.
28. Toffoli, M., Higgins, A., Lee, C., Koletsi, S., Chen, X., Eberle, M., Sedlazeck, F.J., Mullin, S., Proukakis, C., and Schapira, A.H.V. (2021). Intronic Haplotypes in the GBA Gene Do Not Predict Age at Diagnosis of Parkinson's Disease. Movement Disorders *36*, 1456–1460.

29. Martin, M., Patterson, M., Garg, S., O Fischer, S., Pisanti, N., Klau, G., Schöenhuth, A., and Marschall, T. (2016). WhatsHap: fast and accurate read-based phasing. BioRxiv 085050.
30. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
31. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.
32. Huang, Y.-T., Liu, P.-Y., and Shih, P.-W. (2021). Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing. Genome Biology *22*, 95.
33. Stone, D.L., Tayebi, N., Orvisky, E., Stubblefield, B., Madike, V., and Sidransky, E. (2000). Glucocerebrosidase gene mutations in patients with type 2 Gaucher disease. Human Mutation *15*, 181–188.
34. Martin Frith / last.
35. Kovaka, S., Fan, Y., Ni, B., Timp, W., and Schatz, M.C. (2020). Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. Nature Biotechnology 1–11.
36. Beutler, E., Gelbart, T., and Scott, C.R. (2005). Hematologically important mutations: Gaucher disease. Blood Cells, Molecules, and Diseases *35*, 355–364.
37. Spataro, N., Roca-Umbert, A., Cervera-Carles, L., Vallès, M., Anglada, R., Pagonabarraga, J., Pascual-Sedano, B., Campolongo, A., Kulisevsky, J., Casals, F., et al. (2017). Detection of genomic rearrangements from targeted resequencing data in Parkinson's disease patients. Movement Disorders : Official Journal of the Movement Disorder Society *32*, 165–169.
38. Ebert, P., Audano, P.A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M.J., Sulovari, A., Ebler, J., Zhou, W., Mari, R.S., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science *372*,.
39. Dehghani, N., Bras, J., and Guerreiro, R. (2021). How understudied populations have contributed to our understanding of Alzheimer's disease genetics. Brain *144*, 1067–1081.
40. Program, T.G.P.G. (2021). GP2: The Global Parkinson's Genetics Program. Movement Disorders *36*, 842–851.
41. Alcalay, R.N., Caccappolo, E., Mejia-Santana, H., Tang, M.X., Rosado, L., Orbe Reilly, M., Ruiz, D., Ross, B., Verbitsky, M., Kisselev, S., et al. (2012). Cognitive performance of GBA mutation carriers with early-onset PD: The CORE-PD study. Neurology *78*, 1434–1440.
42. Nalls, M.A., Duran, R., Lopez, G., Kurzawa-Akanbi, M., McKeith, I.G., Chinnery, P.F., Morris, C.M., Theuns, J., Crosiers, D., Cras, P., et al. (2013). A Multicenter Study of Glucocerebrosidase Mutations in Dementia With Lewy Bodies. JAMA Neurology *70*, 727–735.
43. Gan-Or, Z., Amshalom, I., Kilarski, L.L., Bar-Shira, A., Gana-Weisz, M., Mirelman, A., Marder, K., Bressman, S., Giladi, N., and Orr-Urtreger, A. (2015). Differential effects of severe vs mild GBA mutations on Parkinson disease. Neurology *84*, 880–887.
44. Liu, G., Boot, B., Locascio, J.J., Jansen, I.E., Winder-Rhodes, S., Eberly, S., Elbaz, A., Brice, A., Ravina, B., van Hilten, J.J., et al. (2016). Specifically neuropathic Gaucher's mutations accelerate cognitive decline in Parkinson's. Annals of Neurology *80*, 674–685.
45. Blauwendraat, C., Reed, X., Krohn, L., Heilbron, K., Bandres-Ciga, S., Tan, M., Gibbs, J.R., Hernandez, D.G., Kumaran, R., Langston, R., et al. (2020). Genetic modifiers of risk and age at onset in GBA associated Parkinson's disease and Lewy body dementia. Brain *143*, 234–248.
46. Schierding, W., Farrow, S., Fadason, T., Graham, O., Pitcher, T., Qubisi, S., Davidson, A.J., Perry, J.K., Anderson, T., Kennedy, M., et al. (2020). Common variants co-regulate expression of GBA and modifier genes to delay Parkinson's disease onset. Movement Disorders mds.28144.
47. Thomson, D.W., and Dinger, M.E. (2016). Endogenous microRNA sponges: evidence and controversy. Nat Rev Genet *17*, 272–283.

48. Mullin, S., Smith, L., Lee, K., D'Souza, G., Woodgate, P., Elflein, J., Hällqvist, J., Toffoli, M., Streeter, A., Hosking, J., et al. (2020). Ambroxol for the Treatment of Patients With Parkinson Disease With and Without Glucocerebrosidase Gene Mutations. JAMA Neurology *77*, 427.

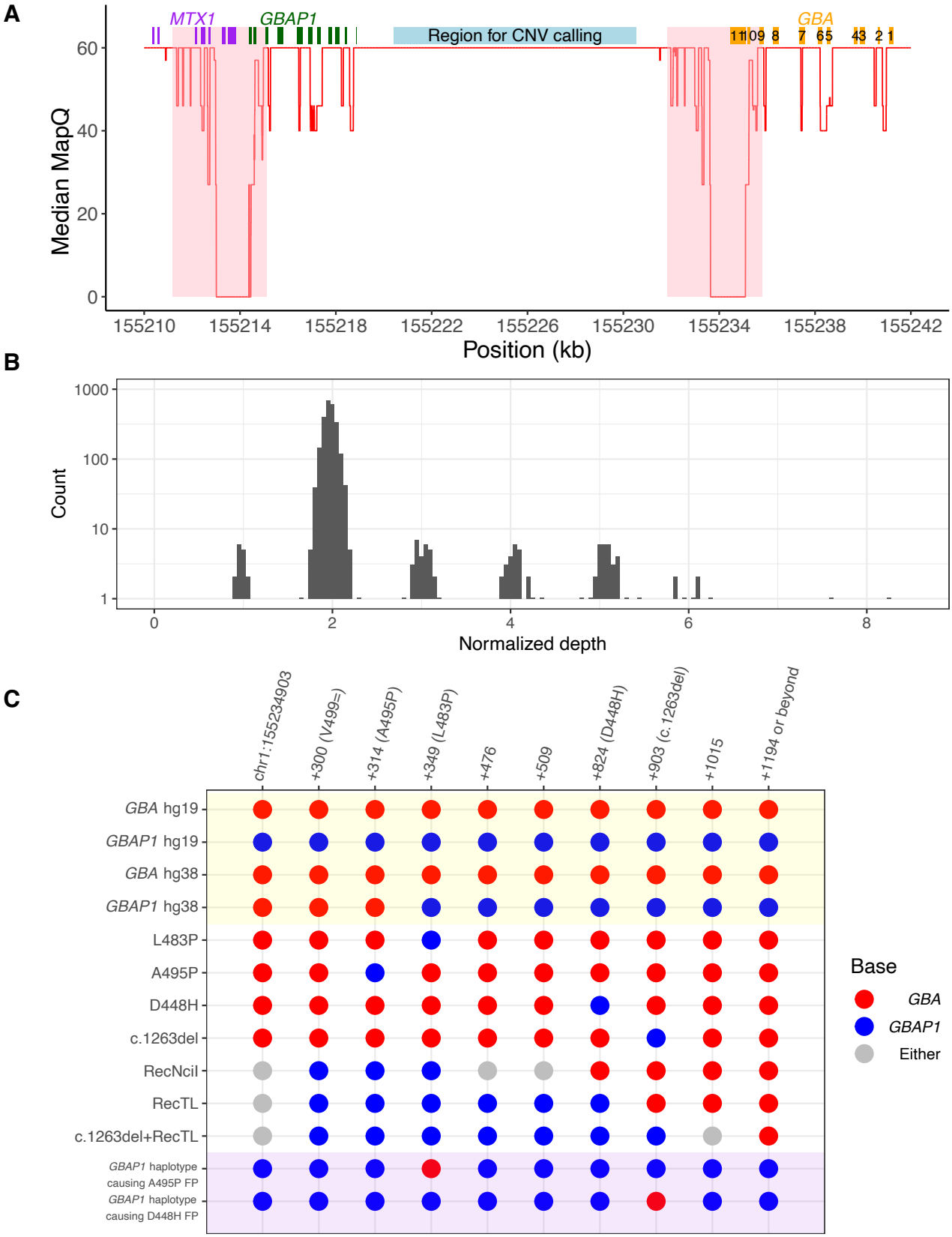
Figure titles and legends

Figure 1: Schematic illustration of the different types of *GBA* recombinant alleles and positions of PCR primers used to detect them with ONT.



Not to scale. **A.** Wild-type allele. Only primer pair 1 will produce an amplicon. **B.** Non-reciprocal recombination (gene conversion). Similar to non-recombinant alleles, only primer pair 1 will produce an amplicon. **C.** Reciprocal crossover between gene and pseudogene resulting in a 20.6 kb deletion (CNL). Only primer pair 2 will produce an amplicon. **D.** Reciprocal crossover between gene and pseudogene resulting in a 20.6 kb duplication (CNG). Both primer pair 1 and primer pair 3 will produce amplicons. Note that the normal allele is present and that amplification with primer pair 3 will produce an amplicon independently of the number of copy number gains.

Figure 2: Gauchian detects challenging *GBA* variants through targeted copy number calling and haplotype phasing.



A. Median mapping quality (red line) across 2504 1kGP samples plotted for each position in the *GBA*/*GBAP1* region (hg38). A median filter is applied in a 50 bp window. The eleven exons of *GBA* are shown as orange boxes. *GBAP1* and *MTX1* exons are shown as green and purple boxes, respectively. The 4kb major homology region (98.1% sequence similarity, exons 9-11) between *GBA* and *GBAP1* is shaded in pink and highlights an area of low mapping accuracy. The light blue box shows the 10kb unique region between the two genes in which copy number calling is performed in Gauchian. **B.** Distribution of normalized depth in the 10 kb CN calling region in 2504 1kGP samples, showing peaks at CN1 (CNL), 2 (no CNV), and 3-8 (CNG). **C.** Recombinant haplotypes in the exons 9-11 homology region, distinguished by *GBA*/*GBAP1* differentiating bases (x-axis). Reference genome sequences are shaded in yellow. There is an error in hg38 where the first three sites of *GBAP1* show *GBA* bases, which could lead to alignment errors. The *GBA* recombinant haplotypes are shown in the white background, including those where one or a few nearby sites are mutated to the corresponding *GBAP1* base, resulting from either gene conversion or CNL. Gray bases indicate that the base can be either *GBA* or *GBAP1* depending on the breakpoint position of the CNL/conversion. Shaded in purple are two example *GBAP1* haplotypes, found by Gauchian, that have been partially converted to *GBA* and can cause false positive *GBA* variant calls by standard secondary analysis pipelines. For the first example, the reverse-p.L483P variant on *GBAP1* directs aligners to align *GBAP1* reads to *GBA*, causing the nearby p.A495P false-positive call. For the second example, the reverse-c.1263del variant inserts 55bp to *GBAP1*, driving *GBAP1* reads to align to *GBA*, causing the nearby p.D448H false-positive call.

Table titles and legends

Table 1: Details of cross-validation between Gauchian and ONT.

Sample[a]	CN change	Other variants	CNV	Variant type	Number of samples
NA20756	1		Gain	CNG with no other variant	11
HG01912	3				
HG01889	5				
HG02284	6				
HG03547	3				
NA19909	4				
HG03895	1				
NA18917	2				
NA19711	2				
HG03575	4				
Brain-S1	1				
PP-3307*	1	p.L483P		CNG + SNV	1
Brain-S2	4	c.1263del+RecTL		CNG + c.1263del+RecTL conversion	2
Brain-S3	4	c.1263del+RecTL			
HG03428	-1		Loss	CNL, non-pathogenic	3
NA19024	-1				
PP-12224	-1				
HG00422	-1	RecNciI		Pathogenic CNL (RecNciI CNL)	1
Brain-S4	-1	p.L483P		Non-pathogenic CNL+ p.L483P	1
HG00119	0	c.1263del+RecTL	No CN change	Gene conversion	2
HG00115	0	c.1263del+RecTL			
PP-3420	0	p.L483P		SNV	9
PP-3700	0	p.L483P			
PP-57787	0	p.L483P			
PP-59343	0	p.L483P			
PP-59926	0	p.L483P			
PP-60060	0	p.L483P			
Brain-S5	0	p.L483P			
PP-41342*	0	p.L483P/p.E365K			
PP-3429	0	p.A495P			

PP-3762*,PP-42378*,PP-3476,PP-3179,PP-3001,PP-3173,PP-3023,PP-42444,PP-3406,PP-56534,PP-52772,PP-41705	0	No *GBA* variants	12

Samples with * were discordant with BWA-GATK.

[a]Samples with IDs starting with NA- and HG- were obtained from NHGRI; samples with IDs starting with PP were obtained from PPMI; samples marked as brain were obtained from QSBB.

Table 2: Non-pathogenic CNVs in 1kGP and AMP-PD cohorts.

	1kGP			PD						LBD	
	European	African	Other	European		African		Other or Unknown		European	
	Control	Control	Control	Case	Control	Case	Control	Case	Control	Case	Control
CNL	2	3	9	8[a]	7	0	0	0	0	13[c]	13
CNG	1	74	18	11[b]	6	1	2	0	1	21[d]	11[e]
Total	503	661	1340	2227	1213	22	27	76	15	2598	1941

[a] Three out of the 8 PD cases with non-pathogenic CN losses also have a pathogenic *GBA* variant (2 samples have p.L483P and one samples has p.N409S).

[b] Four out of the 11 PD cases with CN gains also have a pathogenic *GBA* variant (3 samples have p.L483P and one sample has p.N409S).

[c] One out of the 13 LBD cases with non-pathogenic CN losses also has a pathogenic *GBA* variant, p.L483P.

[d] Five out of the 21 LBD cases with CN gains also have a pathogenic or PD-related *GBA* variant (p.L483P, p.D448H, c.1263del+RecTL, p.T408M, and compound heterozygote p.L483P / p.D448H).

[e] One out of 11 LBD controls with CN gains also has a PD-related *GBA* variant, p.T408M.

Table 3: *GBAP1*-like variants in the exons 9-11 homology region in 1kGP, PD, and LBD cohorts.

		p.A495P	p.L483P	p.D448H	c.1263del	RecNcil		c.1263del+RecTL		Total
						CNL	Conversion	CNL	Conversion	
1kGP	N=2504	1	5	0	2	1	0	0	2	11
PD	Case (N=2325)	3	14	1	0	1	2	1	0	22
	Control (N=1255)	0	6	1	0	0	0	0	0	7
LBD	Case (N=2598)	4	23	4	6	10	3	2	2	54*
	Control (N=1941)	2	0	1	0	0	0	0	0	3
PD+LBD called by Gauchian		9	43	7	6	11	5	3	2	86
PD+LBD called by BWA-GATK		9 (+11 FP)	27	7 (+2 FP)	0	0	1	0	0	44

*One sample is compound heterozygous for p.L483P and p.D448H.

Table 4: Summary of samples carrying *GBA* coding variants detected in 1kGP, PD, and LBD cohorts

		p.N409S	Severe* variants	Total	Total excluding p.N409S
1kGP	N=2504	3	14	53	50
PD	case N=2325	128	38	296	171
	control N=1255	158	9	200	43
	OR (95% CI)	n/a	2.12 (1.07-4.71)	n/a	2.07 (1.48-2.95)
LBD	case N=2598	59	79	353	298
	control N=1941	19	2	86	67
	OR (95% CI)	n/a	30.83 (9.71-187.55)	n/a	3.68 (2.82 – 4.87)

*Severe and mild variants are defined in Table S8