Reply to: Is Gauchian genotyping of GBA1 variants reliable?

Check for updates

Marco Toffoli [1,2] ✉, Anthony H. V. Schapira [1,2], Fritz J. Sedlazeck [3] & Christos Proukakis [1,2] ✉

We recently described two methods for *GBA1* analysis, which is hampered by the adjacent highly homologous pseudogene: Gauchian, a novel algorithm for analysis of short-read WGS, and targeted long-read sequencing[1]. Tayebi et al. have applied the former to 95 individuals, and compared it to Sanger sequencing[2]. They reported correct genotypes in 84, while 11 had discrepant calls. In addition, they report false Gauchian calls in 1000 Genomes Project (1 kGP) samples. Gauchian was developed because the homology of the *GBA1* region requires a variant caller that does not rely solely on the read alignments, and can identify specific variants known to be pathogenic. To understand the cause of these discrepancies, we reviewed their data and concluded that they are misinterpreting Gauchian results in 8 of the 11 discrepant samples and incorrectly using Gauchian to analyze low-coverage 1kGP samples.

Among the 11 (11.5%) samples with inconsistent calls with Sanger (Table 1), four (Pat_08, Pat_26, Pat_28, and Pat_58) were not called as the variants are not on Gauchian's target variant list, which includes all ClinVar variants in December 2021. These variants, and any others, can be easily added (see Supplementary Information). Three other samples (Pat_75, Pat_76, and Pat_79) had low data quality, resulting in large variation in sequencing depth across the genome, as shown by the median absolute deviation (MAD) of genome coverage: 0.269, 0.128, and 0.127 (three highest values among all samples). Gauchian recommends trusting calls in samples with MAD values < 0.11, and produces a warning message if this is exceeded. In all three samples, the *GBA1 + GBAP1* copy number was a no-call (marked as "None" in the output file), indicating that Gauchian could not determine the copy number due to high coverage variation. Variants were not called because no further analysis was done beyond copy number calling. These should not be viewed as false negatives, as the warning message and the report of no-calls should prompt the user to obtain higher-quality data or consider alternative sequencing. Tayebi et al. then mention that when one of the three cases with no-calls was aligned to hg38, Gauchian no longer reported a no-call ("None"), but the different call in hg38 was due to a different user error (wrong alignment settings that led to low MAPQs throughout the region, which then led to incorrect copy number calls by Gauchian, see discussion on hg19 vs. hg38 below). Moreover, they state that this test demonstrated that the sequencing depth was adequate, while the issue is high coverage variation instead of low depth. Among the remaining four samples with inconsistent results: Pat_03 had a Gauchian call of

heterozygosity for p.Asn409Ser (traditionally referred to as N370S), while Sanger reports this as homozygous. A review of the IGV trace (Tayebi et al. Supp Fig. 1) shows that at least 10 reads (around a fifth of the total) have the reference base, and therefore it is hard to conclude this is homozygous. A review of the Sanger trace (not provided) could determine whether there is a low peak representing the reference allele. We cannot provide a conclusion, and additional analysis is recommended. Mosaicism could be a plausible explanation, and this has been reported in *GBA1*[3,4], albeit not at this position. Pat_47 had a false negative p.Leu483Pro call. Pat_16 was indeed wrongly genotyped as homozygous for p.Asn409Ser, related to the adjacent c.1263del+RecTL deletion. Pat_92 had all expected variants called, but the heterozygous p.Asp448His was mis-genotyped as homozygous. In summary, there is one false negative and two wrongly genotyped variants (heterozygous variants called homozygous). Gauchian's precision is, therefore, 98.9% (175 out of 177 calls are correct). Its allele-level recall/sensitivity is 99.4% after excluding alleles not on Gauchian's target list and samples that could not be analyzed due to high coverage variation. Alternatively, it can be calculated as 97.2% if only samples with high coverage variation are excluded, 96.2% if only alleles not on the target list are excluded, and 94.1% if all these samples are considered. We note that the precision calculated by Tayebi et al. is significantly lower (93.7%) as they considered a false negative for an expected variant allele as a false positive for a "negative WT" allele, thus double counting false negatives as false positives (see their Supplementary Table 2).

Tayebi et al. concluded that Gauchian is not able to call recombinant variants without providing orthogonal evidence. In Pat_95, Pat_71, and Pat_16, they examined alignments in IGV and reported the absence of supporting reads for Gauchian calls, but all recombinant alleles called by Gauchian were consistent with Sanger. This highlights that read mapping in this region is unreliable (variant supporting reads may align to the pseudogene), making interpretation of alignments in IGV very challenging. Gauchian is designed to untangle ambiguous alignments, locally phase haplotypes and make correct calls. Particularly, in Pat_95, they claimed that Gauchian called the expected RecNciI variant but got the mechanism of the recombinant allele wrong (gene conversion vs. gene fusion). This claim appears to be based on an incorrect interpretation of IGV alignments, i.e., seeing 3' UTR mismatches associated with *GBAP1* does not necessarily indicate gene fusion, as they can be misalignments or even part of the gene

[1]Department of Clinical and Movement Neurosciences, Queen Square Institute of Neurology, University College London, London, UK. [2]Aligning Science Across Parkinson's (ASAP) Collaborative Research Network, Chevy Chase, MD, USA. [3]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. ✉e-mail: m.toffoli@nhs.net; c.proukakis@ucl.ac.uk

Table 1 | Details on the 11 samples where Gauchian and Sanger are inconsistent

Sample	Gauchian calls			Sanger Assessment, Tayebi et al.		Our assessment
	Copy Number of *GBA1* and *GBAP1*	*GBAP1*-like variant in exons 9-11	Other unphased variants	Genotype	Prediction	
Pat_08	4	None	p.Asn409Ser	p.Asn409Ser/p.Gln389Ter	False Negative	Missed variant Gln389Ter is not on Gauchian's target list
Pat_28	4	None	p.Arg535His	p.Arg535His/Cys381Tyr	False Negative	Missed variant Cys381Tyr is not on Gauchian's target list
Pat_58	4	None	p.Asn409Ser, p.Arg296Ter	p.Asn409Ser, p.Arg296Ter, c.203delC	False Negative	Missed variant c.203delC is not on Gauchian's target list
Pat_26	4	None	p.Asn409Ser	p.Asn409Ser/p.Arg502Cys	False Negative	Missed variant Arg502Cys (should be Arg502His) is not on Gauchian's target list. Tayebi et al.'s Supplementary Fig. 1 shows no variant at p.Arg502Cys (c.1504 C > T), but a different variant at the neighboring position, p.Arg502His (c.1505 G > A), which is not on Gauchian's target list.
Pat_75	None (No Call)	NA	NA	p.Arg502Cys/p.Arg159Trp	Missed	Copy number is a no-call due to high variation in depth so no further variant calling was performed. Coverage MAD 0.269
Pat_76	None (No Call)	NA	NA	p.Asn409Ser/p.Asn409Ser	Missed	Copy number is a no-call due to high variation in depth so no further variant calling was performed. Coverage MAD 0.128
Pat_79	None (No Call)	NA	NA	p.Leu483Pro/p.Arg502Cys	Missed	Copy number is a no-call due to high variation in depth so no further variant calling was performed. Coverage MAD 0.127
Pat_03	4	None	p.Asn409Ser	p.Asn409Ser/p.Asn409Ser	False Negative	Gauchian call is supported by reads, see Tayebi et al.'s Supplementary Fig. 1.
Pat_47	4	None	p.Asn409Ser	p.Asn409Ser/p.Leu483Pro	False Negative	True false negative
Pat_16	3	c.1263del+RecTL/	p.Asn409Ser, p.Asn409Ser	p.Asn409Ser, c.1263del+RecTL	False Positive	Heterozygous p.Asn409Ser misgenotyped as homozygous as Gauchian did not know the exact breakpoint of the c.1263del+RecTL deletion, which is very close to p.Asn409Ser.
Pat_92	7	p.Asp448His/ p.Leu483Pro,p.Asp448His		p.Asp448His/ p.Leu483Pro+Rec7	False Negative	There is no false negative. Rec7 is reflected in the copy number call (copy number gain). This *GBAP1* duplication does not have any functional impact on *GBA1*, so Gauchian does not report it as a *GBA1* variant. Heterozygous p.Asp448His misgenotyped as homozygous.

conversion. The RecNciI in Pat_95 is a gene conversion, as indicated by the normal copy number between *GBAP1* and *GBA1*. Tayebi et al. claimed that this is a gene fusion without orthogonal evidence. In addition, they claimed that Gauchian misreported copy numbers in Pat_92, Pat_42, and Pat_72, again without orthogonal evidence. We validated Gauchian copy number gains by digital PCR in four cases[1]. While particular recombinants could be prone to erroneous copy number calling, we do not know what "other techniques" identified a different copy number in Pat_92. Orthogonal validation using digital PCR would resolve this. Finally, it is true that Gauchian does not have all possible recombinants on its target list, as it is designed to focus on recombinant variants in exons 9–11, because others are rare and detectable with standard callers.

Tayebi et al. reported 4 samples where Gauchian missed variants in GRCh38 compared to GRCh37. Among these, two (Pat_35, Pat_75) were due to incorrect alignment settings that resulted in abnormally low mapping quality throughout the region in the GRCh38 BAMs, leading to incorrect copy number calls by Gauchian. It is likely that ALT-aware alignment was not turned on for these two samples when aligned to GRCh38. So the inconsistency in the calls was not due to Gauchian. The remaining two (Pat_16, Pat_78) reflected an area of improvement for Gauchian to better call p.Asn409Ser, which is not a *GBAP1*-like variant, and can thus be called well by standard callers.

We reported Gauchian calls of 1000 Genomes Project (1kGP) samples, validating some by targeted long reads[1]. Gauchian called zero samples with a biallelic variant in exons 9–11. However, Tayebi et al. reported a completely different set of Gauchian calls in the same samples (in their Supplemental Table 4). This was caused by incorrect use of Gauchian on old low coverage WGS (median coverage <10×, https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/), rather than 30× (https://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/1000G_2504_high_coverage/data/). Indeed, when they re-ran the analysis on the high-coverage dataset, no biallelic calls were detected. With respect to the suggestion for an input filter to prevent unsuitable data from being processed, noisy data will already result in a no-call ("None") and, therefore, leave no ambiguity. For the issue of coverage, we have added to GitHub documentation a clear warning that coverage should be ≥30, which would be a standard QC metric for data before downstream analysis is attempted.

We are grateful to Tayebi et al. for assessing the Gauchian analysis of this very challenging gene[2] but note that most discrepancies were due to incorrect use or misinterpretation of results. "No call" samples due to inadequate data quality cannot be considered false negative, as no calls are provided, and warnings of noisy coverage are given where applicable. Samples with inadequate coverage should obviously be avoided, as Gauchian is expected to perform at coverage >30×. Gauchian does not call variants not on its target list, which can be expanded. We provide updated recall (99.4%) and precision (98.9%) values. We have not seen any evidence of the alleged inability of Gauchian to call recombinant variants and would welcome orthogonal copy number assessment of discrepancies. We show that Gauchian can be used for *GBA1* assessment when coverage and data quality are adequate. We do note a limitation in genotyping p.Asn409Ser, a non-recombinant variant that can be called by standard variant callers, which we recommend running together with Gauchian for a complete call set. Finally, in clinical cases where absolute certainty is required, Sanger sequencing could be considered, with targeted long-read sequencing as another option[1,5–7].

Data availability
Gauchian and its documentation can be downloaded at https://github.com/Illumina/Gauchian. All additional data and material used have been reported and provided in the original paper.

References
1. 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).
2. Tayebi, N., Lichtenberg, J., Hertz, E. & Sidransky, E. Is Gauchian genotyping of GBA1 variants reliable? *medRxiv* https://doi.org/10.1101/2023.10.26.23297627 (2023).
3. Filocamo, M. et al. Somatic mosaicism in a patient with Gaucher disease type 2: implication for genetic counseling and therapeutic decision-making. *Blood Cells Mol. Dis.* **26**, 611–612 (2000).
4. Hagege, E. et al. Type 2 Gaucher disease in an infant despite a normal maternal glucocerebrosidase gene. *Am. J. Med. Genet. A* **173**, 3211–3215 (2017).
5. Pachchek, S. et al. Accurate long-read sequencing identified GBA1 as major risk factor in the Luxembourgish Parkinson's study. *npj Parkinsons Dis.* **9**, 156 (2023).
6. Graham, O. E. E. et al. Nanopore sequencing of the glucocerebrosidase (GBA) gene in a New Zealand Parkinson's disease cohort. *Parkinsonism Relat. Disord.* **70**, 36–41 (2020).
7. Leija-Salazar, M. et al. Evaluation of the detection of GBA missense mutations and other variants using the Oxford Nanopore MinION. *Mol. Genet. Genom. Med.* **7**, e564 (2019).

Author contributions
M.T.—Conceptualization, data analysis, paper draft, and review. A.H.V.S.—paper review. F.J.S.—Conceptualization, paper review. C.P.—Conceptualization, data analysis, paper draft and review.

Competing interests
FJS receives research support from PacBio and Oxford Nanopore. AHVS has received consulting fees from AvroBio, Auxilius, Coave, Destin, Enterin, Escape Bio, Genilac, and Sanofi and speaking fees from Prada Foundation. MT and CP declare no competing interests.

IP rights notice

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-08060-5.

Correspondence and requests for materials should be addressed to Marco Toffoli or Christos Proukakis.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Christina Karlsson Rosenthal.

Reprints and permissions information is available at http://www.nature.com/reprints

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