

# UCLA

## UCLA Previously Published Works

### Title

African ancestry neurodegeneration risk variant disrupts an intronic branchpoint in GBA1.

### Permalink

<https://escholarship.org/uc/item/3668224k>

### Journal

Nature structural biology, 31(12)

### Authors

Álvarez Jerez, Pilar

Wild Crea, Peter

Ramos, Daniel

et al.

### Publication Date

2024-12-01

### DOI

10.1038/s41594-024-01423-2

Peer reviewed

# African ancestry neurodegeneration risk variant disrupts an intronic branchpoint in *GBA1*

Received: 25 March 2024

Accepted: 10 October 2024

Published online: 12 December 2024

 Check for updates

A list of authors and their affiliations appears at the end of the paper

Recently, an African ancestry-specific Parkinson disease (PD) risk signal was identified at the gene encoding glucocerebrosidase (*GBA1*). This variant (**rs3115534-G**) is carried by ~50% of West African PD cases and imparts a dose-dependent increase in risk for disease. The risk variant has varied frequencies across African ancestry groups but is almost absent in European and Asian ancestry populations. *GBA1* is a gene of high clinical and therapeutic interest. Damaging biallelic protein-coding variants cause Gaucher disease and monoallelic variants confer risk for PD and dementia with Lewy bodies, likely by reducing the function of glucocerebrosidase. Interestingly, the African ancestry-specific *GBA1* risk variant is a noncoding variant, suggesting a different mechanism of action. Using full-length RNA transcript sequencing, we identified partial intron 8 expression in risk variant carriers (G) but not in nonvariant carriers (T). Antibodies targeting the N terminus of glucocerebrosidase showed that this intron-retained isoform is likely not protein coding and subsequent proteomics did not identify a shorter protein isoform, suggesting that the disease mechanism is RNA based. Clustered regularly interspaced short palindromic repeats editing of the reported index variant (**rs3115534**) revealed that this is the sequence alteration responsible for driving the production of these transcripts containing intron 8. Follow-up analysis of this variant showed that it is in a key intronic branchpoint sequence and, therefore, has important implications in splicing and disease. In addition, when measuring glucocerebrosidase activity, we identified a dose-dependent reduction in risk variant carriers. Overall, we report the functional effect of a *GBA1* noncoding risk variant, which acts by interfering with the splicing of functional *GBA1* transcripts, resulting in reduced protein levels and reduced glucocerebrosidase activity. This understanding reveals a potential therapeutic target in an underserved and underrepresented population.

Dementia with Lewy bodies (DLB) and Parkinson disease (PD) are believed to be caused by a combination of aging, environmental factors and genetics. Genetics has provided valuable insights into the underlying biology of disease. Damaging variants in multiple genes have been shown to cause disease and numerous variants have been associated with increased risk<sup>1–3</sup>. One particular gene of interest is *GBAI* (previously known as *GBA*), which encodes the lysosomal enzyme glucocerebrosidase (GCase). Damaging coding variants in *GBAI* increase the risk for PD and DLB across a wide spectrum of odds ratios (ORs)<sup>4,5</sup>. Interestingly, the phenotype of individuals with PD carrying *GBAI* variants is characterized by faster progression and a higher frequency of dementia compared to noncarriers<sup>6,7</sup>. It is most commonly hypothesized that *GBAI* mutations confer risk by reducing GCase activity. Furthermore, damaging *GBAI* variants are enriched in certain populations (for example, p.E365K in Northern Europeans and p.N409S in Ashkenazi Jews)<sup>8</sup>. Biallelic *GBAI* variants cause Gaucher disease, a lysosomal storage disorder that leads to a variety of clinical presentations.

In the human genome, *GBAI* is located on chromosome 1q22 and is adjacent to its pseudogene, known as *GBAILP*. *GBAILP* has very high sequence homology with *GBAI* (~96%). Consequently, accurately mapping short-read RNA and DNA sequencing in this region is a complex task<sup>9,10</sup>. Much is unknown about the potential function of *GBAILP*; however, using long-read sequencing, which overcomes mapping issues in this genomic region, it is clear that *GBAILP* is expressed at the RNA level and there is some evidence for protein expression<sup>10</sup>.

Recently, a PD *GBAI* risk signal was identified in the first African ancestry PD genome-wide association study (GWAS)<sup>11</sup>. The main index variant was remarkably common in West African populations with an estimated frequency of ~50% in West African PD cases and a reported OR of 1.58 (95% confidence interval (CI) = 1.37–1.80,  $P = 2.397 \times 10^{-14}$ ) per allele. In addition, this variant was associated with an earlier age at onset of 2 years per allele ( $\beta = -2.004$ , s.e.m. = 0.57,  $P = 0.0005$ ). Strikingly, the main index variant (**rs3115534**, NM\_000157.4 (*GBAI*): c.1225-34C>A) is a noncoding variant reported to be the strongest expression (eQTL) and protein (pQTL) quantitative trait locus for *GBAI* in the African ancestry population<sup>12,13</sup>. We showed previously that there are no common coding variants or structural variants in linkage disequilibrium with **rs3115543**, implying a disease mechanism independent of protein-coding or genomic structural variants at this locus<sup>11</sup>.

Here, we elucidate the disease mechanism of an intronic *GBAI* PD risk variant seen as the first and major genetic risk factor in African ancestry populations. Additionally, we show that the index variant (**rs3115534**) causes abnormal splicing and processing of *GBAI* transcripts, only present in risk variant carriers.

## Results

### Functional dissection of the *GBAI* African ancestry locus

Recently, a noncoding *GBAI* variant was reported to be associated with increased risk for PD in African ancestry individuals (Fig. 1a)<sup>11</sup>. This variant was also reported to be an eQTL and pQTL resulting in increased gene expression (Fig. 1b) and decreased protein expression (Fig. 1c)<sup>12,13</sup>. Using UK Biobank Olink data, we replicated the previously reported association between **rs3115534** and *GBAI* protein levels. After filtering for African ancestry, a total of 1147 samples remained with a *GBAI* protein measure (43 GG, 351 GT and 753 TT). The G allele is significantly associated with lower *GBAI* protein levels ( $P = 0.006$ ,  $\beta = -0.074$ , s.e.m. = 0.027). (Supplementary Fig. 1). It is important to note that, in the reference genome used here (hg38), G is the reference allele for **rs3115534**, although G is the risk allele biologically. In contrast, the alternative allele in hg38 (**rs3115534-T**) functions as the nonrisk allele and the more common allele globally.

Given the complexity of short-read DNA and RNA mapping in the *GBAI* region because of high sequence homology with *GBAILP*, we generated Oxford Nanopore Technologies (ONT) long-read RNA sequencing (RNAseq) data to investigate this region in a more accurate

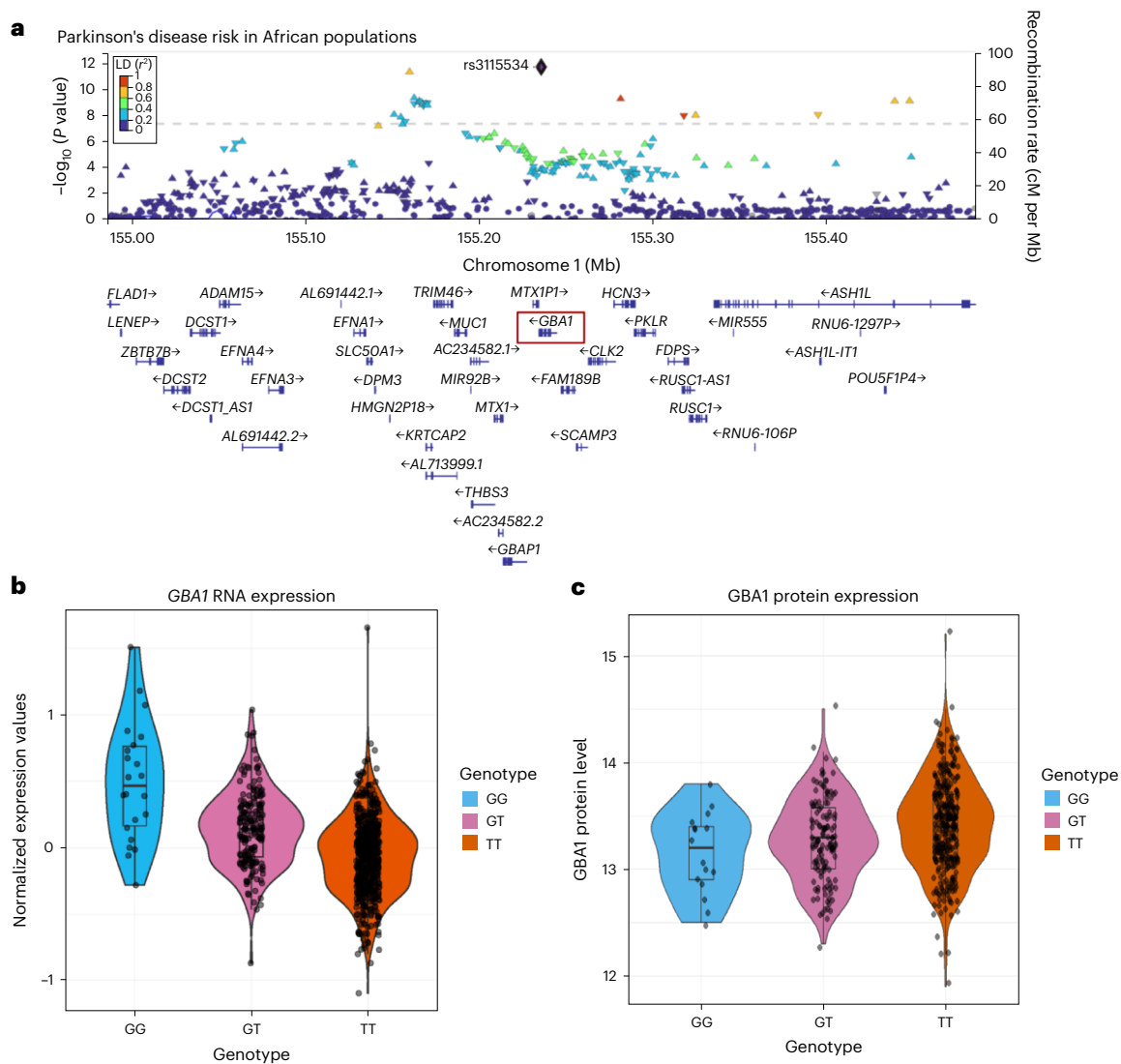
and comprehensive manner. Long-read RNAseq data were generated from eight African ancestry lymphoblastoid cell lines (LCLs) from Coriell across risk genotypes (one homozygous risk, GG; four heterozygous risk, GT; three homozygous nonrisk, TT). Surprisingly, we identified that there was a clear enrichment of sequence reads in the intron 8 region proximal to exon 9 that was specific to carriers of the G risk allele (Fig. 2a and Supplementary Table 11).

Subsequently, de novo StringTie2 isoform calling suggested that there were multiple unannotated transcripts including one starting approximately 40 bp before exon 9 and two full-length *GBAI* isoforms including all of intron 8, hereafter collectively referred to as ‘intron 8 expression’ (Supplementary Fig. 2). Interestingly, the reported index variant **rs3115534** is included in intron 8 expression (Fig. 2a, black arrow). We quantified the expression levels of this region and determined that expression levels of intron 8 were highly correlated with the presence of the G genotype (Fig. 2a–c and Supplementary Table 12). While the expression of the whole intron 8 (Fig. 2C) did not seem to follow a dose-dependent effect, this was perhaps because of only having a single homozygous G sample. In addition, no clear differences were observed for neighboring exons 8 and 9 (Fig. 2d,e) nor for total *GBAI* or *GBAILP* expression (Supplementary Fig. 3). Importantly, every mapped sequence read exhibited a G allele, indicating their origin from the G risk haplotype. We manually changed the base in G reads to T to confirm that these reads mapped uniquely to *GBAI* and that the striking difference we observed was not an artifact of mismapping to *GBAILP* driven by the **rs3115534** variant. No differences in mapping were observed (Supplementary Fig. 4).

Next, we examined whether intron 8 is also expressed in the human brain. Using ONT long-read RNAseq of eight African ancestry Human Brain Collection Core (HBCC) frontal cortex samples from different genotypes (four GG, two GT and three TT), we ran StringTie2 with the same methods as the LCLs. However, even when using our custom transcript models with our transcripts of interest, StringTie2 only identified the shorter intron 8 transcript in one GG carrier, likely because of lower expression of *GBAI* in the brain and lower RNA quality, as evidenced by lower RNA integrity number (RIN) values (Supplementary Tables 1 and 12). However, when looking at expression coverage plots, intron 8 expression was observed in all carriers of the G allele (GG and GT) but not in TT carriers (Fig. 3a, Supplementary Fig. 5 and Supplementary Table 11). Next, we generated Illumina short-read RNAseq for 18 LCLs (including the initial eight from above) and identified a similar albeit less pronounced increased read coverage across intron 8 (Fig. 3b and Supplementary Fig. 6). Likewise, similar enrichment in the intron 8 region correlated with the G allele of **rs3115534** was seen in Illumina RNAseq data from the HBCC frontal cortex (Fig. 3c and Supplementary Fig. 7), 1000 Genomes Project LCL RNAseq (Fig. 3d) and Accelerating Medicine Partnership (AMP) PD blood-based RNAseq data (Fig. 3e). Importantly, with increased numbers, a likely allelic dosage effect becomes visible (that is, GG has significantly more intron 8 expression compared to GT carriers), although larger GG numbers in future datasets would be helpful to confirm this pattern (Supplementary Figs. 8 and 9 and Supplementary Table 11).

### *GBAI* intron 8 expression unlikely encodes new protein product

Because of the comparatively lower *GBAI* expression in frontal cortex tissues and the greater accessibility of LCLs, the majority of subsequent experiments and analyses were conducted using LCLs. To validate the inclusion of *GBAI* intron 8, we designed primers specific to the most highly expressed region proximal to exon 9. Reverse transcription (RT)–PCR and subsequent Sanger sequencing validated the presence of this short transcript containing intron 8 in LCLs only in G risk variant carriers, confirming our initial findings (Supplementary Figs. 10 and 11). Next, we used cap analysis of gene expression and sequencing (CAGEseq), a technique often used to accurately pinpoint transcription start sites. The CAGEseq library preparation is dependent on the



**Fig. 1 | Overview of the African ancestry PD *GBA1* GWAS locus. a**, LocusZoom plot showing **rs3115534** as index variant (purple diamond) and located in intron 8 of *GBA1*. **b**, **rs3115534** as eQTL for *GBA1* RNA expression from Kachuri et al. showing increased *GBA1* expression with G risk genotypes<sup>12</sup>. Genotype counts: GG,  $n = 22$ ; GT,  $n = 185$ ; TT,  $n = 537$ . **c**, **rs3115534** as pQTL for *GBA1* protein

expression from Surapaneni et al. showing decreased protein levels with G risk genotypes<sup>13</sup>. Genotype counts: GG,  $n = 16$ ; GT,  $n = 134$ ; TT,  $n = 317$ . For all box plots, the center line represents the median, edges of the box represent the first and third quartiles and ends of bars represent the maximum and minimum (not including outliers).

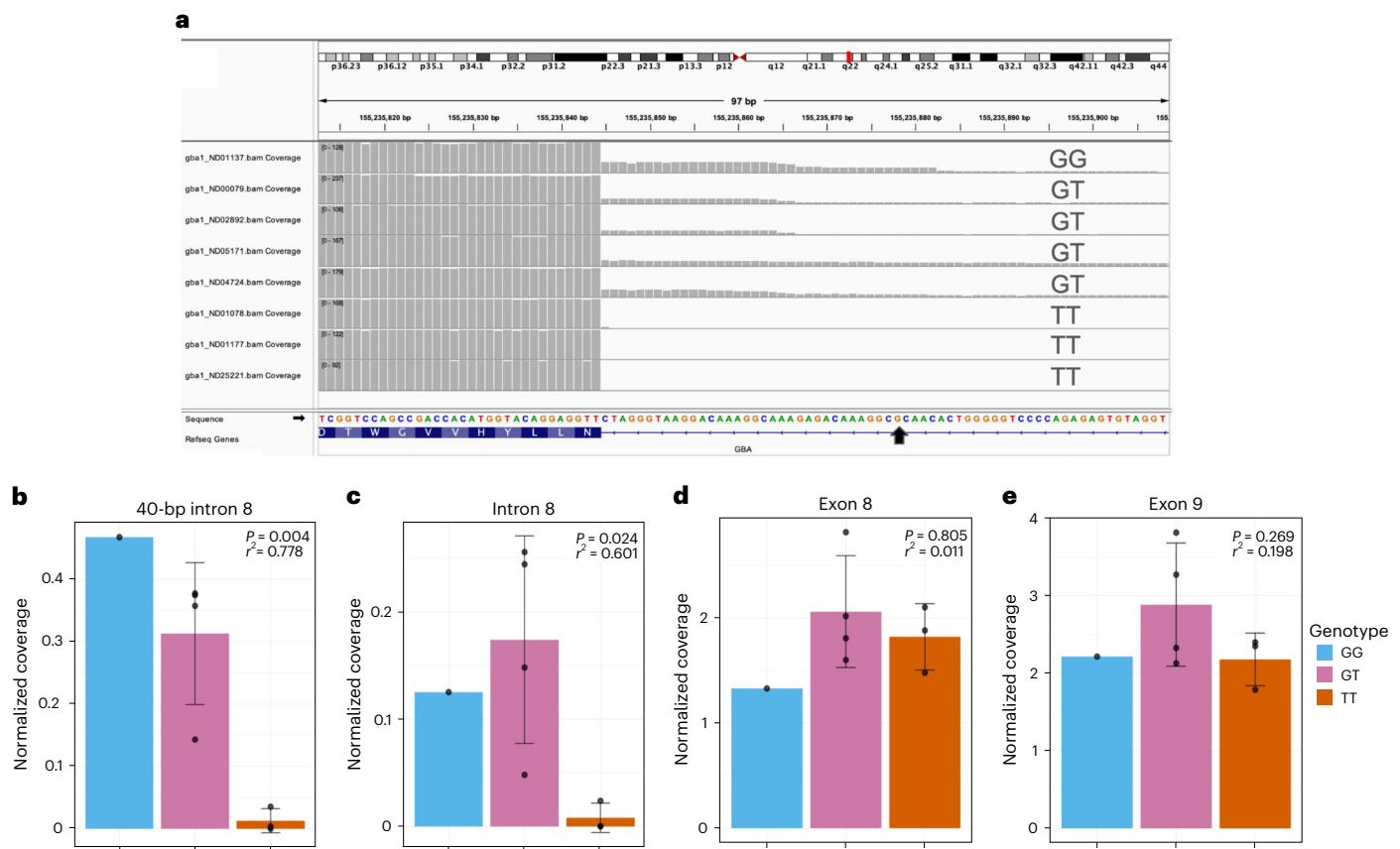
presence of an RNA cap, which is used to capture RNAs. We detected the canonical *GBA1* transcription start site; however, we did not identify a unique transcription start site within intron 8 (Supplementary Fig. 12). This indicates that the transcript containing intron 8 does not have a 5' cap. However, the transcripts do have a poly(A) tail that can be detected with ONT long-read RNAseq.

The protein-coding capacity of this transcript was assessed using multiplexed and enhanced chemiluminescence (ECL) western blotting. Given that the highest intronic expression was observed close to exon 9, an C terminus antibody specific to amino acids 517–536 of the functional GCse protein was selected to ensure the capture of proteins that had a more downstream start site. Both multiplex and ECL methods consistently identified the native GCse protein but showed no other bands, indicating a lack of novel protein (Supplementary Fig. 13). Subsequent mass-spectrometry-based analysis of the gel region that was predicted to harbor the protein of the short transcript did not identify any known *GBA1* peptides (Supplementary Table 13). These results suggest that the transcript is not protein coding and that the disease mechanism is likely to be RNA based.

### **rs3115534 is the functional risk variant at the *GBA1* locus**

The close proximity of the index risk variant to the short transcript and the lack of other variants in linkage disequilibrium nearby made **rs3115534** a strong candidate to be the functional effect variant for disease risk (Fig. 1a). To investigate this, we performed clustered regularly interspaced short palindromic repeats (CRISPR) editing on two LCLs: one in which the homozygous risk genotype (GG) was edited to be homozygous nonrisk (TT) and one in which the homozygous nonrisk genotype (TT) was edited to be homozygous risk (GG) (Supplementary Table 10). Genotyping qPCR was used to confirm the **rs3115534** genotype in the edited lines. All lines contained the genotype as expected except one, in which full conversion from GG to TT was only partially successful and led to a heterozygous GT line. This edited line was excluded from downstream analysis.

Next, we performed ONT long-read DNA and RNAseq to confirm successful CRISPR editing and to assess the presence or absence of the intron 8 expression. ONT long-read DNA showed successful editing of line ND22789 from TT to GG and partial editing of line ND01137 from GG to GT, validating the qPCR results. ONT long-read RNAseq



**Fig. 2** *GBA1* intron 8 expression is correlated with *rs3115534* genotype. **a**, ONT long-read RNAseq of eight LCLs shows a consistent pattern where the *rs3115534*-G risk allele is associated with intron 8 expression and absent in homozygous T (nonrisk allele) individuals generated using IGV. **b,c**, Quantification of intron 8 expression is significantly associated with the G allele in both the 40-bp region before exon 9 (**b**) and the full intron 8 (**c**) (linear regression,  $P < 0.05$ ). **d,e**, No

significant differences were identified in the two neighboring exons 8 (**d**) and 9 (**e**). Coverage for all panels was normalized by dividing the mean depth by the total number of mapped reads per million (Methods). In **b–e**, a linear regression was run with GG + GT in one group versus TT. For all panels, genotype counts are as follows: GG,  $n = 1$ ; GT,  $n = 4$ ; TT,  $n = 3$ . Error bars represent the s.d. for all panels with the center at the mean. The unadjusted  $r^2$  is displayed.

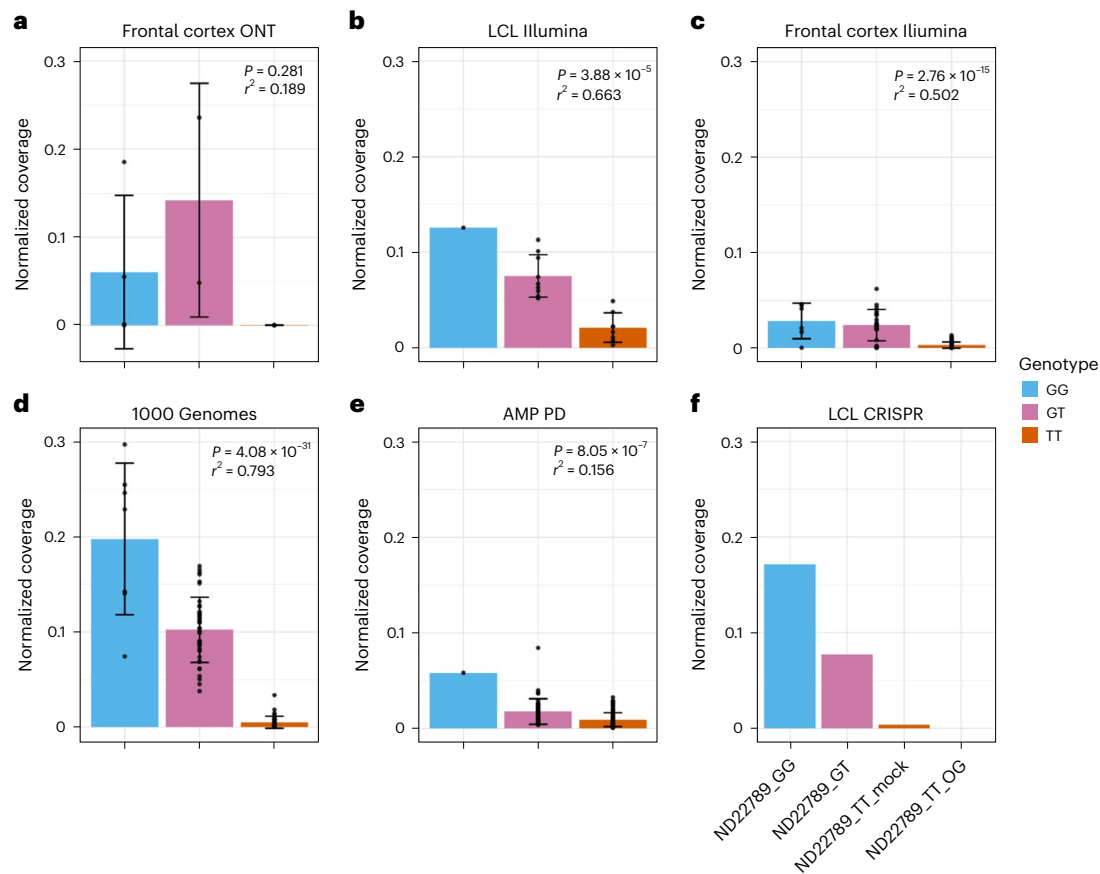
again showed the presence of intron 8 sequence reads but only in LCLs carrying a G allele. Quantification of the intronic reads for the full allelic series of the ND22789 edited line displayed an effect specific to the G allele (Fig. 3f), showing that the *rs3115534*-G allele is solely responsible for intron 8 transcription in LCLs. Quantification of both edited lines individually and collapsed by genotype is provided in Supplementary Fig. 14.

### Intron 8 is expressed across cell types in the human brain

Next, we assessed in which brain cell type this short transcript is expressed. Initial screening of frontal cortex brain single-nucleus RNAseq showed that overall *GBA1* expression is too low to accurately assess transcript expression and intron 8 expression was not observed (Supplementary Fig. 15). Therefore, we used an enrichment strategy with probes targeting the *GBA1* region in the single-nucleus complementary DNA (cDNA) library. After enrichment, we extracted *GBA1* transcripts and performed long-read ONT RNAseq on the single-nucleus libraries from one GG carrier and one TT carrier. Sequencing showed clear enrichment for *GBA1* and *GBA1LP* transcripts (Supplementary Table 14). *GBA1* and *GBA1LP* were ubiquitously expressed across major brain cell types (Supplementary Fig. 16a). Supporting our previous results, reads covering intron 8 were predominantly identified in the GG carrier as shown above (Supplementary Fig. 16a,b). When assessing the expression of the intron 8 region across cell types, the region was ubiquitously expressed across cell types, similar to the full *GBA1* transcript.

### *rs3115534* is located in a key *GBA1* intron 8 branchpoint

To investigate potential functional downstream consequences of *rs3115534*, we explored several in silico algorithms. RegulomeDB analysis to assess the effect of *rs3115534* on motifs and genome accessibility reported that this variant is an eQTL and is located in a region of open chromatin. However, no transcription factor motifs are affected by this variant. Sequence conservation analysis showed that the T allele (nonrisk) is highly conserved across vertebrates and humans are the only vertebrates harboring G as a reference allele (Supplementary Fig. 17). Interestingly, when analyzing the 5-methylcytosine DNA modifications in the ONT long-read DNA data of CRISPR-edited LCLs, we identified that the G allele is methylated. When *rs3115534* is a T, the methylation is lost (Supplementary Fig. 18). This change in methylation status was confirmed when ONT sequencing was performed on the initial LCLs (Supplementary Fig. 19) and two additional frontal cortical brain samples (Supplementary Fig. 20). Lastly, given that *rs3115534* was close to an exon (34 bp), we also assessed the variant's potential to disrupt splicing. We used two complementary approaches: (1) SpliceAI, which is based on a deep neural network that accurately predicts splice junctions from pre-mRNA transcript sequences, and (2) Branchpointer and AGAIN, which are algorithms that are driven by an understanding of splicing biology. Using SpliceAI, no significant score of interest ( $< 0.2$ ) was identified. However, using Branchpointer and AGAIN, we identified that this variant is located within the key intronic branchpoint sequence of intron 8. The *rs3115534*-G allele (risk, C in the minus strand) is likely



**Fig. 3 | Increased intron 8 expression across datasets in G allele carriers.**

**a**, Intron 8 coverage from human frontal cortex sequenced with ONT ( $n = 8$ ). Intron 8 expression is only present in G allele carriers but does not reach statistical significance ( $P = 0.281$ ) likely because of a smaller sample size. **b**, Intron 8 coverage from LCLs ( $n = 18$ ) sequenced with Illumina. Expression is significantly associated with the G allele ( $P = 3.88 \times 10^{-5}$ ). **c**, Intron 8 expression from human frontal cortex sequenced with Illumina ( $n = 92$ ). Expression is significantly associated with the G allele ( $P = 2.76 \times 10^{-15}$ ). **d**, Intron 8 coverage from LCLs in the 1000 Genomes Project cohort ( $n = 88$ ). Expression is significantly associated with the G allele ( $P = 4.08 \times 10^{-31}$ ). **e**, Intron 8 coverage from blood in the AMP

PD cohort ( $n = 148$ ). Expression is significantly associated with the G allele ( $P = 8.05 \times 10^{-7}$ ). **f**, CRISPR editing of LCLs showed that the *rs3115534*-G risk allele is significantly associated with intron 8 expression. Here, coverage is shown for the full allelic series of ND22789, a TT line originally CRISPR edited through to a GG line. Coverage for all panels normalized by dividing the mean regional depth by the total number of mapped reads per million (Methods). In **a, b, e**, a linear regression was run with GG + GT in one group versus TT. In **c, d**, a linear regression was run with GG, GT and TT in separate groups. Error bars represent the s.d. for all panels with the center at the mean. The unadjusted  $r^2$  is displayed.

disrupting the splicing process by replacing the key A base (nonrisk, T in the positive strand) (Fig. 4).

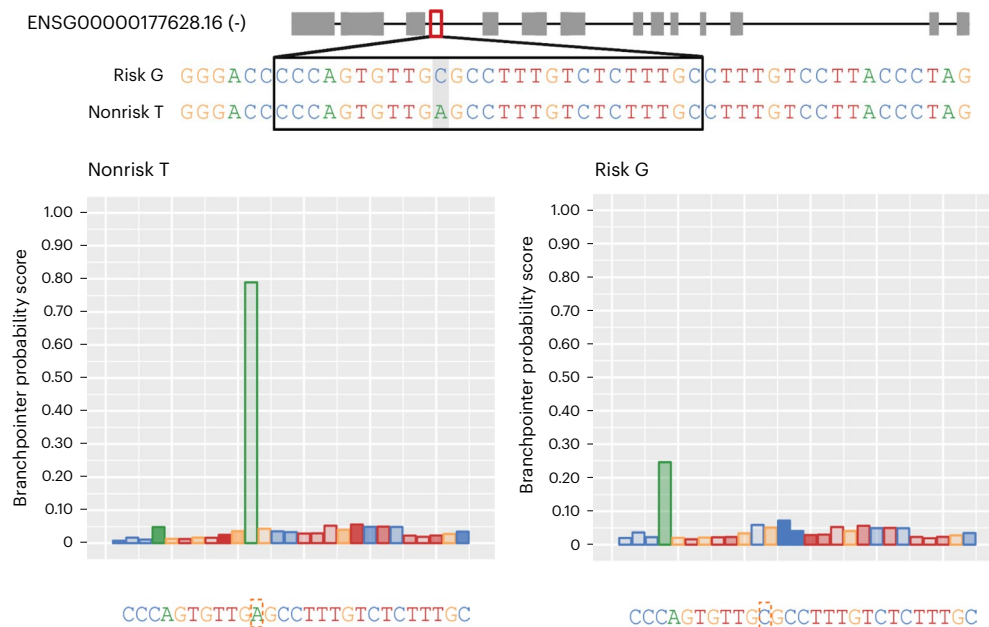
This explains our initial finding from the ONT long-read RNAseq, which suggested that new transcripts (intron 8 containing) were transcribed only from the G risk haplotype (Supplementary Fig. 2). The ‘shorter’ transcripts, containing the part of intron 8 close to exon 9, are potentially spliced at the AG sequence (CT on sense strand) 6 nt upstream of *rs3115534* and are sequenced as a short transcript given that they retain the poly(A) tail. The longer transcripts, containing the full nonspliced intron 8, are sequenced in full and also arise exclusively from the G risk haplotype. Here, intron 8 is partially retained, likely because of premature splicing of the intron occurring as a result of branchpoint disruption in the presence of the G (risk) allele in a highly conserved nucleotide position.

Next, we investigated the branch sites engaged with U2 small nuclear ribonucleoprotein (snRNP) in human 293Flp-in cells, which we determined to be homozygous for *rs3115534*-T (nonrisk). Using previously generated data<sup>14</sup>, we found further evidence that *rs3115534*-T is the main branchpoint nucleotide in intron 8 of *GBA1*. Different branch site datasets, obtained with constitutive (SF3A2) or associated (RBM5 and RBM10) U2 snRNP proteins, all predominantly used the adenine nucleotide at *rs3115534*-T as a branchpoint

(Supplementary Fig. 21), confirming the bioinformatic branchpoint prediction algorithms.

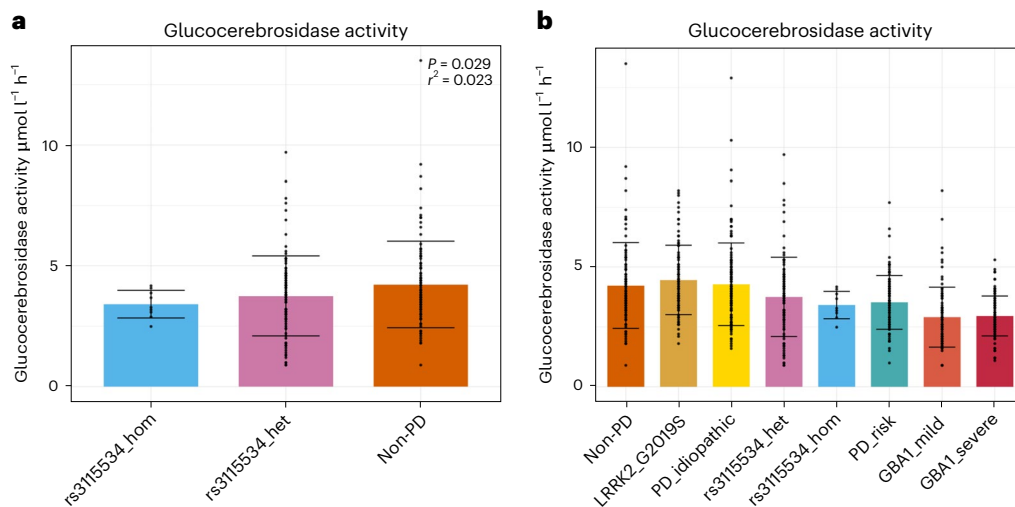
#### Assessing downstream effects of the branchpoint disruption

Given these results, we also wanted to assess potential consequences on GCase activity levels. Using Centogene’s CentoCards we compared GCase activity across individuals without PD, individuals with idiopathic PD, heterozygous *GBA1* carriers with PD-coding risk variants (p.E365K, p.T408M), heterozygous *GBA1* carriers with mild Gaucher disease-coding risk variants (p.N409S), heterozygous *GBA1* carriers with severe Gaucher disease-coding risk variants (including p.L483P) and *rs3115534* variant carriers in heterozygous (GT) and homozygous state (GG). Using these data, a significant dose-dependent reduction in GCase activity correlated with the *rs3115534*-G risk genotype was observed when running a linear regression across all three groups ( $P = 0.029$ ,  $\beta = -0.449$ , s.e.m. = 0.205) (Fig. 5a). Notably, the G-allele-associated reduction in GCase activity is similar to PD-coding risk variants that do not cause Gaucher disease (*rs3115534*.het versus PD\_risk,  $P = 0.1333$ ,  $\beta = -0.224$ , s.e.m. = 0.201; one-sided  $P$  value test) and higher than the Gaucher disease-causing variants (*rs3115534*.het versus *GBA1*\_mild,  $P = 1.00 \times 10^{-4}$ ,  $\beta = -0.846$ , s.e.m. = 0.215 and *rs3115534*.het versus *GBA1*\_severe  $P = 1.52 \times 10^{-5}$ ,  $\beta = -0.800$ , s.e.m. = 0.186; one-sided



**Fig. 4 | The *GBA1* intronic rs3115534 variant acts as a splicing branchpoint.** The causal variant rs3115534 (highlighted in gray on the top and in red dashed box on the bottom) in intron 8 is located in the key splicing branchpoint according to Branchpointer. When rs3115534 is in a nonrisk state (T), on the antisense strand,

the A allele functions as a branch site for the spliceosome, whereas, in the risk state (G), on the antisense strand, the C allele disrupts this branch site resulting in abnormal splicing.



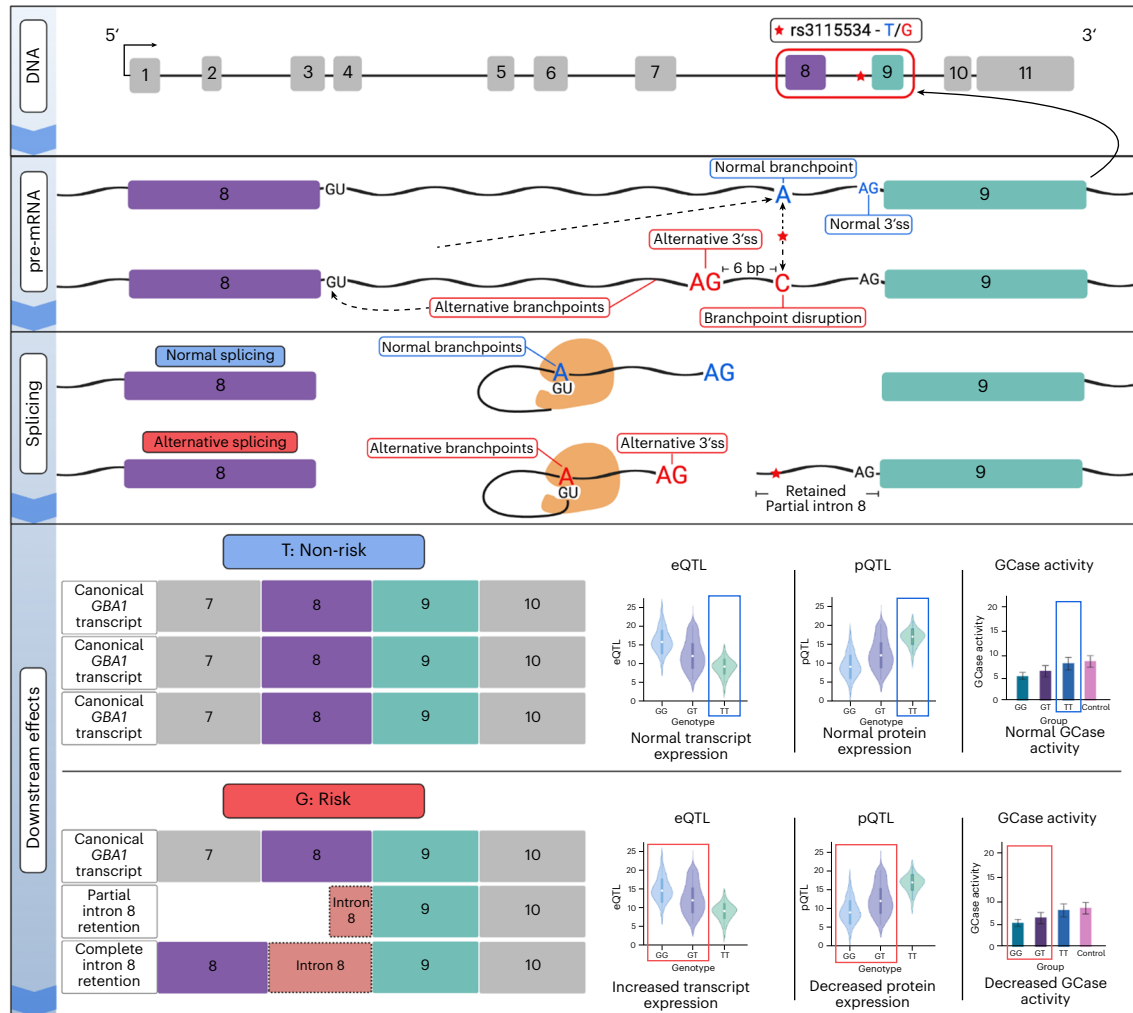
**Fig. 5 | Measuring GCase activity across *GBA1* genotypes.** **a**, GCase activity was measured across rs3115534 *GBA1* genotypes showing a significant dose G-allele-dependent reduction across genotypes ( $P = 0.029$ ). A linear regression was run with GG, GT and TT in separate groups. The unadjusted  $r^2$  is displayed. **b**, When measuring GCase activity across multiple heterozygous *GBA1* genotypes, it appears that rs3115534-GT and rs3115534-GG reduce GCase activity to similar levels to heterozygous PD risk variants (p.E365K and p.T408M) but

remain higher than heterozygous *GBA1* Gaucher disease-causing variants (*GBA1\_mild*, such as p.N409S, and *GBA1\_severe*, such as p.L483P). Note that **a** was extracted from **b** and the same data are included in both. Group counts are as follows: non-PD,  $n = 97$ ; LRRK2\_G2019S,  $n = 95$ ; PD\_idiopathic,  $n = 122$ ; rs3115534\_het,  $n = 99$ ; rs3115534\_hom,  $n = 99$ ; PD\_risk,  $n = 99$ ; *GBA1\_mild*,  $n = 90$ ; *GBA1\_severe*,  $n = 99$ . Error bars represent the s.d. for all panels.

$P$  value tests), confirming that the reduction in GCase is not enough to cause Gaucher disease (Fig. 5b). Supplementary Table 11 provides full details on statistical testing. Combined with our other results, these data suggest that, while no new protein product is made, the risk allele disrupts normal splicing and creates a new transcript that is likely non-functional, leading to a decrease in overall protein levels and, therefore, lower GCase activity (Fig. 6).

Lastly, we wanted to assess whether other variants in the *GBA1* region potentially affect branchpoint sequences. When exploring all

*GBA1* variants from gnomAD ( $n = 1,625$ ), three variants were identified to be of interest by the AGAIN algorithm, including the rs3115534 variant (Supplementary Table 15). The two other variants, rs140335079 and rs745734072, did not show a similar Branchpointer pattern (Supplementary Fig. 22). When exploring intron expression for rs140335079 in AMP PD, we did not identify any differences across genotypes (Supplementary Fig. 23). No RNAseq data were available for rs745734072, given that it is a rare variant and only identified in the South Asian population.



**Fig. 6 | Suggested variant-to-function hypothesis of rs3115534.** rs3115534-T is transcribed to pre-mRNA as A, a highly conserved branchpoint nucleotide. However, rs3115534-G, which confers elevated PD risk, is instead transcribed to C, causing the observed branchpoint disruption (rs3115534, denoted as red star). This single-nucleotide change in intron 8 impacts splicing by disrupting the normal binding from the adenosine branchpoint nucleotide and the 5' splice site

(GU). Subsequently, an alternative branchpoint is used, uncovering an alternative 3' splice site upstream of the normal splice site immediately proximal to exon 9 and resulting in partial and complete intron retention and fewer functional *GBA1* mRNA transcripts. Downstream, abnormal splicing of *GBA1* leads to reduced GCCase protein and subsequent lower GCCase activity, which is a known pathomechanism of PD and DLB. ss, splice site. Generated with BioRender.com.

## Discussion

Most of the current genetic, genomic and functional knowledge in the neurodegeneration field is based on European ancestry findings. Recent efforts are showing great progress in making genomics more diverse, including GWAS reporting in East Asia, South Asia, Latin America and Africa<sup>11,15-17</sup>. In the past decade, GWAS has been the workhorse in genetics and has shown tremendous progress in the identification of genomic regions associated with traits and diseases<sup>18</sup>. However, one of the main limitations of GWAS is that it cannot pinpoint the exact mechanism, causal gene or variant in associated genomic regions. Therefore, follow-up methods such as QTL analysis and functional studies are used to fill this gap.

Recent progress in African ancestry PD genetics identified a non-coding *GBA1* risk variant associated with disease risk and an earlier age at onset<sup>11</sup>. Here, we show extensive follow-up of genomic and transcriptomic data from this region, pinpoint the functional variant and highlight the likely variant-associated mechanism. Using long-read RNAseq, we identified intron 8 retention, which was enriched in close proximity to the reported index variant rs3115534. Expression levels of intron 8 were correlated with G allele dosage and very low or absent in TT carriers. Given the high sequence overlap between *GBA1* and

*GBA1LP*, it is unlikely that this expression is simply a sequence mapping artifact. The absence of a CAGEseq peak in this intronic region and the lack of a detectable additional protein isoform suggest that the risk mechanism is RNA based. CRISPR experiments showed that the reported index variant rs3115534 is the variant responsible for directly changing transcription. Importantly, given the notable in silico evidence for an intronic branchpoint disruption of rs3115534, it is clear that this is the cause of the intron 8 expression.

RNA splicing is the process where pre-mRNA is transformed into mRNA by removing introns. This complex process depends on the donor site (GT), the acceptor site (AG) and the branch site (often a sequence motif including an adenine base, known as the branchpoint). Disruption of any of these sequences causes missplicing, which typically results in reduced functional mRNA. Mutations in the branchpoint sequence motif are known to contribute to disease<sup>19-22</sup>. In the case of rs3115534, the A allele (nonrisk, T in the sense strand) is part of the key branch site sequence of intron 8; therefore, when rs3115534 is mutated to the C allele (risk, G in the sense strand) partial splicing disruption occurs (Fig. 6). This disruption results in partial intron 8 retention in some *GBA1* transcripts. This becomes particularly clear when investigating the long-read sequencing data showing enrichment



of sequence reads close to exon 9 that start at the next intronic AG acceptor site 39 nt before exon 9, which is often seen in other branch site sequence disruptions.

Our finding that branchpoint disruption may confer disease risk highlights a potential mechanism for increased disease risk. All prior knowledge of *GBAI* disease mechanisms has attributed damaging coding variation to reduced GCcase activity, leading to disease risk. However, there is some evidence that supports potential alternative mechanisms<sup>23</sup>. Here, we show that the branchpoint disruption likely causes splicing dysregulation, which results in increased risk by lowering the protein levels (pQTL) and, therefore, reducing the GCcase activity correlated with genotype dosage. This is consistent with the hypothesis that reduction in GCcase activity is the pathomechanism underlying PD and DLB and highlights a potential therapeutic target for African ancestry individuals.

Importantly, there are several inherent limitations of the study, many of which are driven by the lack of ancestrally diverse tissue, cellular and data resources. First, given the high sequence homology between *GBAI* and *GBAILP* and the low expression levels of the potential short transcript isoforms, we cannot exclude its protein-coding potential. It could be possible that this isoform is too lowly expressed or is degraded too quickly for the antibody to detect the protein expression or to be seen on generalized mass spectrometry analysis. Second, although we measured GCcase activity to assess potential downstream effects of the rs3115534 variant and several other *GBAI* genotypes, this assay remains very variable across samples and tissues. Here, we include a large collection of samples and observed expected effects across *GBAI* genotypes and, thus, believe that the results are robust, although access to additional biological samples in the future will allow us to repeat the experiment at higher power and greater resolution. In addition to larger sample collections, research across different cell types, especially including brain-related cell types, will allow us to measure the exact reduction in GCcase levels, assess how the branchpoint disruption behaves across cell types and determine what the downstream consequences are of the reduction in GCcase protein level. Ongoing recruitment efforts in the Global Parkinson's Genetics Program (GP2; <https://gp2.org/>) aim to fill that gap<sup>24</sup>.

In summary, we report the rapid translation of a previously identified African ancestry GWAS locus and show the power of genetic diversity, which can result in valuable new biological insights into well-studied genes such as *GBAI*. We provide compelling evidence that rs3115534-G is the causal variant and show that the likely functional mechanism is a disruptive allelic change in the intronic branchpoint sequence that disrupts splicing, resulting in reduced protein levels. This shows how a common GWAS variant can be functionally explained by an intronic branchpoint sequence alteration, a potentially underexplored mechanism for GWAS. Overall, this implies that rs3115534-G is the most common damaging *GBAI* variant with a minor allele frequency of over 20% in certain populations and for which a substantial number of West African PD cases are heterozygous (40%) or homozygous (13%) risk carriers<sup>11</sup>. There is no evidence that this variant causes Gaucher disease, a condition that is remarkably infrequent in African ancestry groups, which is likely because of the magnitude of the biological effect of this variant, where the result is only a partial reduction of protein levels and GCcase activity. Interestingly, this is an alternative mechanism for increased disease risk at *GBAI* and no other variants in *GBAI* were identified to act through a similar mechanism. The frequency of this risk variant, its mapping as the functional allele at this risk locus and the variant's mode of action at an intronic branchpoint make this an attractive candidate for precision-based therapeutics in a remarkably underserved population. This work underscores the scientific and societal importance of working in groups underrepresented in research, as well as the need for the generation of biological and data resources in these groups, and marks the start of the realization of the mission of GP2.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41594-024-01423-2>.

## References

- Blauwendraat, C., Nalls, M. A. & Singleton, A. B. The genetic architecture of Parkinson's disease. *Lancet Neurol.* **19**, 170–178 (2020).
- Chia, R. et al. Genome sequencing analysis identifies new loci associated with Lewy body dementia and provides insights into its genetic architecture. *Nat. Genet.* **53**, 294–303 (2021).
- Nalls, M. A. 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.* **18**, 1091–1102 (2019).
- Sidransky, E. et al. Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. *N. Engl. J. Med.* **361**, 1651–1661 (2009).
- Nalls, M. A. et al. A multicenter study of glucocerebrosidase mutations in dementia with Lewy bodies. *JAMA Neurol.* **70**, 727–735 (2013).
- Malek, N. et al. Features of *GBA*-associated Parkinson's disease at presentation in the UK Tracking Parkinson's study. *J. Neurol. Neurosurg. Psychiatry* **89**, 702–709 (2018).
- Iwaki, H. et al. Genetic risk of Parkinson disease and progression: an analysis of 13 longitudinal cohorts. *Neurol. Genet.* **5**, e348 (2019).
- Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
- Toffoli, M. et al. Comprehensive short and long read sequencing analysis for the Gaucher and Parkinson's disease-associated *GBA* gene. *Commun. Biol.* **5**, 670 (2022).
- Gustavsson, E.K. et al. The annotation of *GBA1* has been concealed by its protein-coding pseudogene *GBAP1*. *Sci. Adv.* **10**, eadk1296 (2024).
- Rizig, M. et al. Identification of genetic risk loci and causal insights associated with Parkinson's disease in African and African admixed populations: a genome-wide association study. *Lancet Neurol.* **22**, 1015–1025 (2023).
- Kachuri, L. et al. Gene expression in African Americans, Puerto Ricans and Mexican Americans reveals ancestry-specific patterns of genetic architecture. *Nat. Genet.* **55**, 952–963 (2023).
- Surapaneni, A. et al. Identification of 969 protein quantitative trait loci in an African American population with kidney disease attributed to hypertension. *Kidney Int.* **102**, 1167–1177 (2022).
- Damianov, A. et al. The splicing regulators RBM5 and RBM10 are subunits of the U2 snRNP engaged with intron branch sites on chromatin. *Mol. Cell* **84**, 1496–1511 (2024).
- Foo, J. N. 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.* **77**, 746–754 (2020).
- Andrews, S. V. et al. The genetic drivers of juvenile, young, and early-onset Parkinson's disease in India. *Mov. Disord.* **39**, 339–349 (2024).
- Loesch, D. P. et al. Characterizing the genetic architecture of Parkinson's disease in Latinos. *Ann. Neurol.* **90**, 353–365 (2021).
- Sollis, E. et al. The NHGRI-EBI GWAS catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
- Zhang, P. et al. Genome-wide detection of human variants that disrupt intronic branchpoints. *Proc. Natl Acad. Sci. USA* **119**, e2211194119 (2022).


20. Xie, J., Wang, L. & Lin, R.-J. Variations of intronic branchpoint motif: identification and functional implications in splicing and disease. *Commun. Biol.* **6**, 1142 (2023).
21. Kadri, N. K., Mapel, X. M. & Pausch, H. The intronic branch point sequence is under strong evolutionary constraint in the bovine and human genome. *Commun. Biol.* **4**, 1206 (2021).
22. Kuivenhoven, J. A. et al. An intronic mutation in a lariet branchpoint sequence is a direct cause of an inherited human disorder (fish-eye disease). *J. Clin. Invest.* **98**, 358–364 (1996).
23. Kuo, S.-H. et al. Mutant glucocerebrosidase impairs  $\alpha$ -synuclein degradation by blockade of chaperone-mediated autophagy. *Sci. Adv.* **8**, eabm6393 (2022).
24. Global Parkinson's Genetics Program GP2: The Global Parkinson's Genetics Program. *Mov. Disord.* **36**, 842–851 (2021).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2024

**Pilar Álvarez Jerez** <sup>1,2,3,19</sup>, **Peter Wild Crea** <sup>2,19</sup>, **Daniel M. Ramos** <sup>1</sup>, **Emil K. Gustavsson** <sup>4</sup>, **Mandy Radefeldt**<sup>5</sup>, **Andrey Damianov**<sup>6</sup>, **Mary B. Makarios** <sup>1,7</sup>, **Oluwadamilola O. Ojo** <sup>8,9</sup>, **Kimberley J. Billingsley**<sup>1,2</sup>, **Laksh Malik** <sup>1</sup>, **Kensuke Daida** <sup>2</sup>, **Sarah Bromberek**<sup>1</sup>, **Fangle Hu**<sup>1</sup>, **Zachary Schneider**<sup>2</sup>, **Aditya L. Surapaneni**<sup>10</sup>, **Julia Stadler**<sup>1</sup>, **Mie Rizig** <sup>11</sup>, **Huw R. Morris** <sup>12</sup>, **Caroline B. Pantazis** <sup>1</sup>, **Hampton L. Leonard**<sup>1,7</sup>, **Laurel Screven**<sup>1</sup>, **Yue A. Qi** <sup>1</sup>, **Mike A. Nalls** <sup>1,7</sup>, **Sara Bandres-Ciga**<sup>1</sup>, **John Hardy**<sup>3</sup>, **Henry Houlden** <sup>11</sup>, **Celeste Eng**<sup>13</sup>, **Esteban González Burchard** <sup>13</sup>, **Linda Kachuri** <sup>14,15</sup>, **Chia-Ho Lin**<sup>6</sup>, **Douglas L. Black** <sup>6</sup>, **Global Parkinson's Genetics Program (GP2)\***, **Andrew B. Singleton**<sup>1,2</sup>, **Steffen Fischer** <sup>5</sup>, **Peter Bauer**<sup>5</sup>, **Xylena Reed** <sup>1</sup>, **Mina Ryten** <sup>4,16,17</sup>, **Christian Beetz**<sup>5</sup>, **Michael Ward** <sup>1,18</sup>, **Njideka U. Okubadejo** <sup>8,9</sup> & **Cornelis Blauwendraat** <sup>1,2</sup> 

<sup>1</sup>Center for Alzheimer's and Related Dementias, National Institute on Aging and National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA. <sup>2</sup>Laboratory of Neurogenetics, National Institute on Aging, Bethesda, MD, USA. <sup>3</sup>Department of Neurodegenerative Disease, UCL Queen Square Institute of Neurology, University College London, London, UK. <sup>4</sup>Genetics and Genomic Medicine, Great Ormond Street Institute of Child Health, University College London, London, UK. <sup>5</sup>Centogene, Rostock, Germany. <sup>6</sup>Department of Microbiology, Immunology and Molecular Genetics, The David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. <sup>7</sup>DataTecnica, Washington, DC, USA. <sup>8</sup>College of Medicine, University of Lagos, Lagos, Nigeria. <sup>9</sup>Lagos University Teaching Hospital, Lagos, Nigeria. <sup>10</sup>Department of Medicine, New York University Langone School of Medicine, New York, NY, USA. <sup>11</sup>Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, University College London, London, UK. <sup>12</sup>UCL Movement Disorders Centre, University College London, London, UK. <sup>13</sup>Department of Biotherapeutic Sciences and Department of Medicine, University of California, San Francisco, San Francisco, CA, USA. <sup>14</sup>Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, CA, USA. <sup>15</sup>Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA. <sup>16</sup>Aligning Science Across Parkinson's (ASAP) Collaborative Research Network, Chevy Chase, MD, USA. <sup>17</sup>UK Dementia Research Institute and Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK. <sup>18</sup>Neurogenetics Branch, National Institute of Neurological Disorders and Stroke, Bethesda, MD, USA. <sup>19</sup>These authors contributed equally: Pilar Álvarez Jerez, Peter Wild Crea. \*A list of authors and their affiliations appears at the end of the paper.  e-mail: [cornelis.blauwendraat@nih.gov](mailto:cornelis.blauwendraat@nih.gov)

### Global Parkinson's Genetics Program (GP2)

**Mary B. Makarios**<sup>1,7</sup>, **Oluwadamilola O. Ojo**<sup>8,9</sup>, **Mie Rizig**<sup>11</sup>, **Caroline B. Pantazis**<sup>1</sup>, **Hampton L. Leonard**<sup>1,7</sup>, **Mike A. Nalls**<sup>1</sup>, **Sara Bandres-Ciga**<sup>1</sup>, **John Hardy**<sup>3</sup>, **Henry Houlden**<sup>11</sup>, **Andrew B. Singleton**<sup>1,2</sup>, **Mina Ryten**<sup>4,16,17</sup>, **Njideka U. Okubadejo**<sup>8,9</sup> & **Cornelis Blauwendraat**<sup>1,2</sup>

A full list of members and their affiliations appears in the Supplementary Information.

## Methods

All our research complies with the relevant ethical regulations. The work was covered by local institutional review board approval at each site involved.

### Biosamples used for assessment of effects of *GBAI* rs3115534

To assess the potential molecular downstream effects of the non-coding *GBAI* variant rs3115534, we used data from GP2 data release 5 (<https://gp2.org/>, <https://doi.org/10.5281/zenodo.7904832>) and identified variant carriers of interest with matching LCLs at the Coriell Institute for Medical Research (<https://www.coriell.org/>). In addition, we accessed brain tissue samples from the HBCC with and without the noncoding *GBAI* variant (rs3115534). A full overview of included samples and their demographics is provided in Supplementary Table 1. Note that the genome reference allele (hg38) is G for rs3115534, which is also the risk allele, and T for the alternative nonrisk allele. Given that *GBAI* is transcribed from the antisense strand, C is risk and A is nonrisk on the RNA level.

### DNA extraction from cell pellets and brain tissue samples

DNA was extracted from LCLs ( $2 \times 10^6$  cells) following the Pacific Biosciences (PacBio) high-molecular-weight (HMW) DNA extraction cultured cells protocol with the Nanobind tissue kit (PacBio, 102-203-100). The extraction occurred on a KingFisher Apex System (Thermo Fisher Scientific, 5400920). For brain tissue, 40 mg of frontal cortex tissue was cut and manually homogenized with a TissueRuptor (Qiagen, 9002755) in buffer CT (PacBio, 102-280-300). Then, the DNA was extracted following the PacBio Apex Nanobind Tissue Big DNA protocol with the Nanobind tissue kit (PacBio, 102-203-100) using the KingFisher Apex System (Thermo Fisher Scientific, 5400920). The DNA from LCLs and brain tissue was quantified using the Qubit double-stranded DNA (dsDNA) BR assay (Invitrogen, Q32850) and sized with a Femto Pulse System (Agilent, M5330AA). The samples then underwent a size selection to remove fragments under 15 kb using the short-read eliminator kit (PacBio, 102-208-300). After size selection, the DNA was sheared to a target size of 30 kb using the Megaruptor 3 (Diagenode, B060100003) with Fluid+ needles (Diagenode, E07020001) at speed 45 for two cycles. DNA was then quantified and resized with the Qubit dsDNA BR assay and the Femto Pulse system. Samples needed to have at least 4.5  $\mu$ g of DNA and be 20–40 kb in size to take forward into library preparation, as previously described<sup>25,26</sup>.

### ONT DNA library preparation and sequencing

Sequencing libraries were prepared with the ONT SQK-LSK110 kit and 400 ng of prepared library per sample was loaded onto R9.4.1 flow cells on ONT's PromethION device with Minknow 22.10.7 software. Samples were sequenced over 72 h with 1–2 additional library loads per flow cell. Sequencing data resulted in an average coverage of 30 $\times$  and an N50 of around 30 kb per sample.

### RNA extraction from cell pellets and brain tissue samples

RNA was extracted from LCLs ( $5 \times 10^6$  cells) and brain tissue (40 mg) using the RNA Direct-zol miniprep kit and its corresponding protocol (Zymo Research, R2050, [https://files.zymoresearch.com/protocols/r2050\\_r2051\\_r2052\\_r2053\\_direct-zol\\_rna\\_miniprep.pdf](https://files.zymoresearch.com/protocols/r2050_r2051_r2052_r2053_direct-zol_rna_miniprep.pdf)). In short, samples were resuspended in 600  $\mu$ l of TRI reagent and an equal volume of 100% ethanol. Brain tissue needed additional homogenization with a Dounce homogenizer during the resuspension steps. Each mixture was then transferred into a Zymo-Spin IICR Column, centrifuged and transferred to a new collection tube. DNase I treatment was performed using the recommended guidelines. Washing steps were performed with Zymo Research's Direct-zol RNA prewash and RNA wash buffer according to the protocol. Finally, RNA was eluted in RNase-free water and quality control was performed on Agilent's TapeStation 4200.

All cell lines had an RIN > 9, while brain RNA had RINs between 5.1 and 8.6 (Supplementary Table 1).

### ONT cDNA library preparation and sequencing

First, 200 ng of total RNA from the LCLs and brain tissue was prepared for sequencing using ONT's cDNA-PCR SQK-PCS111 library preparation kit and protocol with modifications. This library preparation relies on an oligo(dT)-based poly(A) selection. The protocol modifications included an additional bead cleanup after RT with 11.25  $\mu$ l of RNase-free XP beads (Beckman Coulter, A63987), short fragment buffer washes (ONT, PCS111 kit) and elution into 22.5  $\mu$ l of elution buffer. PCR settings during amplification were adjusted to set the annealing steps to 12 cycles. After library preparation, total RNA was quantified using the Qubit dsDNA high-sensitivity (HS) assay it (Invitrogen, Q32851). Then, 22 fmol of prepared library was loaded onto R9.4.1 PromethION flow cells and sequenced for 72 h using the Minknow 22.10.7 software.

### Single-nucleus cDNA library preparation

Nuclei from frozen brain tissue were prepared using a modified homogenization protocol<sup>27,28</sup>. Briefly, 30–50 mg of tissue was homogenized in a Dounce homogenizer with 20 strokes of the loose pestle and 20 strokes of the tight pestle with 1 $\times$  lysis buffer (10 mM Tris-HCl (Sigma-Aldrich, T2194-1L), 10 mM NaCl (Sigma-Aldrich, 5922C-500ML), 3 mM MgCl<sub>2</sub> (Sigma-Aldrich, M1028-100ML), 0.1% Nonidet P40 substitute/IGEPAL CA-630 (Sigma-Aldrich, I8896-50ML), 1 mM DTT (Sigma-Aldrich, 646563-10X.5 ML) and 1 U per  $\mu$ l RNase inhibitors (Sigma-Aldrich, 03335402001)). Nuclei were filtered and collected by centrifugation through a sucrose cushion gradient (Sigma-Aldrich, NUC201-1KT). Myelin and debris were removed and nuclei were washed (10 mM Tris-HCl, 10 mM NaCl, 3 mM MgCl<sub>2</sub>, 1% BSA (Miltényi Biotec, 130-091-376), 0.1% Tween-20 (Bio-Rad, 1610781) and 1 mM DTT), pelleted and permeabilized in 0.1 $\times$  lysis buffer 2 (10 mM Tris-HCl, 10 mM NaCl, 3 mM MgCl<sub>2</sub>, 1% BSA, 0.1% Nonidet P40 substitute/IGEPAL CA-630 and 0.01% digitonin (Invitrogen, BN20061)), washed and counted. The resulting nuclei were then prepared using the 10x Genomics single-cell Multiome ATAC + gene expression kit (10x Genomics, 1000283) and loaded on single-cell Chip J (10x Genomics, 1000234) for the recovery of 10,000 nuclei. After Tn5 transposition, single-cell isolation, barcoding and preamplification, cDNA was generated as directed in the user guide with poly(dT) primers for RT. The resulting cDNA libraries were quantified with Qubit dsDNA HS reagents (Invitrogen, Q33231) and the average fragment size was determined using high-sensitivity DNA D5000 screentape analysis (Agilent Technologies, 5067-5593 and 5067-5592).

### Targeted transcript capture for long-read RNAseq

Next, to identify the cell type(s) expressing the short intron-containing *GBAI* transcript, we performed long-read sequencing on the single-cell cDNA libraries generated from the 10x Genomics single-cell Multiome ATAC + gene expression kit. We first targeted *GBAI* in the single-cell cDNA libraries using PacBio's customer collaboration Iso-Seq express capture using Integrated DNA Technologies (IDT) xGEN lockdown probes protocol ([https://ostr.ccr.cancer.gov/wp-content/uploads/2023/06/PacBio\\_TargetedIso-Seq.pdf](https://ostr.ccr.cancer.gov/wp-content/uploads/2023/06/PacBio_TargetedIso-Seq.pdf)) with primer modifications to tailor it to ONT sequencing<sup>10</sup>. In short, 10 ng of each single-cell cDNA library was amplified with ONT cDNA 10x primers taken from the literature<sup>29</sup> and the NEBNext single-cell, low-input cDNA synthesis and amplification module (New England Biolabs, E6421S). After amplification, the cDNA was cleaned with ProNex beads (Promega, NG2001). A total of 500 ng of cDNA was then hybridized using custom *GBAI* xGen lockdown probes (IDT) (Supplementary Table 2) and washed with the xGen lockdown hybridization and wash kit (IDT, 1080577). The captured cDNA was then amplified using the same primers as above. Full details of reagents and PCR conditions are provided in Supplementary Table 3. After the *GBAI* capture, 10 ng of cDNA was taken into library preparation using an optimized version of the ONT PCS111

kit. The cDNA first underwent a biotin tagging reaction with custom oligos and PCR amplification, after which it was cleaned using AMPure XP beads (Beckman Coulter, A63881). Next, the cDNA was bound to M280 streptavidin beads (Invitrogen, 11205D) and amplified with the ONT cDNA primer. The cDNA underwent AMPure XP bead cleaning once more and was quantified using the Qubit dsDNA HS reagents. Then, 35 fmol was taken forward, the RAP T adaptor was added and the sample underwent standard loading on an R.9.4.1 flow cell on a PromethION with Minknow 22.10.7 software.

### Whole-genome DNA long-read sequencing analysis

All DNA sequencing runs were basecalled on the National Institutes of Health (NIH) high-performance computing (HPC) cluster (Biowulf) using Guppy (version 6.1.2) in super-accuracy mode with the 'dna\_r9.4.1\_450bps\_modbases\_5mc\_cg\_sup\_prom.cfg' configuration file. Basecalled files were mapped to hg38 using Minimap2 (version 2.24/2.26)<sup>30</sup> preserving methylation tags ('samtools fastq -TMm, MI \${FASTQ\_PATH}/\${BAM\_FILE} | minimap2 -y -x map-ont -t 20 -a --eqx -k 17 -K 10 g'). SNVs were called using Clair3 (version 1.0.4)<sup>31</sup> and structural variants were called using Sniffles (version 2.2)<sup>32</sup>.

### Untargeted long-read RNAseq analysis

All RNAseq runs were basecalled on the NIH HPC cluster (Biowulf) using Guppy (version 6.1.2) in super-accuracy mode with the 'dna\_r9.4.1\_450bps\_sup\_prom.cfg' configuration file for cDNA.

Pychopper version 2.7.1 (<https://github.com/epi2me-labs/pychopper>) was run on the cDNA FASTQ files with a minimum mean read quality of 7.0 and a minimum segment length of 50 for kit SQK-PCS111. Reads passing quality control were mapped to hg38 using Minimap2 (version 2.26)<sup>30</sup> with splice-aware parameters ('-t 10 -ax splice -k14 -uf'). Only reads with a minimum mapping quality of 40 and flagged as primary alignment were kept. For transcript calling and quantification, we used StringTie2 (version 2.2.1)<sup>33</sup>. StringTie2 was first run in long-read reference free mode ('stringtie -L -R -m 50') to capture our unannotated intronic region. Then, we generated a reference GTF file with the canonical *GBAI* transcript (ENST00000368373.8) and the transcript containing part of intron 8 and reran our samples against this reference ('stringtie --rf -G \${reference\_annotation} -L -v -p 10') for quantification of the intron-containing transcript and canonical transcript. Each mapped BAM and StringTie2 annotation was then manually inspected on Integrative Genomics Viewer (IGV; version 2.16.0)<sup>34</sup>.

Additionally, we calculated regional sequencing depth for different *GBAI* regions. For this, we used SAMtools (version 1.17)<sup>35</sup> to subset the *GBAI* regions from each hg38-mapped BAM and used 'samtools coverage -q5 -Q20 --ff UNMAP,SECONDARY,QCFAIL,DUP -r \$chr:\$start-\$end \${IN}' to calculate the mean depth at each region of interest. These regions of interest included (1) exons 8 and 9; (2) intron 8; (3) the -40-bp short transcript region in intron 8; (4) intron 8 minus the -40-bp short intron 8 transcript region; and (5) *GBAI* and *GBAILP*. These coordinates were based on ENST00000368373.8 for *GBAI* and the exact BED file can be found in Supplementary Table 4. To normalize coverage metrics across samples, we extracted the number of uniquely mapped reads per sample over the whole genome using 'samtools view -c -F 260 \${IN}' and divided these reads by a million. We then divided our mean depth metric by the number of reads for a normalized coverage per million metric. Normalized coverage counts were averaged across rs3115534 genotypes and plotted using ggplot2 (ref. 36) in R (version 4.3.0).

Lastly, to check that these reads mapped uniquely to *GBAI* and not to *GBAILP*, we manually edited the rs3115534 variant in the FASTQ data and looked for differences in mapping. To do this, we subsetted the *GBAI* and *GBAILP* regions for one GG and one TT from the hg38-aligned BAM files and converted the subsets back to a FASTQ using SAMtools (version 1.17) bam2fq. Then, using Visual Studio Code (<https://github.com/microsoft/vscode>), we manually changed the

rs3115534 variant from a G to a T and vice versa and remapped the edited FASTQ. Mapped BAM files were then inspected on IGV 2.16.0.

### Transcript capture long-read RNAseq analysis

Long-read single-nucleus RNAseq data were basecalled using Guppy (version 6.1.2) with the 'dna\_r9.4.1\_450bps\_sup\_prom.cfg' configuration file. The run FASTQ was then mapped using Minimap2 (version 2.26) with splice-aware parameters to hg38, subset for *GBAI* ± 1 Mb and then the subset region reverted to a FASTQ using SAMtools (version 1.17) bam2fq. This subset FASTQ was then split into a FASTQ per cell type using corresponding Illumina unique cell type barcodes. Each barcode was matched against the FASTQ allowing for one mismatch and the resulting FASTQs were sorted to only keep unique read ids. Each cell FASTQ was quality-controlled using Pychopper (version 2.7.1) with standard parameters and mapped using Minimap2 (version 2.26) with splice-aware parameters. Only reads with a minimum mapping quality of 40 and flagged as primary alignment were kept. Then, we performed the same depth calculations as above using SAMtools (version 1.17) and used the number of cells in the original 10x library as the normalization factor for our depth per region across cell types. Regional *GBAI* and transcript coverage was plotted using ggplot2. As an additional quality control, we checked barcode sequence diversity in each of our cell types by calculating the ratio of barcodes in our data divided by total Illumina barcodes. Density plots of barcode usage were generated with ggplot2.

### Short-read sequencing data generation and processing

Illumina short-read data were generated for the same LCLs to complement the ONT data. RNA was extracted using the same method described above and library preparation and sequencing were completed by Psomagen (<https://www.psomagen.com/>). The ribosomal RNA removed total was fragmented and primed for cDNA synthesis using TruSeq stranded total RNA library prep kit reagents (96 samples; Illumina, 20020597) before incubating for 8 min at 95 °C (C1000 Touch Thermal Cycler). The cleaved and primed RNA was reverse-transcribed into first-strand cDNA using SuperScript II reverse transcriptase (Thermo Fisher Scientific, 18064-014). Actinomycin D and first-strand synthesis act D mix were added to enhance strand specificity. The second strand was synthesized using the second-strand master mix from the same TruSeq stranded total RNA kit (16 °C incubation for 1 h). To enable adaptor ligation, the dscDNA was adenylated at the 3' end and RNA adaptors were subsequently ligated to the dA-tailed dscDNA. Finally, additional amplification steps were carried out to enrich the library material. The final library was validated (D1000 ScreenTape System) and quantified (Quant-iT PicoGreen dsDNA assay kit).

The sequencing library was then loaded onto a flow cell containing surface-bound oligos complementary to the adaptors in the library and amplified into distinct clusters. Following cluster generation, the Illumina sequencing by synthesis (SBS) technology (Illumina Novaseq 1.5 5000/6000 S4 reagent kit, 300 cycles, 20028312; Illumina Novaseq 1.5 Xp 4-lane kit, 20043131) was used to accurately sequence each base pair. Real-time analysis software (RTA version 3) was used to basecall data from raw images generated by the Illumina SBS technology. The binary BCL/cBCL files were then converted to FASTQ files using bcl2fastq (bcl2fastq version 2.20.0.422), a package provided by Illumina. Illumina FASTQ data were aligned to hg38 using STAR (version 2.7.10)<sup>37</sup>. We then calculated regional depth following the same steps as for the bulk RNA ONT long-read data.

### Accessing publicly available whole-genome sequencing (WGS) and RNAseq

We additionally looked into Illumina transcriptomic data from the AMP PD (<https://www.amp-pd.org/>) and 1000 Genomes Project (<https://www.internationalgenome.org/>) datasets for homozygous reference, heterozygous and alternative allele carriers of the rs3115534 variant.

For AMP PD, we extracted data for 1 homozygous G carrier, 47 heterozygous G carriers and 98 non-European controls. The ancestry of extracted data was divided as follows: 12 African admixed (7 GT and 5 TT), 15 African (1 GG, 5 GT and 9 TT), 118 European (35 GT and 83 TT) and 1 Asian (1 TT). Details on AMP PD samples are provided in Supplementary Table 5. Full details on data generation and processing of WGS and RNAseq data were provided in previous studies<sup>38,39</sup>. Full details on ancestry predictions were provided in a previous study<sup>11</sup>. Using these files, we calculated regional depth following the same steps as for the bulk RNA ONT long-read data.

The 1000 Genomes Project WGS and RNAseq data were downloaded and samples were demultiplexed to individual sample FASTQ files and aligned to hg38 using STAR version (version 2.6.1)<sup>37</sup>. We extracted data for 7 homozygous G carriers, 40 heterozygous G carriers and 41 homozygous T carriers, all of African ancestry. Details on 1000 Genomes Project samples are provided in Supplementary Table 6. We calculated regional depth following the same steps as for the bulk RNA ONT long-read data. Additionally, we accessed African American ancestry Illumina RNAseq data from the HBCC ( $n = 92$ ). Within the 92 samples, there were 6 GG, 20 GT and 66 TT samples. These data were accessed through the National Institute of Mental Health (NIMH) Data Archive (<https://nda.nih.gov/>). FASTQ files were aligned to hg38 using STAR version (version 2.6.1)<sup>37</sup>. Details on Illumina HBCC samples are provided in Supplementary Table 7. The tissue used in this research was obtained from the HBCC Intramural Research Program (IRP; <http://www.nimh.nih.gov/hbcc>).

To calculate the importance of the *rs3115534*-G allele with respect to the depth per region, we ran linear regressions in R (version 4.3.0) with genotypes as the predictor and normalized depth as the outcome for each dataset and region. Because of the low numbers of the GG groups, we combined the GG and GT genotypes for the regression in each dataset (Illumina Coriell, ONT Coriell, ONT HBCC, ONT CRISPR and AMP PD) except for the 1000 Genomes Project and Illumina HBCC datasets, where there were >5 samples with a GG genotype. For the 1000 Genomes Project and Illumina HBCC datasets, we ran the regression with GG, GT and TT split into three groups.

### Validation of the intronic expression *GBAI* transcript

To validate the presence of the potential *GBAI* intron-containing transcript we generated custom primers that bind to the most highly expressed part of the transcript. The forward (*GBAI\_X11\_F9*, 5'-GCGACGCCACAGGTAG-3') and reverse (*GBAI\_X11\_R9*, 5'-CTTTGCCTTACCCTAGAACCCTC-3') primers specifically designed to start before exon 9 and end at exon 11 of *GBAI* were used at a final concentration of 0.8  $\mu$ M (IDT); an additional reverse primer specific to the 3' untranslated region of *GBAI* (*GBAI\_UTR\_R4*, 5'-CCTTTGCCTTACCCTAGAACC-3') was also used in conjunction with *GBAI\_X11\_F9* (Supplementary Table 8). As input, we used RNA from LCLs with and without the *rs3115534*-G allele. RNA was quantified using Qubit RNA HS assay (Invitrogen, Q32852) and reverse-transcribed to cDNA. cDNA was synthesized from RNA using a high-capacity cDNA RT kit (Thermo Fisher Scientific, 4368814) following the manufacturer's recommendations. The cDNA then underwent PCR using REDTaq ReadyMix (Millipore Sigma, R2523-20RXN). PCR products were mixed with 6 $\times$  loading dye (New England Biolabs, B7024S), loaded onto a 1% agarose gel containing SYBR safe DNA gel stain (Thermo Fisher Scientific, S33102), sized with a 1 kb plus DNA ladder (New England Biolabs, N3200L) and imaged on a ChemiDoc Imaging System (Bio-Rad, 12003153). PCR bands were excised from 1% agarose gel and DNA was purified using the NucleoSpin gel and PCR cleanup kit (Takara Bio, 740609) according to the manufacturer's instructions. Sanger sequencing was performed by Psomagen using conventional protocols (Supplementary Table 9).

In addition, we aimed to assess whether the short *GBAI* transcript is capped. To check the presence of a 5' cap, we selected three

Coriell lines (ND01137-GG, ND02892-GT and ND22789-TT) for CAGE library prep and sequencing performed by DNAFORM (<https://www.dnaform.jp/en/>). In brief, RNA quality was assessed with a Bioanalyzer (Agilent) and all samples had an RIN above 8.3. First-strand cDNAs were transcribed to the 5' ends of capped RNAs and attached to CAGE 'barcode' tags. The samples were then sequenced on an Illumina NextSeq 500 and the sequenced CAGE tags were mapped to the human hg38 genome using BWA software (version 0.5.9). Mapped BAM files were inspected for transcription start site clusters using IGV (version 2.16.0).

### CRISPR editing of *rs3115534*

To determine whether *rs3115534* is the functional variant in this GWAS locus, CRISPR editing was performed by Synthego (<https://www.synthego.com/>). CRISPR editing was performed using two LCLs (ND01137-GG and ND22789-TT) with the aim to edit both LCLs to the opposite genotype (Supplementary Table 10). In brief, cell pools were created using high-quality chemically modified synthetic single guide RNA (sgRNA) and SpCas9 transfected as RNPs to ensure high editing efficiencies without the use of any selection markers that could negatively affect cell biology. Knock-ins were generated using either single-stranded DNA or plasmid, depending on the insert size. The parental cells were electroporated with SpCas9 and target-specific sgRNA to generate the edited cell pool. Similarly, mock-transfected cell pools were made by electroporating the parental cells with SpCas9 only and confirmed to be unedited at target locus. After editing, cells (mock-transfected pools, intermediate pools and final fully edited LCLs) were processed for subsequent assays. A predesigned genotyping assay specific to *rs3115534* (Thermo Fisher Scientific, C\_57592022\_20) was used to confirm the CRISPR-edited genotypes using an allelic discrimination qPCR assay (QuantStudio 6 Pro, Applied Biosystems, A43159). Additionally, the CRISPR lines were manually inspected after long-read sequencing on IGV (v2.16.0) to confirm no accidental editing of *GBA1P*.

### Bioinformatic annotation of *rs3115534*

To assess the potential functional effect of *rs3115534*, we investigated several annotation resources. Summary statistics from the largest African PD GWAS were used to generate a LocusZoom plot using African linkage disequilibrium patterns<sup>11,40</sup>. RegulomeDB (version 2.2) was explored for *rs3115534* to assess its effect on motifs and genome accessibility<sup>41</sup>. The UCSC Genome Browser was accessed to investigate the conservation of this allele across vertebrates. To evaluate whether *rs3115534* was involved in splicing, we assessed the following algorithms: AGAIN, SpliceAI and Branchpointer<sup>42-44</sup>. Branchpointer was used in R (version 4.3.0) to evaluate the impact of *rs3115534* on branchpoint architecture. We ran *rs3115534* as the query file and calculated branchpoint predictions using the 'queryType = SNP' option. Gencode's hg38 version 44 release was used as our reference file. Branchpoint predictions were plotted through Branchpointer's plotBranchpointWindow script. In addition, we assessed all coding and noncoding *GBAI* variants ( $n = 1,625$ ) present in gnomAD (version 4) for their potential to disrupt intronic branchpoint sequences using AGAIN. To identify functional evidence for branchpoint usage, 293Flp-in cells (Thermo Fisher Scientific, R78007) were genotyped by aligning 100-nt paired-end RNAseq reads from chromatin-associated RNA (Damianov et al., unpublished) to *GBAI* exon 9 and the flanking upstream region. The alignment confirmed that these cells are homozygous for *rs3115534*-T. Similar sequence analysis indicated the presence of *rs2990223*-G variation in all reads mapping to the highly homologous *GBA1P* and these reads were kept separate. Branch site reads from 293Flp-in cells, obtained by RBM5 and SF3A3 RNPseq or RBM10 and SF3A3 IPseq, were then aligned to *GBAI* and *GBA1P*. Reads that could originate from either of these two genes were kept and analyzed separately. Branchpoint prediction was performed as described previously<sup>14</sup>.

### Assessing the protein-coding ability of short transcript

**Protein extraction.** LCLs from Coriell Biorepository were maintained in suspension with RPMI-1640 medium (Thermo Fisher Scientific, 11875093) containing 2 mmol L<sup>-1</sup> GlutaMAX (Thermo Fisher Scientific, 35050061) and 15% FBS (Thermo Fisher Scientific, A5256701) at 37 °C in 5% carbon dioxide. Protein was extracted from LCLs (5 × 10<sup>6</sup> cells) using a Tris-HCl cell lysis buffer (Cell Signaling Technology, 9803) containing a protease and phosphatase inhibitor cocktail (Cell Signaling Technology, 5872) on ice.

**Multiplex western blotting.** Protein was normalized to 30 µg and loaded into a 4–20% precast polyacrylamide gel (Bio-Rad, 4561094) before being transferred to a nitrocellulose membrane (Bio-Rad, 1704270). The membrane was blocked for 1 h (LiCor Biosciences, 927-60001) and incubated (4 °C) with primary anti-GCase (1 µg ml<sup>-1</sup> working concentration, 1:1,000 dilution; Sigma-Aldrich, polyclonal clone G4171, RRID: AB\_1078958) and anti-β-actin (1 µg ml<sup>-1</sup> working concentration, 1:1,000 dilution; Abcam, monoclonal clone mAbcam 8224, RRID: AB\_449644) antibodies on a shaker overnight. Finally, the membrane was incubated with donkey anti-rabbit (LiCor Biosciences, lot D30328-05, 926-68073, RRID: AB\_10954442) and donkey anti-mouse (LiCor Biosciences, lot D30124-05, 926-32212, RRID: AB\_621847) secondary antibodies (1:20,000 dilution for both secondary antibodies) at room temperature for 1 h before imaging (Odyssey DLx, LiCor Biosciences). The entire procedure was repeated with identical results.

**ECL western blotting.** Protein was normalized to 30 µg and loaded into a 4–20% precast polyacrylamide gel (Bio-Rad, 4561094) before being transferred to a nitrocellulose membrane (Bio-Rad, 1704270). The membrane was blocked for 1 h with 5% blotting-grade blocker (Bio-Rad, 1706404) and incubated (4 °C) with primary anti-GCase (1 µg ml<sup>-1</sup> working concentration, 1:1,000 dilution; Sigma-Aldrich, polyclonal clone G4171, RRID: AB\_1078958) antibody on a shaker overnight in 5% blotting-grade blocker. After washing with TBS-T (0.1% Tween-20), the membrane was incubated with goat anti-rabbit IgG (H + L) cross-adsorbed secondary antibody conjugated to horseradish peroxidase (HRP; 1:1,000 dilution; Invitrogen, 31462, RRID: AB\_228338) for 1 h. Following washing with TBS-T, the membrane was incubated with anti-β-actin antibody (1 µg ml<sup>-1</sup> working concentration, 1:1,000 dilution; Abcam, monoclonal clone mAbcam 8224, RRID: AB\_449644) for 1 h and then probed with goat anti-mouse IgG (H + L) cross-adsorbed secondary antibody conjugated to HRP (1:1,000 dilution; Invitrogen, 31432, RRID: AB\_228302). The membrane was then probed with Clarity Max Western ECL substrate for 1 h (Bio-Rad, 1705062). The blot was imaged on the ChemiDoc MP Imaging System (Bio-Rad, 12003153).

**Mass spectrometry analysis of western blot.** Protein extraction, normalization and gel electrophoresis were performed as detailed above. The bands were visualized with Coomassie blue stain (Bio-Rad, 1610786) and manually excised. Gel bands between 2 and 15 kDa were excised, reduced with 5 mM TCEP (Sigma-Aldrich, 580560) and alkylated with 5 mM *N*-ethylmaleimide (Sigma-Aldrich, 04259). Samples were digested with trypsin (Promega, V5280) at a 1:20 (w/w) ratio of trypsin to sample at 37 °C for 18 h. Peptides were extracted then desalted using Oasis HLB plate (Waters, WAT058951). Liquid chromatography–tandem mass spectrometry data acquisition was performed on an Orbitrap Lumos mass spectrometer (Thermo Fisher Scientific) coupled to an Ultimate 3000 high-performance liquid chromatograph. Peptides were separated on a ES902 Easy-Spray column (75-µm inner diameter, 25-cm length, 3-µm C18 beads; Thermo Fisher Scientific). Mobile phase B was increased from 3% to 20% in 39 min. Lumos was operated in data-dependent mode. Peptides were fragmented with a higher-energy collisional dissociation method at a fixed collision energy of 35. The Proteome Discoverer 2.4 software was used for

database search using the Mascot search engine. Data were searched against the SWISS-PROT human database.

**Assessing GBA1 protein expression in UK Biobank.** To replicate previous pQTL results of [rs3115534](https://doi.org/10.1038/s41594-024-01423-2), we accessed the UK Biobank 500K whole-genome sequencing data through the UK Biobank Research Analysis Platform (<https://ukbiobank.dnaxexus.com/>). We used the population-level variant data produced using Illumina DRAGEN version 3.7.8. Genotypes for [rs3114435](https://doi.org/10.1038/s41594-024-01423-2) were extracted using Plink (version 2.0; <https://www.cog-genomics.org/plink/2.0/>) and only those individuals with African or African admixed ancestry were kept. We then merged the genotype information with available Olink proteomic measures to get the GBA1 protein level per individual. Any unavailable data from the protein information were dropped. We then generated a violin plot of GBA1 protein expression per genotype using ggplot2 (ref. 26) in R (version 4.3.0). We ran a linear regression in R (version 4.3.0) with genotypes as the predictor and GBA1 protein levels as the outcome to test for significance.

### Assessing GCase activity across GBA1 genotypes

GCase activity was assessed across *GBA1* genotypes in 710 samples containing 97 non-PD (used as controls), 122 idiopathic PD without a known *GBA1* mutation, 95 *LRRK2* p.G2019S carriers, 99 [rs3115534](https://doi.org/10.1038/s41594-024-01423-2)-GT, 9 [rs3115534](https://doi.org/10.1038/s41594-024-01423-2)-GG and 99 PD risk variant carriers (p.E365K or p.T408M), 90 *GBA1* p.N409S (mild) and 99 *GBA1* severe mutations (for example, *GBA1* p.L483P). Exonic *GBA1* mutations are written on the basis of current recommendations, which include the 39-aa signal peptide at the start of the protein (for example, p.E365K = p.E326K; p.T408M = p.TM; p.N409S = p.N370S; p.L483P = p.L444P). The GCase enzyme was extracted from one dried blood spot punch (Ø 3.2 mm) per sample and measured by incubating for 1 h at 37 °C under agitation with an aqueous buffer containing citrate, phosphate, taurocholic acid sodium salt, NaN<sub>3</sub> and Triton X-100 (pH 5.2). Next, an aqueous solution with the synthetic substrate 4-methylumbelliferyl β-D-glucopyranoside and NaN<sub>3</sub> was added followed by a second incubation step for 16 h at 37 °C under agitation. The enzymatic reaction was quenched by addition of stop buffer (aqueous glycine solution, pH 10.5 adjusted by NaOH). The enzymatic product, 4-methylumbelliferone was quantified by fluorimetry on a microplate reader (Victor X2, PerkinElmer). The instrument was calibrated using an external calibration curve. The enzymatic activity is specified in units of µmol L<sup>-1</sup> h<sup>-1</sup> (amount of product per blood volume per incubation time). Each sample was measured in duplicate and a third replicate was used for background correction. The background of the chemical blank was determined by the addition of stop buffer before the substrate. As quality parameters for the assay, standard blood samples were added to each batch to ensure the accuracy of the determination. To analyze the data, we ran a linear regression in R (version 4.3.0) with genotypes as the predictor and GCase activity levels as the outcome to test for significance. We also compared each group to each other with a one-sided *P* value test in R.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

Unedited Coriell LCL lines are available online (<https://www.coriell.org/>). CRISPR-edited Coriell LCL lines are available upon request and the establishment of a material transfer agreement (MTA) with Coriell and NIH and CARD abiding by the Coriell NINDS Human Genetics Repository MTA for biospecimens. All generated LCL Coriell ONT DNAseq, CAGEseq and RNAseq data (Illumina and ONT) are available online (<https://www.amp-pd.org/>) through GP2 tier 2 access, which is obtainable by filling in the form (<https://www.amp-pd.org/researchers/data-use-agreement>). It is part of GP2 release 7

(<https://doi.org/10.5281/zenodo.10962119>; [https://console.cloud.google.com/storage/browser/gp2tier2/release7\\_30042024/gp2\\_omics/Alvarez\\_Jerez\\_et\\_al\\_2024](https://console.cloud.google.com/storage/browser/gp2tier2/release7_30042024/gp2_omics/Alvarez_Jerez_et_al_2024); data path: gp2tier2/release7\_30042024/gp2\_omics/Alvarez\_Jerez\_et\_al\_2024). AMP PD Illumina blood-based RNAseq data are available online (<https://www.amp-pd.org/>) after signing the data use agreement. The 1000 Genomes Project data are publicly available online (<https://www.internationalgenome.org/>). Brain tissue bulk RNAseq data ([https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000979.v3.p2](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000979.v3.p2)) and frontal cortex data ([https://nda.nih.gov/edit\\_collection.html?id=3151](https://nda.nih.gov/edit_collection.html?id=3151)) are available online. Summary statistics for *cis*-eQTLs and a catalog of ancestry-specific eQTLs were obtained from Kachuri et al.<sup>12</sup> (<https://doi.org/10.5281/zenodo.7735723>).

## Code availability

All scripts and code for this project, including all the identifiers and version numbers for software used, can be found on GitHub (<https://github.com/GP2code/GBA1-rs3115534-branchpoint>) and Zenodo (<https://doi.org/10.5281/zenodo.10484208>) (ref. 45).

## References

- Cogan, G., Jerez, P. A., Malik, L., Blauwendraat, C. & Billingsley, K. J. Processing frozen cells for population-scale SQK-LSK114 Oxford Nanopore long-read DNA sequencing SOP. *protocols.io* <https://doi.org/10.17504/protocols.io.6qpvr347bvmk/v1> (2023).
- Baker, B. et al. Processing human frontal cortex brain tissue for population-scale Oxford Nanopore long-read DNA sequencing SOP. *protocols.io* <https://doi.org/10.17504/protocols.io.kxygx3zzog8j/v1> (2023).
- Morabito, S., Miyoshi, E., Michael, N. & Swarup, V. Integrative genomics approach identifies conserved transcriptomic networks in Alzheimer's disease. *Hum. Mol. Genet.* **29**, 2899–2919 (2020).
- Shi, Z., Das, S., Morabito, S., Miyoshi, E. & Swarup, V. Protocol for single-nucleus ATAC sequencing and bioinformatic analysis in frozen human brain tissue. *STAR Protoc.* **3**, 101491 (2022).
- Lebrigand, K., Magnone, V., Barbry, P. & Waldmann, R. High throughput error corrected Nanopore single cell transcriptome sequencing. *Nat. Commun.* **11**, 4025 (2020).
- Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
- Zheng, Z. et al. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nat. Comput. Sci.* **2**, 797–803 (2022).
- Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
- Kovaka, S. et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
- Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
- Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
- Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* 1st edn (Springer, 2009).
- Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- Craig, D. W. et al. RNA sequencing of whole blood reveals early alterations in immune cells and gene expression in Parkinson's disease. *Nat. Aging* **1**, 734–747 (2021).
- Iwaki, H. et al. Accelerating Medicines Partnership: Parkinson's disease. Genetic resource. *Mov. Disord.* **36**, 1795–1804 (2021).
- Boughton, A. P. et al. LocusZoom.js: interactive and embeddable visualization of genetic association study results. *Bioinformatics* **37**, 3017–3018 (2021).
- Dong, S. et al. Annotating and prioritizing human non-coding variants with RegulomeDB v.2. *Nat. Genet.* **55**, 724–726 (2023).
- Signal, B., Gloss, B. S., Dinger, M. E. & Mercer, T. R. Machine learning annotation of human branchpoints. *Bioinformatics* **34**, 920–927 (2018).
- Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548 (2019).
- Zhang, P. et al. Genome-wide detection of human intronic AG-gain variants located between splicing branchpoints and canonical splice acceptor sites. *Proc. Natl Acad. Sci. USA* **120**, e2314225120 (2023).
- Makarious, M. B. GP2code/GBA1-rs3115534-branchpoint: Initial Release. *Zenodo* <https://zenodo.org/records/10484209> (2024).

## Acknowledgements

We thank all of the participants who donated their time and biological samples to be a part of this study. This work used the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>). We also thank the Coriell cell repository and 1000 Genomes Project; without their valuable resources including the genetic diversity presence, this work would not be possible. This work was supported in part by the IRP of the NIH including the Center for Alzheimer's and Related Dementias (CARD), within the IRP of the National Institute on Aging and the National Institute of Neurological Disorders and Stroke (NINDS) (1ZIAAG000538-03, ZIAAG000542-01 and 1ZIAAG000543-01, to A.B.S. and C. Blauwendraat). GP2 is funded by the Aligning Science Across Parkinson's (ASAP) initiative and implemented by The Michael J. Fox Foundation for Parkinson's Research (MJFF) (<https://gp2.org>). This research was funded in part by ASAP MJFF-024547 (to P.B. and C. Beetz) through the MJFF. K.D. reports receiving grants from the KSPS Research Fellowship for Japanese Biomedical and Behavioral Researchers at the NIH. Branch site discovery in 293Flp-in cells was supported by NIH grant R01GM127473 (A.D.). The tissue used in this research was obtained from the HBCC IRP (<http://www.nimh.nih.gov/hbcc>), supported by project ZIC MH002903. Protein QTL data were generated using National Institute of Diabetes and Digestive and Kidney Diseases grant R01DK108803 (to A.S.). This project was supported by GP2. GP2 is funded by the ASAP initiative and implemented by the MJFF (<https://gp2.org>). A complete list of GP2 members can be found online (<https://gp2.org>). J.H., H.R.M. and M.R. report receiving funding in part by ASAP (000478). This research was conducted using the UK Biobank Resource under application no. 33601. Data used in the preparation of this article were obtained from the AMP PD Knowledge Platform. Up-to-date information on the study is available online (<https://www.amp-pd.org>). The AMP PD program is a public-private partnership managed by the Foundation for the NIH and funded by the NINDS in partnership with the ASAP initiative, Celgene Corporation (a subsidiary of Bristol-Myers Squibb Company), GlaxoSmithKline (GSK), the MJFF, Pfizer, AbbVie, Sanofi US Services and Verily Life Sciences. 'Accelerating Medicines Partnership' and 'AMP' are registered service marks of the US Department of Health and Human Services. The generation of molecular data for the TOPMed program was supported by the National Heart, Lung and Blood Institute (NHLBI). RNAseq for the NHLBI TOPMed Genes-Environments and Admixture in Latino Asthmatics Study (GALA II; phs000920) and Study of African Americans, Asthma, Genes and Environments (SAGE; phs000921) was performed at the Broad Institute Genomics Platform (HHSN2682016000341). WGS for the same studies was performed at the New York Genome Center (NYGC; 3R01HL117004-02S3) and Northwest Genomics Center (HHSN2682016000321). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering, was provided by the TOPMed Informatics Research Center (3R01HL117626-02S1; contract HHSN2682018000021). Core support including phenotype harmonization, data management, sample identity quality control

and general program coordination was provided by the TOPMed Data Coordinating Center (R01HL-120393 and U01HL-120393; contract HHSN268201800001). WGS as part of GALA II was performed by the NYGC under a grant from the Centers for Common Disease Genomics of the Genome Sequencing Program (GSP) (UM1 HG008901). The GSP coordinating center (U24 HG008956) contributed to cross-program scientific initiatives and provided logistical and general study coordination. The GSP is funded by the National Human Genome Research Institute, NHLBI and National Eye Institute. This work and E.G.B. were supported in part by the Sandler Family Foundation, American Asthma Foundation, Robert Wood Johnson Foundation Amos Medical Faculty Development Program, Harry Wm. and Diana V. Hind Distinguished Professor in Pharmaceutical Sciences II, NHLBI (R01HL117004, R01HL135156, X01HL134589 and U01HL138626), National Institute of Environmental Health Sciences (R01ES015794), National Institute on Minority Health and Health Disparities (R56MD013312 and P60MD006902), Tobacco-Related Disease Research Program (24RT-0025 and 27IR-0030) and National Human Genome Research Institute (U01HG009080). The Parkinson's Progression Markers Initiative (PPMI) is sponsored by the MJFF and supported by a consortium of scientific partners: 4D Pharma, AbbVie, AcureX Therapeutics, Allergan, Amathus Therapeutics, ASAP, Avid Radiopharmaceuticals, Bial Biotech, Biogen, BioLegend, BlueRock Therapeutics, Bristol-Myers Squibb, Calico Life Sciences, Celgene Corporation, DaCapo Brainscience, Denali Therapeutics, The Edmond J. Safra Foundation, Eli Lilly and Company, Gain Therapeutics, GE Healthcare, GSK, Golub Capital, Handl Therapeutics, Insitro, Janssen Pharmaceuticals, Lundbeck, Merck & Co., Meso Scale Diagnostics, Neurocrine Biosciences, Pfizer, Piramal Imaging, Preval Therapeutics, F. Hoffmann-La Roche and its affiliated company Genentech, Sanofi Genzyme, Servier, Takeda Pharmaceutical Company, Teva Neuroscience, UCB, Vanqua Bio, Verily Life Sciences, Voyager Therapeutics and Yumanity Therapeutics. The PPMI investigators did not participate in reviewing the data analysis or content of the manuscript. Up-to-date information on the study is available online ([www.ppmi-info.org](http://www.ppmi-info.org)). The PD Biomarker Program (PDBP) consortium is supported by the NINDS at the NIH. A full list of PDBP investigators can be found online (<https://pdbp.ninds.nih.gov/policy>). The PDBP investigators did not participate in reviewing the data analysis or content of the manuscript. We thank Synthego for the CRISPR editing and Psomagen for RNAseq of the Illumina libraries and Sanger sequencing. We thank B. Fubara and Y. Li at the NINDS Proteomics Core Facility for Protein Identification Analysis for efficiently performing the mass spectrometry analysis. We also thank

H. Macpherson from the University of College of London for guidance during the *GBA1* single-nucleus capture protocol.

The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

## Author contributions

The study was conceptualized by P.A.J., P.A.W.C., C. Beetz, M.W., N.U.O. and C. Blauwendraat. Experiments were performed by P.A.J., P.A.W.C., D.M.R., M.R., A.D., K.J.B., L.M., S.B., C.H., Z.S., J.S., Y.A.Q. and C.E. Data were provided by A.L.S., H.L.L., L.K., C. Beetz, P.B. and S.F. Data were analyzed by P.A.J., P.A.W.C., D.R., E.K.G., M.R., M.B.M., O.O.O., K.J.B., L.M., K.D., S.B., C.H., Z.S., A.L.S., J.S., M.Z., H.R.M., C.B.P., H.L.L., L.S., Y.A.Q., M.A.N., S.B.C., J.H., H.H., C.E., E.G.B., L.K., A.B.S., S.F., P.B., X.R., M.R., C. Beetz, M.W., N.U.O., C. Blauwendraat and A.D. The original draft was written by P.A.J., P.A.W.C. and C. Blauwendraat. The manuscript was reviewed and edited by P.A.J., P.A.W.C., D.R., E.K.G., M.R., M.B.M., O.O.O., K.J.B., L.M., K.D., S.B., C.H., Z.S., A.L.S., J.S., M.Z., H.R.M., C.B.P., H.L.L., L.S., Y.A.Q., M.A.N., S.B.C., J.H., H.H., C.E., E.G.B., L.K., C.H.L., D.L.B., A.B.S., S.F., P.B., X.R., M.R., C. Beetz, M.W., N.U.O. and C. Blauwendraat.

## Competing interests

M.M.B.'s, H.L.'s and M.A.N.'s participation in this project was part of a competitive contract awarded to Data Tecnica International by the NIH to support open science research. M.A.N. also currently serves on the scientific advisory board for Character Bio and is a scientific founder at Neuron23. M.R., S.F., C. Beetz and P.B. are employees of Centogene. The remaining authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41594-024-01423-2>.

**Correspondence and requests for materials** should be addressed to Cornelis Blauwendraat.

**Peer review information** *Nature Structural & Molecular Biology* thanks Christos Proukakis and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Dimitris Typas, in collaboration with the *Nature Structural & Molecular Biology* team.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).



## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

All scripts and code for this project can be found at: <https://github.com/GP2code/GBA1-rs3115534-branchpoint>  
Tools used for data collection involve:  
Minknow 22.10.7  
RTA v3  
Illumina NextSeq500  
QuantStudio 6 Pro

Data analysis

All scripts and code for this project can be found at: <https://github.com/GP2code/GBA1-rs3115534-branchpoint>  
Tools used for data analysis involve:  
Guppy 6.1.2  
Minimap2 2.24  
Minimap2 2.26  
Sniffles 2.2  
Samtools 1.17  
Clair3 1.0.4  
PyChopper 2.7.1  
Stringtie 2.2.1  
IGV 2.16.0  
R 4.3.0  
ggplot2  
Virtual Studio Code

bcl2fast1 2.20.0.422  
 STAR 2.7.10  
 STAR 2.6.1  
 BWA 0.5.9  
 RegulomeDB 2.2  
 AGAIN  
 SpliceAI  
 Branchpointer 4.3.0  
 Proteome Discoverer 2.4  
 Illumina DRAGEN 3.7.8  
 Plink 2.0

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Unedited Coriell LCL lines are available at <https://www.coriell.org/>. CRISPR edited Coriell LCL lines are available upon request and establishment of an MTA with Coriell and NIH/CARD abiding by the Coriell NINDS Human Genetics Repository Material Transfer Agreement For Biospecimens. All generated LCL Coriell ONT DNaseq, CAGEseq and RNAseq data (ILM and ONT) is available at <https://www.amp-pd.org/> via GP2 tier 2 access which is obtainable via filling in the form <https://www.amp-pd.org/researchers/data-use-agreement>. It is part of the following release DOI 10.5281/zenodo.10962119; release 7; [https://console.cloud.google.com/storage/browser/gp2tier2/release7\\_30042024/gp2\\_omics/Alvarez\\_Jerez\\_et\\_al\\_2024](https://console.cloud.google.com/storage/browser/gp2tier2/release7_30042024/gp2_omics/Alvarez_Jerez_et_al_2024). Additionally, data path is as follows: gp2tier2/release7\_30042024/gp2\_omics/Alvarez\_Jerez\_et\_al\_2024. AMP-PD ILM blood based RNAseq is available at <https://www.amp-pd.org/> after signing the data use agreement. 1000 Genomes project data is publicly available at <https://www.internationalgenome.org/>. Brain tissue bulk RNAseq is available at [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000979.v3.p2](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000979.v3.p2) and frontal cortex data at [https://nda.nih.gov/edit\\_collection.html?id=3151](https://nda.nih.gov/edit_collection.html?id=3151). Summary statistics for cis-eQTLs and a catalog of ancestry-specific eQTLs from Kachuri et al.12 were obtained from <https://doi.org/10.5281/zenodo.7735723>.

## Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

Sex was used exclusively in this study, and was either self-reported or confirmed through genetic data analysis. No conclusions of this manuscript pertain to only one sex nor where any sex-specific analyses performed.

Reporting on race, ethnicity, or other socially relevant groupings

Ancestries referenced in this study were determined through genetic data analysis and are specified throughout this paper.

Population characteristics

Population characteristics, where available, are specified in the supplementary information for each study cohort. In summary:

LCL Coriell samples included 9 females and 9 males with an average age of 60 yo.  
 ONT HBCC Samples included 3 females and 6 males with an average age of 38 yo.  
 Illumina HBCC samples included 33 females and 59 males with an average age of 44 yo.  
 AMP-PD samples included 65 females and 81 males with an average age of 62yo.  
 1000 Genomes samples included 48 females and 40 males. Age is not available for these samples due to consent.

All samples are of African, African Admixed, or African American ancestry with the exception of 118 AMP-PD samples that are European.

Recruitment

No specific participant recruitment was utilized in this study. We accessed publicly available data or biosamples from entities such as the UK Biobank, 1000Genomes, AMP-PD, Coriell Institute of Medicine, or the Human Brain Collection Core.

Ethics oversight

All our research complies with the relevant ethical regulations. The work is covered by local IRB approval at each site involved.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences  Behavioural & social sciences  Ecological, evolutionary & environmental sciences

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were determined by availability of samples with different with different rs3115534 genotypes. Sample sizes were also determined by available genetic data in samples of African or African Admixed ancestry.
Data exclusions	Samples were excluded based on ancestry where available. The only exception was the inclusion of European samples in AMP-PD data to increase the control group. Once samples were picked based on ancestry no other exclusions were applied.
Replication	Conclusions drawn for rs3115534 investigation in available biosamples were replicated in multiple publicly accessible datasets all of which are detailed in manuscript. For original finding of intron 8 retention in the LCL ONT data, that was replicated in LCL Illumina, CRISPR ONT sequencing, AMP-PD samples, frontal cortex ONT sequencing, and 1000Genomes cohort. Single cell sequencing was only performed once. The CRISPR PCR was run twice. Sanger sequencing was only performed once. Wester blot experiments were run three times with both ECL and fluorescent methods.
Randomization	Samples were allocated to experimental groups based on their rs3115534 genotype for all analyses.
Blinding	Blinding was not relevant to our study, as we had to choose samples with specific genotypes.

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants

## Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Antibodies

Antibodies used	Primary glucocerebrosidase (1 ug/mL working concentration, 1:1000 dilution, Sigma-Aldrich, Polyclonal Clone G4171) $\beta$ -actin (1 ug/mL working concentration, 1:1000 dilution, Abcam, Monoclonal Clone mAbcam 8224) Donkey anti-Rabbit (LiCor Biosciences, Lot No. D30328-05, 926-68073, 1:20,000 dilution) secondary antibody Donkey anti-Mouse (LiCor Biosciences, Lot No. D30124-05, 926-32212, 1:20,000 dilution) secondary antibody Goat anti-Rabbit IgG (H+L) Cross-Adsorbed Secondary Antibody, HRP (1:1000 dilution, Invitrogen, 31462) Goat anti-Mouse IgG (H+L) Cross-Adsorbed Secondary Antibody, HRP (1:1000 dilution, Invitrogen, 31432)
Validation	Per Sigma Aldrich, primary glucocerebrosidase antibody (G4171, Sigma Aldrich) was validated as follows: "Enhanced antibody validation is an assay or method that provides researchers with additional assurance that the antibody specificity for the target antigen agrees with previously defined expression data". Additionally, guide on antibody validation for Sigma Aldrich can be found here: <a href="https://www.sigmaaldrich.com/US/en/technical-documents/technical-article/protein-biology/elisa/antibody-standard-validation">https://www.sigmaaldrich.com/US/en/technical-documents/technical-article/protein-biology/elisa/antibody-standard-validation</a> Per Abcam, primary $\beta$ -actin (mAbcam 8224, Abcam) was validated and tested by the manufacturer in multiple cell lines with consistent and accurate results. Data from experiments performed by the manufacturer is available here: <a href="https://www.abcam.com/en-us/products/primary-antibodies/beta-actin-antibody-mabcam-8224-loading-control-ab8224#application=wb">https://www.abcam.com/en-us/products/primary-antibodies/beta-actin-antibody-mabcam-8224-loading-control-ab8224#application=wb</a>

## Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)	All cell lines came from the Coriell Institute for Medical Research. Sex for each line is reported in Supplementary Tab:e 1 of the manuscript. In Summary, 9 were male and 9 were female. The 293Flp-In cells are a commercially available: <a href="https://www.thermofisher.com/order/catalog/product/R78007">https://www.thermofisher.com/order/catalog/product/R78007</a> . These are derived from human HEK293, which is of female origin: <a href="https://www.synthego.com/hek293#:~:text=HEK293%20is">https://www.synthego.com/hek293#:~:text=HEK293%20is</a>
---------------------	---

%20a%20hypotriploid%20human,genome%2C%20which%20displays%20cytogenetic%20instability.

Authentication

Cell line authentication and QC was done by Coriell Institute for Medical Research. Additionally, Coriell cell lines underwent qPCR to confirm the annotated genotype at rs3115534.

For the 293Flp-In cells, they are periodically tested as describe here Damianov, A. et al. The splicing regulators RBM5 and RBM10 are subunits of the U2 snRNP engaged with intron branch sites on chromatin. Mol. Cell 84, 1496–1511.e7 (2024).

Mycoplasma contamination

All cell line checks were performed by the manufacturers and tested negative for mycoplasma contamination.

Commonly misidentified lines  
(See [ICLAC](#) register)

No commonly misidentified lines were used in the study.

## Plants

Seed stocks

NA

Novel plant genotypes

NA

Authentication

NA