Advancing the knowledge on the *GBA* gene and its role on the pathogenesis of Parkinson disease

Marco Toffoli

A thesis submitted for the degree of Doctor of Philosophy
Department of Clinical and Movement Neurosciences,
UCL Queen Square Institute of Neurology
University College London
2021

Declaration

I, Marco Toffoli, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Since its first description in the XIV century, the understanding of Parkinson disease (PD) has advanced significantly. However, a considerable part of its pathogenesis remains elusive, and no disease modifying therapy has been successfully developed yet and partly because of this lack of knowledge. The genetic background of PD is diverse, accounting for both mendelian, familial and sporadic forms of the disease. One of the main contributors to the genetic risk of PD is the *GBA* gene, which encodes for a lysosomal enzyme and is also linked to Gaucher disease, an autosomal recessive storage disorder. *GBA* variants are relatively frequent in sporadic PD, making it a promising target for disease modifying therapies. However, the penetrance of *GBA* variants is low and mostly unexplained, with only a minority of *GBA* variants carriers developing PD. In addition, the study of *GBA* is complicated by difficulties in sequencing the gene, due to the presence of a highly homologous pseudogene (*GBAP1*) in close proximity to *GBA*.

The aim of this PhD research is to study potential modifiers of risk of PD among *GBA* variant carriers. First, I collaborated in the development of RAPSODI, an online tool to assess and follow-up a cohort of *GBA* carriers, both with and without PD, and analysed the data produced, showing interesting differences between *GBA* carriers and non-carries. The main outcome of this preliminary assessment is that *GBA* carriers with PD have significantly worse motor and non-motor symptoms compared to *GBA*-negative PD patients.

Further, I refined a method to sequence the *GBA* gene with Oxford Nanopore's MinION and developed a novel method for detecting reciprocal recombinants between the gene and pseudogene. By applying these tools, I was able to detect complex structural variants that might modify the risk of PD, as well as explore the role of intronic variants in *GBA*.

Impact statement

RAPSODI is the innovative online portal developed specifically for this study to assess motor and non-motor symptoms of Parkinson Disease (PD) remotely. This tool is capable of generating reliable and reproducible data without the need to meet with participants face to face and will continue to recruit and follow-up participants after the completion of my PhD. The data produced will be helpful to understand clinical differences between *GBA* carriers and non-carriers and will ultimately expand the knowledge of the role of the gene in the pathogenesis of PD. So far, the data generated are promising: they have shown to highlight differences between *GBA* carriers and non-carriers to an extent that was somewhat unexpected and, once published, will hopefully spark more interest in the study of the link between *GBA* variants and PD.

Furthermore, the novel protocol to sequence *GBA* with Oxford Nanopore Technologies (ONT) is both cost-effective and easy to implement, and will provide both current and future researchers with a tool that simplifies *GBA* analysis. It also allows to analyse intronic regions of the gene, and was able to provide useful information about common *GBA* intronic variants.

As a matter of fact, this ONT pipeline has already been applied by independent research groups worldwide to analyse their cohorts of patients.

Moreover, I developed a method to detect complex reciprocal recombinants in *GBA* which was applied to validate an independent *GBA*-caller created by Illumina. While these structural variants do not affect the coding region of *GBA,* they cause extensive deletion or duplication of genomic regions adjacent to the gene, potentially affecting expression. Thanks to the method described in this study, it is now possible to investigate the role of these complex variants in determining the risk of PD - a field that is completely unexplored to date.

List of Abbreviations

α-syn	α-synuclein
AD	Adjusted depth
AMP-PD	Accelerating Medicines Partnership for Parkinson's disease initiative
aQS	Adjusted quality score
BBB	Blood-brain barrier
BRAIN	Bradykinesia Akinesia Incoordination Test
CNG	Copy number gain
DBS	Deep brain stimulation
DIVs	Deep intronic variants
dPCR	Digital polymerase chain reaction
ER	Endoplasmic reticulum
ERT	Enzyme replacement therapy
GC	Glucosylceramide
GCase	Glucocerebrosidase
GD	Gaucher disease
GPU	Graphic processing unit
GWAS	Genome wide association studies
HADS	Hospital Anxiety and Depression Scale
HMM	Hidden Markov Model
IGV	Integrative Genomics Viewer
INDELS	Insertions and deletions
LBD	Lewy body dementia
LINE	Long interspersed nuclear elements
LMW	Low molecular weight
MAD	Median absolute deviation
MDS-UPDRS	Movement Disorders Society - Unified Parkinson Disease Rating Scale
miRNA	micro-ribonucleic acid
Msec	Milliseconds
NHGRI	National Human Genome Research Institute
OLR	Ordinal logistic regression
ONT	Oxford Nanopore Technologies
OR	Odds ratio
PCR	Polymerase chain reaction
PD	Parkinson disease

PD-D	Parkinson disease dementia
PD-MCI	Parkinson disease – mild cognitive impairment
PPMI	Parkinson’s Progression Marker Initiative
QS	Quality score
QSBB	Queen Square Brain Bank
RAPSODI	Remote Assessment of Parkinsonism Supporting Ongoing Development of Intervention
REC	Research Ethics Committee
REM	Rapid eye movement
RBD	REM sleep behavior disorder
RBD1Q	REM Sleep Behavior Disorder One Question Questionnaire
RBDsq	REM sleep behavior disorder questionnaire
Sd	Standard deviation
SINE	Short interspersed nuclear elements
SMCs	Small molecule chaperones
SNpc	*Substantia nigra pars compacta*
SRT	Substrate reduction therapy
SNV(s)	Single nucleotide variant(s)
SV(s)	Structural variant(s)
UCSC	University of Santa Cruz
UNCALLED	Utility for Nanopore Current Alignment to Large Expanses of DNA
UPSIT	University of Pennsylvania Smell Identification Test
WGS	Whole genome sequencing

Acknowledgements

This has been a long and amazing journey. I learnt so much, and there are so many people that contributed to it, that I would need another thesis to name them all.

First, my gratitude goes to my supervisor, Prof. Anthony HV Schapira, who constantly believed in me and provided invaluable guidance that extends beyond the field of academia.

I would also like to thank my secondary supervisor, Prof. Christos Proukakis, whose passion, versatility and attention to details showed me how proper research is done.

My gratitude goes to my other secondary supervisor, Dr. David Chau, for having the patience and skills to teach me how to use a pipette, and all the other fancy lab things.

I would not have survived the long hours in the lab and in front of a computer without the help of the glorious Royal Free lab team: Elisa Menozzi, Dr. Mathew Gegg, Etienne Laverse, Laura Smith, Dr. Manuela Tan, Philip Campbell, Alyssa Costantini, Dr. Edwin Jabbari, Dr. Mark Cooper, Dr. Giuseppe Uras, Dr. Diego Perez Rodriguez, Monica Emili Garcia Segura, Sofia Koletsi, Soraya Rahal, Sara Cable, Chiao Lee, Dr. Antonella Spinazzola, Revi Shahar Golan, Abigail Higgins, Sara Lucas, Dr. Shi-Yu Yang, Dr. Jan-Willem Taanman, Dr. Melissa Salazar and Dr. Stephen Mullin.

Finally, I would like to thank my family, for their endless support, and my partner Dr. Valentina Signorelli, who gives a sense to everything.

Table of content

Declaration **........** ***2***

Abstract **........** ***3***

Impact statement **........** ***4***

List of Abbreviations **........** ***5***

Acknowledgements **........** ***7***

Table of content **........** ***8***

List of figures **........** ***12***

1. Introduction **........** ***15***

1.1. Parkinson Disease, an overview 15

1.2. Cognition in Parkinson disease 16

1.3. The genetic background of Parkinson disease 18

1.4. The heterogenicity of the pathology of PD 19

1.5. The link between the *GBA* gene and PD 20

1.5.1. Epidemiology 20

1.5.2. Biological mechanisms 21

1.5.3. Phenotype of *GBA*-associated PD 23

1.5.4. Therapies targeting the *GBA* pathway........24

1.5.5. Limitations of studies on *GBA*-PD........26

1.6. Detection of genetic variants in the *GBA* gene 26

1.6.1. Structure of the *GBA* gene........26

1.6.1. The challenges of sequencing the *GBA* gene 27

1.6.2. Complex structural variants in *GBA*........28

1.6.3. The role of deep intronic variants in *GBA*........30

1.6.1. A note on the nomenclature........30

1.7. Aims of this PhD 31

2. Remote assessment of a cohort of GBA variants carriers, the RAPSODI study **........** ***33***

2.1. Overview and rationale 33

2.2. Methods 34

2.2.1. Design of the RAPSODI study portal........34

2.2.2. Recruitment of participants 38
2.2.3. Definition of *GBA* carriers 41
2.2.4. Collection of saliva samples 41
2.2.5. Statistical analysis 42

2.3. Results 43
2.3.1. Size and demographics of the cohort 43
2.3.2. Constipation 44
2.3.3. Hospital Anxiety and Depression Scale 47
2.3.4. REM Sleep Behaviour Disorder Screening Questionnaire 50
2.3.5. MDS-UPDRS part 2 52
2.3.6. UPSIT 54
2.3.7. Tap test 57
2.3.8. Cognitive assessment 59
2.3.9. Estimating a risk score in GBA variants carriers 72
2.3.10. Public engagement 74

2.4. Discussion 78

3. Optimisation of a pipeline for sequencing the GBA gene with Oxford Nanopore Technology 83

3.1. Overview and Rationale 83
3.1.1. Oxford Nanopore technology, an overview 83

3.2. Methods 85
3.2.1. Extraction of DNA from Saliva samples 85
3.2.2. PCR amplification of the *GBA* gene 85
3.2.3. Barcoding of DNA samples 86
3.2.4. Purification of PCR products 86
3.2.5. Library preparation and sequencing with MinION 87
3.2.6. Bioinformatics 87
3.2.7. Confirmation of positive and negative results 91

3.3. Results 92
3.3.1. Samples analysed 92
3.3.2. Confirmation of positive and negative results 94
3.3.3. Optimisation of the amplifying PCR 94
3.3.4. Optimisation of the barcoding PCR 97
3.3.5. Purification of PCR product with magnetic beads produces better results 98
3.3.6. Albacore vs Guppy 101
3.3.7. The adjusted quality score improves Nanopolish precision 101

3.3.8. Strand filtering ... 103
3.3.9. Clair is superior to Nanopolish in our dataset ... 105
3.3.1. Detection of suspicious coverage drops in homopolymers ... 108
3.4. Discussion ... 113
4. Characterisation of structural variants in the GBA gene ... 119
4.1. Overview and rationale ... 119
4.2. Methods ... 120
4.2.1. Samples ... 120
4.2.2. PCR approach ... 120
4.2.3. In Silico characterisation of recombinants ... 122
4.2.4. PCR free enrichment of genomic DNA for *GBA* locus – UNCALLED ... 124
4.2.5. Digital PCR to confirm presence and copy number of reciprocal recombinants ... 125
4.2.1. Cross-validation of results with novel Illumina caller for GBA ... 126
4.3. Results ... 126
4.3.1. Recombinants detected with PCR approach ... 126
4.3.1. Prediction of breakpoints of reciprocal recombinants ... 127
4.3.2. Phenotype and other *GBA* variants in carriers of non-pathogenic reciprocal recombinants ... 130
4.3.3. UNCALLED ... 132
4.3.4. Confirmation of copy number in reciprocal recombinants with dPCR ... 137
4.3.5. Cross-platform validation of results with Novel Illumina caller for *GBA* ... 138
4.4. Discussion ... 139
5. Analysis of deep intronic variants in GBA and their effect on the phenotype of Parkinson disease ... 143
5.1. Overview and Rationale ... 143
5.2. Methods ... 144
5.2.1. Participants ... 144
5.2.2. Definition of the haplotypes ... 144
5.2.3. Statistical analysis ... 145
5.3. Results ... 145
5.3.1. Sample size and haplotypes ... 145
5.3.2. Effect of the haplotypes on age at diagnosis of PD ... 146
5.4. Discussion ... 148
6. Final remarks ... 150

Bibliography *152*

Appendix – Code *163*

Script to run Guppy on Myriad 163

The wrapping script used to sequence the GBA gene with ONT 164

Albacore 169

Minimap2 169

LAST 169

Nanopolish 169

Script to detect drops in homopolymers 170

fINDEL 171

Sniffle 171

NanoSV 171

CuteSV 172

UNCALLED 172

List of figures

Figure 1: Mechanisms that link GCase disfunction and α-syn accumulation. 23
Figure 2: Structure of the GBA gene and GBAP pseudogene. Different GBA transcripts are shown, aligned to hg38. 27
Figure 3: Structural variants in the GBA gene. 29
Figure 4: Charts used to visually check the appropriateness of a linear model 42
Figure 5: Constipation among the 5 groups, according to the answers to the RAPSODI questionnaire. 46
Figure 6: Constipation, PD vs non-PD participants. 47
Figure 7: Prevalence of depression and anxiety according to the HADS scale. 49
Figure 8: HADS results, PD vs non-PD participants. 50
Figure 9: Proportion of subjects with RBD with cut-offs of 5 and 6. 52
Figure 10: RBDsq, PD vs non-PD participants. 52
Figure 11: MDS-UPRS part 2 total scores across groups. 53
Figure 12: MDS-UPDRS part 2 total scores, PD vs non-PD participants. 54
Figure 13: UPSIT scores at baseline. 55
Figure 14: UPSIT results in the different groups at baseline. 56
Figure 15: UPSIT results, PD vs non-PD participants. 57
Figure 16: KS30 and AT30 scores. 59
Figure 17: Tap-test results, PD patients and non-PD participants. 59
Figure 18: Picture recognition test scores. 61
Figure 19: Picture recognition test, PD patients vs non-PD participants. 61
Figure 20: Simple reaction time test at baseline. 62
Figure 21: Simple reaction time test, PD patients vs non-affected participants. 63
Figure 22: Results of the Choice reaction time test by groups. 64
Figure 23: Choice reaction time test, PD patients vs non-affected participants. 65
Figure 24: Digit vigilance test results. 66
Figure 25: Digit vigilance test, PD patients vs non-affected participants. 67
Figure 26: Spatial working memory test results. 69
Figure 27: Spatial working memory test, PD patients vs non-affected participants. 69
Figure 28: Results of the numeric working memory test. 72
Figure 29: numeric working memory, PD patients vs non-affected carriers. 72
Figure 30: Risk score vs age in GBA-negative controls and non-affected GBA carriers. 74
Figure 31: Flyer of the 19th May 2019 RAPSODI live event. 76
Figure 32: Participants recruited in the RAPSODI study over time.. 77
Figure 33: Example of an unsuccessful amplifying PCR. 95
Figure 34: Testing different annealing temperatures for the amplifying PCR with LongAmp enzyme 96
Figure 35: Baroding PCR optimisaiton. Two samples are showed in this 0.8% agarose gel (S1 and S2).. 97
Figure 36: TapeStation (Agilent) results of one sample which succeeded barcoding PCR (A, B and C) and one sample which failed barcoding PCR (D, E and F). 100
Figure 37: Comparison of purification methods on 0.8% agarose gel.. 101
Figure 38: Nanopolish Quality score increases linearly with depth of coverage.. 103
Figure 39: Nanopolish adjusted quality score does not change with depth of coverage.. 103
Figure 40: Excess of G in position hg38.chr1:155235011 in 3 randomly selected samples. 104
Figure 41: The exces G in position hg38.chr1:155235011 on the alignment (top row) is present on the forward strand (middle row) but not on the reverse strand (bottom row). 104
Figure 42: Sanger sequencing of the single base deletion that was initially missed by ONT (chr1:155239657). 109
Figure 43: Drop in coverage within the poly-G sequence in 3 wild-type samples (upper 3 samples) and in the sample suspected to be a false negative (bottom sample). 109
Figure 44: Possible single base substitution in poly-G sequence. 111
Figure 45: Sanger confirmation of variant nearby homopolymer. Top row: forward strand. 112
Figure 46: Analysis of AD highlights two variants within homopolymers. 112
Figure 47: Suggestion I made to improve Nanopolish 117
Figure 48: Suggestion I made to improve Clair 118
Figure 49: Primers used to detect the different recombinant alleles within the GBA gene. 121
Figure 50: Visual representation of the sentinel positions at the 3' of GBA. 123
Figure 51: Visual determination of the breakpoint of fusion recombinants within the GBA gene. 123
Figure 52: Regions covered by primers used for dPCR. 125
Figure 53: Reciprocal fusion recombinant, test with different aligners 128

Figure 54: Three examples of reciprocal fusion recombinants 129
Figure 55: Reciprocal duplication recombinant, aligned to hg38 after masking GBAP1.. 129
Figure 56: Repetitive regions in the GBA gene and pseudogene from UCSC genome browser. 133
Figure 57: The developer of UNCALLED talks about the system requirements for the software (from the Github page of UNCALLED). 134
Figure 58: Evidence of enrichment with UNCALLED 135
Figure 59: Enrichment with UNCALLED.. 135
Figure 60: Reciprocal recombinants confirmed by adaptive sampling with UNCALLED. 136
Figure 61: Comparison between different SV callers. The sample displayed is a reciprocal fusion recombinant.. 137
Figure 62: All individual haplotypes detected. 146
Figure 63: Age at diagnosis of PD and haplotypes. 148

List of Tables

Table 1: Genes associated with monogenic forms of PD (PARK genes)[9]. 19
Table 2: Questions about lifestile and environmental factors in the RAPSODI study. 35
Table 3: Online consent form to take part in the RAPSODI study. 41
Table 4: Demographics of the RAPSODI cohort. Age is reported as mean ± standard deviation 44
Table 5: Constipation among the 5 groups, according to the answers to the RAPSODI questionnaire. 45
Table 6: HADS depression and anxiety with cut-offs. 48
Table 7: Prevalence of RBD (reported as number of participants and percentage) according to cut-off 5 and 6. 51
Table 8: MDS-UPDRS part 2 results per each group. 53
Table 9: Results of the UPSIT. 55
Table 10: Inter groups differences of UPSIT scores, estimated with linear regression with age and sex as covariates. 56
Table 11: Severity of hyposmia in the different groups at baseline, according to UPSIT scores. 56
Table 12: KS30 score for dominant and non-dominant hand. SD: standard deviation. 58
Table 13: AT30 score for dominant and non-dominant hand. SD: standard deviation. 58
Table 14: Picture recognition test baseline scores for each group. SD: Standard deviation. 60
Table 15: Simple reaction time test results at baseline. 62
Table 16: Choice reaction time test results. SD: Standard deviation. 64
Table 17: Digit vigilance reaction time test results. SD: Standard deviation. 66
Table 18: Results of spatial working memory test. SD: standard deviation. 68
Table 19: Results of the numeric working memory test. SD: standard deviation. 71
Table 20: Points assigned to each risk factor to calculate a risk score of PD. 73
Table 21: agenda of the 18th June 2020 RAPSODI webinar. 77
Table 22: All pathogenic GBA variants detected with Nanopore sequencing.. 93
Table 23: Optimal conditions for PCR to amplify the GBA gene. 96
Table 24: Optimal conditions for Barcoding PCR. 98
Table 25: Results of comparison of different variants-calling approaches.. 108
Table 26: Results of analysis of 92 samples with AD. 111
Table 27: PCR conditions and primers sequences to detect reciprocal recombinant alleles. 121
Table 28: Sentinel positions. 123
Table 29: primers used for confirmatory dPCR. 125
Table 30: Phenotype and other variants in carriers of non-pathogenic reciprocal recombinants. 132
Table 31: Metrics of the four UNCALLED runs. 132
Table 32: Comparison of different methods for masking the reference for UNCALLED.. 133
Table 33: Enrichment with UNCALLED 135
Table 34: Results of dPCR 138
Table 35: Details of samples used for cross-validation of Gauchian. 138
Table 36: Number of participants carrying each haplotype and age at diagnosis of PD. 147

1. Introduction

1.1. Parkinson Disease, an overview

Parkinson disease (PD) is a progressive neurodegenerative disorder, first described by James Parkinson in 1817[1]. The prevalence of the disease has been steadily increasing due to the aging population, going from 2.5 million people with PD in 1990 to 6.1 million people in 2016 worldwide [2]. PD is one of the major healthcare challenges of our time.

PD is caused by degeneration of dopaminergic neurons in the *substantia nigra pars compacta (SNpc)*[3], associated with the formation of insoluble aggregates of the protein α-synuclein (α-syn) in neuron cell bodies and neurites (Lewy bodies and Lewy neurites)[3]. The neurodegeneration can also involve other areas of the brain, from the brainstem and olfactory nucleus to the cortex, causing the multitude of signs and symptoms associated with PD[4].

The clinical hallmark of PD is the presence bradykinesia (slowness of movements), postural instability, muscular rigidity and 4-7 Hz resting tremor[5]. These are commonly referred to as motor symptoms and are associated with degeneration of neurons in the SNpc. The presence of motor symptoms is required for a clinical diagnosis of PD. Other symptoms, including cognitive dysfunction, rapid eye movemement (REM) sleep behaviour disorder (RBD), constipation, depression and hyposmia, are called non-motor symptoms and mirror the extension of the pathology beyond the substantia nigra[6]. Non-motor symptoms of PD often precede the motor symptoms and might be detectable in PD patients years before the onset of motor symptoms[7].

The therapy of PD is based on medications that control symptoms by increasing dopamine levels and activity, either directly (levodopa containing medications) or indirectly (dopamine agonists and monoamine oxidase inhibitors)[8]. There are currently no disease modifying medications capable of preventing, slowing or reversing the progression of the neurodegeneration in PD. The development of disease modifying therapies is complicated by the diversity of the genetic background of PD and the difficulty of an accurate detection of early cases, possibly before the onset of motor symptoms[9].

1.2. Cognition in Parkinson disease

PD patients can develop cognitive symptoms, which are classified into mild cognitive impairment (PD-MCI) and overt dementia (PD-D)[10]. The Movement Disorders Society provides guidelines for the diagnosis of both PD-MCI and PD-D[11,12] and according to these guidelines 10-20% of PD patients present with PD-MCI at the time of diagnosis[13] and 46% develop PD-D within 10 years from diagnosis[14]. A recent study on a cohort of 141 PD patients found that nearly half of them developed PD-MCI within 2-6 years of follow-up and that all patients with PD-MCI developed PD-D within 5 years[15]. The risk of PD-D is mostly dependent on the age at onset of PD[16]. The pattern of cognitive impairment in the early stages of PD-MCI is characterised by the prominent involvement of executive functions, attention, working memory and, in some instances, visuo-spatial skills and language[17,18]. Executive functions include cognitive flexibility, attention, inhibition, planning and managing daily tasks and are generally directed to adapting to new situations by using previously acquired information to achieve a goal[19]. Executive function can be assessed with tests such as the Tower of London planning test[20] and the trail making tests[21]. Deficits of executive functions are observed in 30%[22,23] of PD patients and are considered the most common cognitive disorder in PD.

Attention, the process of filtering information, is closely related to executive functions. Two separate types of attention can be defined: simple attention, an automatic process of information filtering, and complex attention, more elaborate and controlled by higher level functions[17]. Simple attention is commonly assessed by tests like the Digit span forward and backward and the Trail Making Tests part A[21]. Complex attention deficits can be identified by tests such as the Trail Making Tests part B[21] and the Wisconsin Cards Sorting Test[24]. While there is some debate whether simple attention is primarily affected in PD, complex attention deficits are a common feature of PD cognitive syndromes[17,25,26].

Memory is a broad concept that includes long-term and short-term components. Long-term memory allows the retrieval of information that was stored a long time before, while short-term memory can store information only for a limited amount of time. Working memory is sometimes used synonymously with short-term memory, although some consider the two entities separately, with working memory being the process of manipulating stored information, thus being closely related to executive functions[19,27]. Long-term memory is usually preserved in PD, this being one of the main differences between PD-D and Alzheimer disease. On the other hand, working memory is frequently impaired in PD[28–31].

Visuo-spatial functions are involved in the processing of visual stimuli, including pattern recognition, analysis of space and figure construction[17]. Visuo-spatial deficits have been associated with specific subsets of PD-D[32,33].

Verbal fluency deterioration has also been observed in patients with PD, although language is typically less involved than other cognitive domains[33–35].

According to an interesting hypothesis, two distinct cognitive patterns can be recognised in PD: one involving primarily the cortico-striatal dopaminergic system, associated with a dysexecutive syndrome, and one involving the posterior cortical areas, with deficits mainly in visuo-spatial, memory and language functions. According to this hypothesis, these two different syndromes are associated with groups of genes involved either in dopamine metabolism (cortico-striatal syndrome) or microtubule assembly and stabilization (posterior syndrome)[36].

1.3.The genetic background of Parkinson disease

PD is classified as familial and sporadic, according to whether there is a history of PD among blood relatives. It is estimated that between 5 and 10% of all PD cases are familial, and only 3-5% of the familial PD cases show a clear mendelian pattern[37]. Genes involved in familial monogenic forms of PD are designated with a PARK number, in order of discovery. A list of PARK genes is reported in Table 1. PARK associated PD can be further classified according to its mendelian pattern of inheritance into dominant and recessive. *LRRK2* is the gene most frequently associated with autosomal dominant familial PD, accounting for around 10% of cases[38], while *PRKN* is responsible for 50% of all autosomal recessive, early onset PD cases[39].

The vast majority of PD cases are considered sporadic[37]. Importantly, even sporadic PD shows a clear genetic background, with numerous genes that can increase the risk of PD in a polygenic fashion. Among these genes, the most common, and the ones that are most consistently highlighted by genome wide association studies (GWAS), are *LRRK2*, *GBA*, *SNCA* and *MAPT*[40].

Gene name	Inheritance pattern
SNCA/PARK1-4	Dominant
PRKN/PARK2	Recessive
UCHL1/PARK5	Dominant
PINK1/PARK6	Recessive
DJ-1/PARK7	Recessive
LRRK2/PARK8	Dominant
ATP13A2/PARK9	Recessive
GIGYF2/PARK11	Dominant
HTRA2	Dominant
PLA2G6/PARK14	Recessive
FBXO7/PARK15	Recessive
VPS35/PARK17	Dominant
DNAJC6/PARK19	Recessive
SYNJ1/PARK20	Recessive
DNAJC13/PARK21	Dominant
CHCHD2/PARK22	Dominant
VPS13C/PARK23	Recessive

Table 1: Genes associated with monogenic forms of PD (PARK genes)[9].

1.4. The heterogenicity of the pathology of PD

PD is defined clinically by the presence of motor symptoms and the response to levodopa[41]. However, the pathophysiological mechanisms determining these symptoms are diverse, as suggested by the number of genes associated to the disease. At least three main pathways can be recognised, involving the lysosomes (*GBA*, *LRRK2*), the mitochondria (*PRKN*, *PINK1, DJ1*) and α-syn production and degradation (*SNCA*)[38,39,42,43]. The identification of *MAPT* as a risk factor suggests additional pathways, as this gene is involved in the stabilisation of microtubules[44]. Other mechanisms might also involve the ubiquitin/proteasome pathway, membrane homeostasis, vesicular formation and trafficking.

1.5. The link between the *GBA* gene and PD

1.5.1. Epidemiology

Variants in the *GBA* gene cause Gaucher disease (GD), one of the most common lysosomal storage disorders[45]. GD is a rare autosomal recessive disease, with a prevalence of 1-3:100,000 in European and Australian populations[46–48], but it is significantly more common in the Ashkenazi Jewish community, where its incidence reaches 1:800 live births[49]. GD has been classified into type 1, type 2 and type 3. Type 1 is the milder form of the disease, with most patients reaching adulthood, and is commonly referred to as "non-neuronopathic", as it lacks a clear neurological involvement. The systemic symptoms of type 1 GD include bone abnormalities, anaemia and hepato- and splenomegaly. Type 2 is the "acute neuronopathic" form of the disease. Most patients with type 2 GD die within the first years of life and show severe neurological signs. Finally, type 3 is the "chronic neuronopathic" form, with most patients reaching adulthood, despite carrying significant disability due to neurological involvement[45]. The therapies currently available for GD are very effective in treating the systemic manifestations of the disease, but do not act against the neurological symptoms[50].

A growing body of evidence suggests a clear relationship between GD and PD. Many reports observed a higher than expected prevalence of PD among GD type 1 patients[51–53]. This contradicted the classical paradigm where type 1 GD did not feature neurological involvement and suggested that homozygous and compound heterozygous carrier of *GBA* variants were at increased risk of PD. Following these reports, independent cohort studies discovered that *GBA* variants were significantly more common in sporadic PD cases compared to controls[54–59]. The odds ratios (OR) estimated by these studies spanned from 4 to 20, but all concurred in identifying heterozygous *GBA* variants as a risk factor for PD. One of the most important studies in the field compared the prevalence of *GBA* variants in 5691 PD patients and 4898 controls and estimated an OR of 5.43[60]. Numerous meta analyses and GWAS studies confirmed these findings[44,61–64].

GBA is now considered one of the genes most frequently associated with sporadic PD, and the estimated risk of developing PD among GBA carriers is between 10% and 30% by the age of 80[65].

1.5.2. Biological mechanisms

The mechanisms that link *GBA* and PD have not been fully elucidated and potentially include a number of different cellular mechanisms, possibly mutation specific. Glucocerebrosidase (GCase), the product of expression of *GBA*, is a lysosomal enzyme. Its main substrate is glucosylceramide (GC), that is hydrolysed into glucose and ceramide. The systemic manifestations of GD are caused by the accumulation of GC, especially in macrophages. Accumulation of GC also seems to play a role in the pathology of *GBA*-linked PD[66–68], although evidence of substrate accumulation in PD brain is controversial[69,70]. In support of this hypothesis, human dopaminergic induced pluripotent stem cells carrying *GBA* variants show accumulation of α-syn and GC[71] , that can be reverted by treatment with an inhibitor of the enzyme GC-synthase[66], and rats treated with the GCase inhibitor conduritol B epoxide show increased levels of α-syn[72].

A second proposed mechanism relates to the misfolding of mutant GCase. Usually, GCase travels from the endoplasmic reticulum (ER) to the Golgi and on to the lysosomes, but some *GBA* mutations (e.g. pL483P) might interfere with this process. The misfolded GCase can cause saturation of the ubiquitin-proteasome and ER associated protein degradation systems, leading to the inability of the cell to degrade other proteins, like α-syn[71,73,74].

Other possible mechanisms include mitochondrial dysfunction and inhibition of autophagy (Figure 1).

It is important to note that, while some mechanisms, like those directly related to the reduction of GCase activity, are linked to a loss of function, others, like misfolding of GCase, are more likely to act by gain of function. As a consequence, different *GBA* variants might contribute to the development of PD through more than one mechanism. While severe *GBA* variants like p.L483P bear a higher risk of PD compared to mild variants, reflecting a greater loss of GCase activity[75,76], other PD risk factors like E365K do not cause GD and are associated with normal or only a modest reduction of GCase activity[77].

While *GBA* variants increase α-syn accumulation, α-syn can in turn decrease GCase activity. This bidirectional pathway between GCase and α-syn explains the finding of reduced GCase activity in PD patients without GBA [78,79] and suggests that targeting GCase might have an effect on α-syn pathology even in non-*GBA* PD cases.

Figure 1 provides a visual overview of the mechanisms that link *GBA* variants and PD.

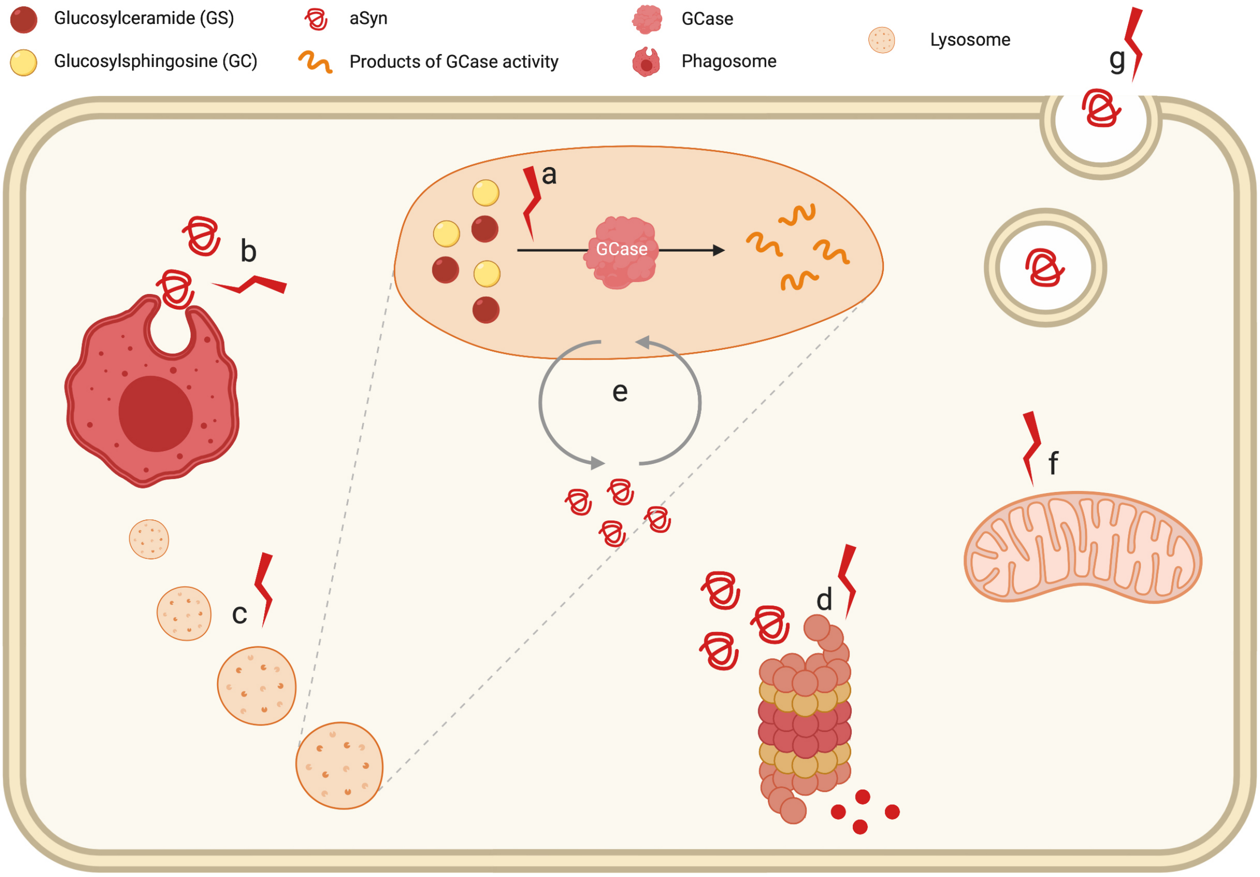


Figure 1: Mechanisms that link GCase dysfunction and α-syn accumulation.

a. Deficit of GCase activity leads to accumulation of its substrates GC and GS. This might alter the homeostasis of α-syn in the lysosomes.

b. GBA deficiency causes downregulation of autophagy, one of the key processes to remove α-syn from cells.

c. After macroautophagy, lysosomes are reformed. GBA mutations interfere with this process, reducing the number of functional lysososmes.

d. Mutant (misfolded) GCase can saturate the ubiquitin-proteasome pathway, reducing its activity of degradation of α-syn

e. Complex structural interactions between α-syn and GCase can lead to a negative feedback where increased α-syn levels impair the activity of GCase and vice versa.

f. Mutant GCase has been shown to damage mitochondrial activity and structure, a process that can lead to/aggravate synucleinopathies

g. Mutant GCase is believed to interfere with the excretion of α-syn, increasing the spread of α-syn pathology.

Abbreviations: GCase Glucocerebrosidase; GC glucosylceramide; GS glucosylsphingosine; α-syn alpha-synuclein. (Figure and caption reproduced from Toffoli et al [80]*, with permission of the publisher).*

1.5.3. Phenotype of *GBA*-associated PD

The phenotype of PD associated with *GBA* variants is similar to that of idiopathic PD, with some significant exceptions[81]. First, age at onset in *GBA*-PD is between 3 and 10 years earlier compared to idiopathic PD[82–86]. The pattern of symptoms is more severe in *GBA*-PD, with a faster rate of motor progression[82], more frequent motor complications like dysphagia, dyskinesia and motor fluctuations[82,85] and a higher prevalence of non-motor symptoms like urinary urgency, constipation, hyposmia, RBD and hallucinations[83–90]. Interestingly, cognitive impairment also seems to be more frequent in *GBA*-PD and to bear a distinctive pattern, with a prominent visuo-spatial component compared to idiopathic PD[90,91]. This is mirrored by a reduction in perfusion and synaptic activity in the parietal and posterior cortices in *GBA*-PD[82,92]. This finding has interesting implications related to the dual syndrome hypothesis in PD cognition (see section 1.2), where a more "posterior" syndrome, with prominent visuo-spatial and language deficits, might be associated with specific cellular pathways[36].

1.5.4. Therapies targeting the *GBA* pathway

The discovery of the link between *GBA* variants and PD generated interest in potential therapeutic applications, as existing therapies for GD which might be applied to slow down or even reverse the neurodegeneration in PD. The cornerstone of GD treatment is the enzyme replacement therapy (ERT)[93], based on the administration of functioning GCase through periodic infusions. ERT is very effective in treating the systemic manifestations of GD, but does not cross the blood brain barrier (BBB), so it does not act on the neurological manifestations of GD and is not a good candidate for the treatment of *GBA*-associated PD. An alternative strategy for treatment of GD is substrate reduction therapy (SRT), which acts by inhibiting the enzyme Glucosylceramide synthase, thus reducing the synthesis of GC and preventing its accumulation due to GCase deficiency[93]. Many SRT molecules can cross the BBB and could potentially be used in *GBA*-PD[94]. Unfortunately, a recent phase 2 clinical trial on venglustat, a promising SRT molecule, failed to meet efficacy endpoints in treating PD[95]. Other approaches with the potential to prevent the neurodegeneration in *GBA*-PD have been investigated in recent years, including small molecule chaperones and gene therapy[96]. Small molecule chaperones (SMCs) act by facilitating the correct folding of mutant GCase and allowing it to reach the lysosomes, thus increasing overall GCase activity. Many SMCs have shown potential to rescue GCase activity in *GBA* variants carriers in pre-clinical and clinical studies[97]. In particular, the SMC ambroxol has recently been studied. Ambroxol is a drug used to treat airway mucus hypersecretion which has showed good GCase chaperoning activity[98]. Its advantage over other molecules is that it has an excellent and well established safety profile, which allows a much more rapid progression through regulatory frameworks to clinical trials. Ambroxol is able to reduce total and phosphorylated α-syn levels in animal models[99] and a clinical trial demonstrated that it can cross the blood brain barrier[100].

On the other hand, gene therapy aims at introducing wild-type copies of *GBA* into the genome of mutation carriers using adenovirus vectors, with the final aim of increasing GCase activity[97].

SRT, SMCs and gene therapy have the potential to modify the course of PD, but their application in a clinical setting is challenging. It is estimated that by the time motor symptoms of PD emerge, up to 50% of dopaminergic neurons in the *substantia nigra* are already lost[101]. This implies that for any therapy capable of arresting the neurodegeneration in PD, early initiation of treatment, even before the onset of motor symptoms, could be more effective. Given that only 10-30% of *GBA* variants carriers will develop PD in their lifetime[65], it would not be reasonable to include them into clinical trials on disease modifying therapies without an appropriate risk stratification. To this end, being able to define who among *GBA* variants carrier is at a higher risk of developing PD is paramount.

1.5.5. Limitations of studies on *GBA*-PD

There is a consistent body of evidence supporting the role of *GBA* variants in the pathogenesis of PD. However, most studies are limited by the methods used to detect these variants. Some studies only detect the most common single nucleotide variants (SNVs) p.N409S and pL483P. This is the case of the multicentre study by Sidransky et al[76]. While providing sound evidence of an association between *GBA* and PD, some interesting variants like E326K and T369M, which are common risk factors for PD but are non-pathogenic for GD, were not included in the analysis. A later GWAS found E365K to be the major author of the association between *GBA* and PD[63,102]. In fact, only a minority of studies report the full sequencing of all *GBA* exons, providing an accurate picture[103]. This disparity can lead to several inaccuracies in estimating the prevalence and the real effect of *GBA* variants towards PD. In the next section, I discuss what are the main challenges in sequencing the G*BA* gene.

1.6. Detection of genetic variants in the *GBA* gene

1.6.1. Structure of the *GBA* gene

The *GBA* gene is located on the long arm of chromosome 1, at coordinates GRCh38.chr1:155234452-155244627 (https://www.ncbi.nlm.nih.gov/gene/2629). It features a 39 amino acid leader peptide that is later cleaved from the protein, while the remaining 11 exons encode the mature GCase. A pseudogene (*GBAP1)* lies 6.9Kb downstream and present some regions of very high homology with *GBA*. Immediately adjacent to the 3' end of *GBAP1* lies *MTX1*, a gene encoding for the protein metaxin-1, which is part of a preprotein import complex in the outer membrane of mitochondria[104]. The highly homologous pseudogene *MTX1P1* is immediately adjacent to the 3' end of *GBA*[105]. As discussed later, this layout causes a high frequency of complex recombination events in the *GBA* locus. A visual representation of the *GBA* gene and pseudogene is provided in Figure 2.

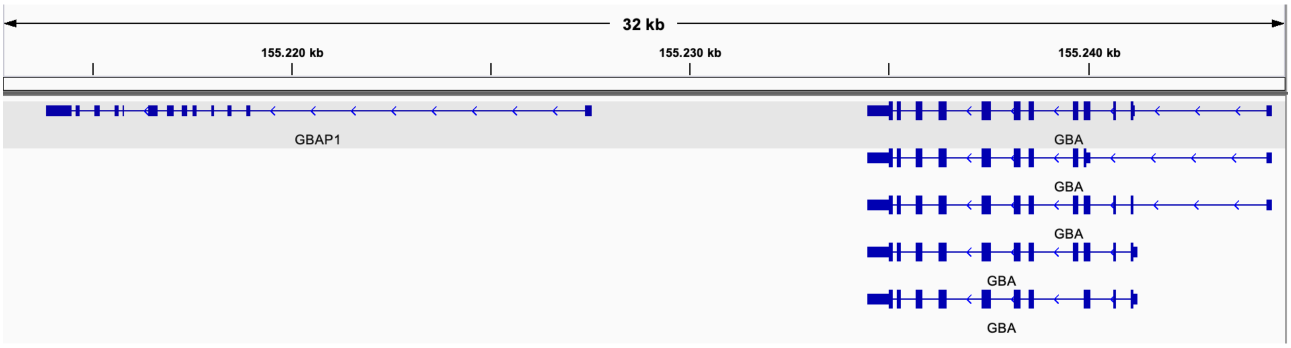


Figure 2: Structure of the GBA gene and GBAP pseudogene. Different GBA transcripts are shown, aligned to hg38.

1.6.1. The challenges of sequencing the *GBA* gene

Sequencing of *GBA* poses many challenges. Sanger sequencing of PCR amplicons is widely used for diagnostic purposes, but the high homology causes unwanted amplification of *GBAP1*. To overcome this problem, a nested PCR approach can be used[106]. Since pathogenic mutations can be found in all 11 exons that encode the mature GCase protein[107], 10 forward and 10 reverse Sanger reactions are required, making this process expensive and time consuming. Moreover, this method is not able to differentiate between non-reciprocal and reciprocal gene fusion recombinants and does not detect reciprocal gene duplication recombinants (as discussed later).

Recently, an allele-specific PCR approach was described, using a multiplex PCR with specific oligonucleotides that can exclusively amplify mutant alleles to detect the 4 most common *GBA* SNVs[108]. However, this method produced a relatively high rate of false positive L483P alleles, possibly because of amplification of *GBAP1*.

For all these reasons, a long-read approach might be preferable, but design of a pipeline capable of detecting both SNVs and structural variants (SVs) reliably is challenging[109]. In chapter 3 of this thesis I describe how I improved and validated a pipeline for sequencing *GBA* with PCR enrichment and long read sequencing with Oxford Nanopore Technology.

Short read whole genome sequencing (WGS) is also impaired by the presence of the pseudogene*,* which causes mis-alignment and mis-call of variants like p.L483P, natively present in *GBAP1*[110,111]. In chapter 4, I discuss my contribution to the validation of a novel caller for *GBA* based on Illumina WGS.

1.6.2. Complex structural variants in *GBA*

Genetic structural variants (SVs) are the product of complex recombination events where large regions of DNA are rearranged. These rearrangements are an essential process of meiosis in mammalian cells and contribute to the genetic variation that is at the base of evolution and natural selection[112]. However, in some cases homologous recombination events can cause genetic disorders[113]. Recombination events can be classified as reciprocal and non-reciprocal. In non-reciprocal recombination, a segment of DNA from a donor site is copied into the acceptor site, without loss of DNA material at the donor site. In reciprocal recombination, the transfer of genetic material between the two sites involved is equal, resulting in two possible patterns, a fusion (copy number loss – CNL) or a duplication (copy number gain – CNG) recombination. In the fusion recombinant, two sites of DNA are merged together and all genetic material between them is deleted, while in duplication recombinants a segment of DNA is duplicated (or multiplicated). Figure 3 illustrates the structure of reciprocal and non-reciprocal recombinants.

The first SV in *GBA* was described in 1990 and consisted of an allele carrying 3 SNVs in exon 10. These 3 SNVs, p.L483P, p.A495P and p.V499=, corresponded to *GBAP1* sequence that was inserted into *GBA* and the allele was called Rec NciI[114]. Subsequently, other recombinants were reported, typically including the RecNciI pattern and some additional variants deriving from *GBAP1* sequence[115]. Pathogenic

SVs in the *GBA* gene can be non-reciprocal events or reciprocal gene fusion events. If the recombination involves exon 10, the Rec NciI pattern will appear. If exon 9 is involved, additional SNVs and a 55 base pairs deletion might be present[115]. If the reciprocal event is downstream of the RecNciI SNVs, no variant will be introduced in *GBA*, as the homology with the pseudogene is complete. In reciprocal duplication alleles, the resulting allele contains a full non-recombinant copy of *GBA* and a recombinant *GBA-GBAP1* structure that carries the *GBAP* 5'UTR sequence (Figure 3). It is unclear whether the reciprocal duplication SVs and the fusions not introducing SNVs in *GBA* are pathogenic for PD.

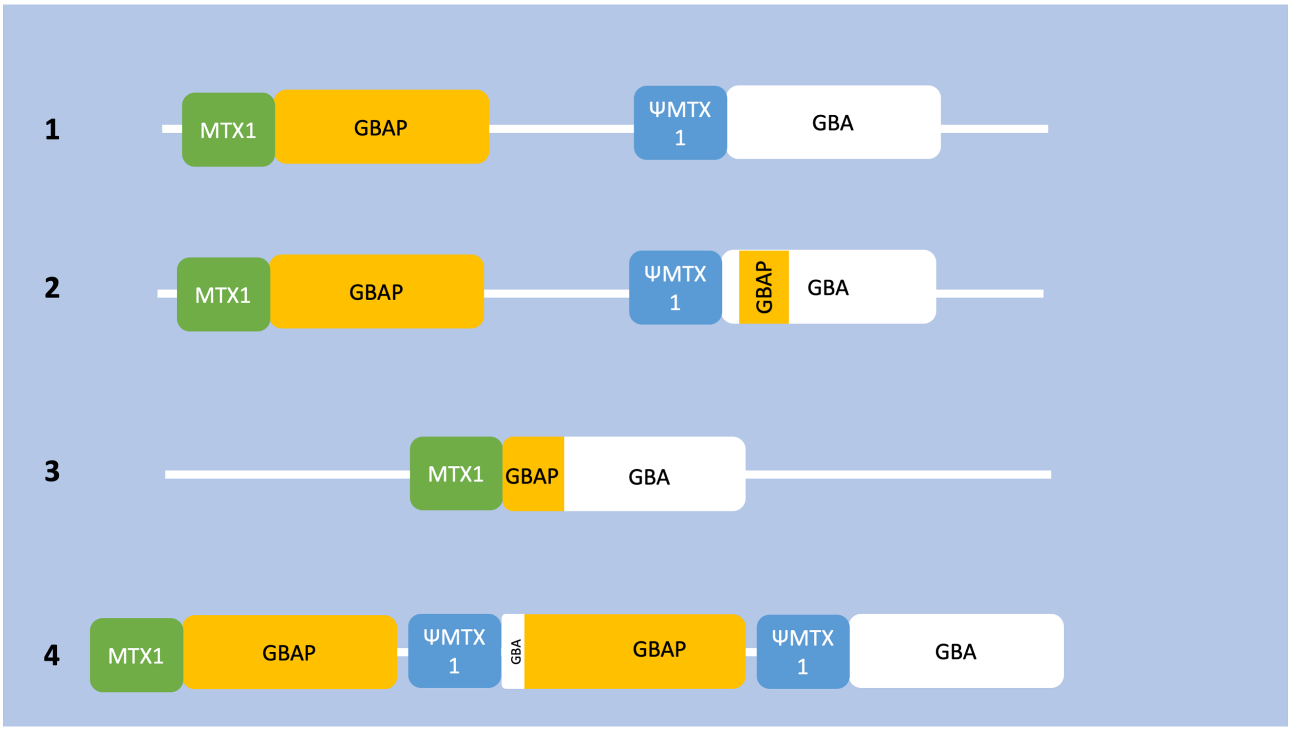


Figure 3: Structural variants in the GBA gene.

1) Wild type GBA and GBAP genes.

2) Non-reciprocal recombinant allele. Part of GBAP is copied into the GBA gene and replaces the corresponding sequence in the gene.

3) Reciprocal fusion recombinant. GBA and GBAP are merged together and all the genetic material between the two breakpoints is deleted.

4) Reciprocal duplication recombinant. GBA and GBAP are merged together to form a recombinant structure, but also preserving normal GBA and GBAP genes in the same allele.

1.6.3. The role of deep intronic variants in *GBA*

While the role of coding *GBA* variants is well established, it is unclear whether deep intronic variants (DIVs) play a role in the pathogenesis of PD. DIVs in other genes have been linked to human diseases like β-Thalassemia[116], Pompe disease[117] and Charcot-Marie-Tooth disease type 1B[118], and recently a DIV in intron 7 of GBA has been linked to GD[119]. Mechanisms of pathogenicity of DIVs include the inclusion of pseudo-exons (by the creation of a novel donor or acceptor splicing site or splicing enhancer or by the disruption of a splicing silencer), the competition with natural splice sites, the disruption of transcription regulatory motifs, the inactivation of non-coding RNA genes and more general genomic rearrangements, often resulting in gene fusions[120].

The reduced penetrance of *GBA* variants toward PD[121] and the finding of reduced GCase activity in PD patients not carrying coding *GBA* variants[78,79] suggest that intronic *GBA* variants might play a role in PD. This is supported by the existence of two common intronic haplotypes in *GBA*, characterised by at least 3 intronic variants, although some authors have subsequently identified more variants associated with the haplotypes and even described sub-haplotypes[122,123]. Interestingly, a recent study showed a correlation between the two intronic haplotypes in *GBA* and age at onset of PD in patients without any other coding *GBA* variant, although I was not able to reproduce this finding[124,125].

1.6.1. A note on the nomenclature

Single nucleotide variants in *GBA* are commonly referred to with the corresponding amino acid change. According to this terminology, p.N370S, the most common GD causing variant, corresponds to a nucleotide substitution at position GRCh28.chr1: 155235843, resulting in the change of the 370th amino acid from an asparagine to a serine. However, some authors consider the 39 aa leader peptide and some others do not. This generates two parallel sets of coordinates for the same variants that can cause confusion to unprepared readers. For example, GRCh28.chr1: 155235843 is referred to as p.N370S or p.N409S by different authors.

The numbering of *GBA* exons is afflicted by a similar issue. While the most widely used transcript has 11 exons, some authors prefer a transcript with 12 exons. Throughout this thesis, the nomenclature that accounts for the 39 aa leader peptide will be used to refer to SNVs, and the exons will be numbered starting from 1 to 11, only accounting for the ones encoding for the mature protein (Figure 2).

Regarding reciprocal recombinants (see section 1.6.2), I opted for the nomenclature used by Tayebi in their paper on structural recombination in *GBA*[115]. So reciprocal recombinants with CNLs are defined fusion recombinants, and those with CNGs duplication recombinants.

I define fusion recombinants as “pathogenic” if they affect the coding region of *GBA*, and “non-pathogenic” if they do not. However, it is important to note that it is unclear whether “non-pathogenic” fusion recombinants actually affect the risk of disease.

The assembly used throughout this thesis is GRCh38.p13.

1.7.Aims of this PhD

Understanding the role of *GBA* variants is crucial to develop disease modifying therapies for PD. Major challenges include the identification of individuals at a higher risk of PD among *GBA* mutation carriers and the detection and characterisation of all *GBA* variants*,* in particular complex SVs.

To address these issues, the aims of my PhD are:

1. Development of RAPSODI, an online portal to assess a large number of *GBA* variant carriers, and utilise it for:

a. Evaluating the feasibility of using remote assessment for cohort studies on *GBA* carriers
b. Detecting early signs of PD among *GBA* variants carriers
c. Detect differences in PD phenotype in *GBA* variants carriers vs non carriers

2. Optimisation of a pipeline for sequencing of *GBA* using Oxford Nanopore Technology and apply it for:
   a. Detecting coding variants in the RAPSODI cohort
   b. Characterising structural variants in *GBA*
   c. Analysing additional genetic information (e.g. intronic regions) provided by long read data
   d. Validating results by a novel caller of SVs developed by Illumina

2. Remote assessment of a cohort of *GBA* variants carriers, the RAPSODI study

2.1. Overview and rationale

The role of *GBA* variants in the pathogenesis of PD is supported by epidemiological and *in vitro* evidence. However, the development of targeted therapies is complicated by the low penetrance of these variants. Since only a minority of *GBA* variant carriers develop PD in their lifetime, the design of clinical trials is challenging. There is indeed a major need for reliable ways to estimate the risk of PD in *GBA* variants carriers and to determine who among them has a higher chance of developing PD. Prior to my PhD, I collaborated with the host laboratory on a longitudinal study on a cohort of *GBA* variant carriers. The scope of the study was to follow these individuals over time in order to create an algorithm for estimation of the risk of PD. This study provided important information, but also identified major challenges of this approach[126–128]. First, a long follow-up period was required to allow a consistent number of *GBA* carriers to develop PD. Second, compliance was a significant issue, as *GBA* carriers are essentially healthy individuals, who are not very inclined to attend to periodic visits. Third, finding *GBA* carriers to include in the study was complicated by technical difficulties in sequencing the gene. To address these issues, we decided to design an online tool capable of assessing a larger number of *GBA* carriers remotely and follow them up over time (the RAPSODI portal). In this section of my PhD, I describe my work in designing and improving the RAPSODI portal and the results of a preliminary baseline assessment.

2.2. Methods

2.2.1. Design of the RAPSODI study portal

RAPSODI (rapsodistudy.com) is an online portal that Dr. Stephen Mullin and I designed between 2017 and 2018. The code for the website was created by the web developer AAH software (https://aahsoftware.uk/) following the instructions that Dr Mullin and I provided.

My specific role in the development of the portal was the selection of the questions and questionnaires used and the creation of the content of the web pages (in collaboration with Dr. Mullin), testing of all portal functionalities and digitalisation of the questionnaires. I managed recruitment, day to day administration of the portal and data analysis from January 2018 onwards. From March 2020 I was helped by a research assistant for day to day administration of the portal and processing of saliva samples.

RAPSODI is designed to assess motor and non-motor early signs and symptoms and to investigate some basic environmental risk factors for PD in participants to the RAPSODI study. This assessment includes questions about family history of PD and lifestyle, clinically validated questionnaires for non-motor symptoms of PD, a tap test, a battery of cognitive tests, a smell test. The assessment is repeated every year for up to 25 years. The study was approved by the London – Queen Square Research Ethics Committee, (REC reference: 15/LO/1155).

2.2.1.1. Questions about family history and lifestyle

These questions aim at gathering information to stratify participants and identify possible risk factors related to their background and the environment they live in. Questions were selected from a similar study run at UCL, PREDICT-PD, which involves the assessment of motor and non-motor symtpoms of PD in the general population[129]. The rationale for applying the same questions as PREDICT-PD is to allow for cross-study comparison in the future, if appropriate. A full list of questions is provided in Table 2.

Do you have any of the following medical conditions? (Parkinson disease, a movement disorder, stroke, motor neuron disease, dementia, none)
How old were you when you noticed your first motor symptom/symptoms of Parkinson disease? E.g. tremor, rigidity, postural instability.
In which year was your Parkinson disease diagnosed?
Have you undergone Deep Brain Stimulation (DBS)?
Do you have any other medical condition?
Does anyone in your family have Parkinson's disease (a blood relative)?
Please list all the medications that you take REGULARLY.
How many bowel movements do you usually have in one week? (1, 2, 3-4, 5-8, 9-12, more than 16)
Do you ever use laxatives?
Does opening your bowels require a lot of effort?
Do you suffer from hard stools?
Do your hands, arms or legs ever shake?
Do you shuffle your feet and take tiny steps when you walk?
Rate your ability in the previous 3 months, without treatment, to have and maintain an erection adequate for intercourse (poor, fair, good, I'd rather not answer)
Do you ever use Viagra or similar to improve ability to achieve an erection?
Do you drink coffee?
Do you or did you smoke? (how many sigarettes/day and for how long?)
Do you drink alcohol?
Do you think you have received significant exposure to pesticides in your lifetime?
Have you ever hit your head so strongly you nearly fainted or nearly lost consciousness, or sustained significant trauma to your face or nose?
Have you ever hit your head so strongly that you did lose consciousness?

Table 2: Questions about lifestile and environmental factors in the RAPSODI study.

2.2.1.2. Clinically validated questionnaires

The questionnaires used are the REM sleep behavior disorder one question questionnaire (RBD1Q)[130], the REM Sleep Behavior Disorder Questionnaire (RBDsq)[131], the Unified Parkinson Disease Rating Scale part 2 (MDS-UPDRS part 2)[132], Hospital Anxiety and Depression Scale (HADS)[133].

2.2.1.3. *Tap test*

The BRadykinesia Akinesia INcoordination (BRAIN) test was used to evaluate hand dexterity and the presence of bradykinesia in participants[134,135].

Participants are asked to press the "S" and ";" keys on their keyboard in succession as fast as they can. First, participants are required to carry out a training session of 5 seconds with each hand, where no data is captured. Then participants are asked to do the task with each hand separately for 30 seconds. Each hand is tested separately. Two composite measures were used for analysis: the kinesia score (KS30) and the akinesia time (AT30). KS30 is the number of key taps in 30 seconds; AT30 is the mean dwell time on each key in milliseconds (msec)[134,135].

2.2.1.4. *Cognitive assessment*

The cognitive tests are hosted on an external web portal called CogTrack (https://www.wesnes.com)[136].

In the RAPSODI study, 6 different tests are used.

- Pattern separation ability. This task measures the ability to encode, store and subsequently retrieve visual information. A series of 20 pictures from separate categories is initially presented one at a time. Around 10 minutes later, the 20 original pictures are presented mixed with 20 matched pictures which are each very similar to the original picture. For each picture the participant must make a response as to whether the picture was the precise picture presented originally or a different one. The accuracy in recognising the original pictures (DPICOACC) and the new pictures (DPICNACC) is calculated as a percentage.
- Simple reaction time. This task assesses alertness and the ability to focus concentration by measuring the speed with which a simple motor response can be made to an expected stimulus, which occurs repeatedly but at unpredictable intervals. Median reaction time (SRT) is calculated.

- Choice reaction time. This task measures the ability to focus concentration and efficiently process information. The participant monitors the screen for the appearance of one of two possible alternative stimuli which occur at unpredictable intervals. Each of the two stimuli requires a separate key to be pressed and the order of presentation is randomised. The percentage of correct answers (CRTACC) and median reaction time (CRT) are calculated.
- Digit vigilance reaction time. This task measures intensive and sustained concentration; also known as vigilance. The participant is required to monitor a series of digits presented singly and rapidly in the centre of the screen, pressing the right arrow when the digit presented matches the target digit displayed constantly on the right of the screen. Median speed (VIGRT) and percentage of targets detected (VIGACC) are measured.
- Spatial working memory. This task measures the ability to hold spatial information in working memory and to retrieve it as quickly as possible. Three rows of 3 light bulbs are initially presented, with 4 of the bulbs being lit, and the subject has to remember the locations of these. The 3 rows of bulbs are then represented, each time with only one of them lit, and the participant must make a response each time as to whether or not the lit bulb was one that was originally lit. Accuracy expressed in percentage (SPMOACC for correct “yes” answers and SPMNACC for correct “no” answers) and median speed (SPMRT) are calculated.
- Numeric working memory. This task measures a participant’s ability to hold a series of 5 different digits in working memory. The task begins with the presentation of the series, one digit at a time. This is followed by a series of presentations of single digits for each of which the participant must make a response as quickly as possible as to whether or not the digit was in the original series. Accuracy expressed in percentage (NWMOACC for correct “yes” answers and NWMNACC for correct “no” answers) and median speed (NWMRT) are calculated.

2.2.1.5. Smell test

The University of Pennsylvania Smell Identification Test (UPSIT) is used to assess olfactory function. Participants are provided with 4 booklets of 10 pages each. For each page, participants need to scratch the section containing the smell with the provided pencil and then select the option among the 4 suggested that best represent their smell experience. The final score ranges from 0 to 40, one point for each correct answer[137]. Participants are then able to record their answers on the RAPSODI web portal and mail the used UPSIT booklets back to the central study team.

2.2.2. Recruitment of participants

Inclusion criteria for RAPSODI are:

- Age 18-90
- GD patients
- Individuals that already know they carry a *GBA* variant
- Relatives of GD patients and *GBA* variant carriers
- PD patients
- Spouses of GD patients, *GBA* variant carriers and PD patients

Exclusion criteria are the presence of dementia and any other neurological conditions that might cause parkinsonism.

The study started active recruitment in January 2018 and will continue recruiting participants on a rolling basis for up to 25 years.

Participants can enroll themselves autonomously via the dedicated web-page on the RAPSODI portal.

GD patients are recruited at 8 patient identification centres with a lysosomal storage disorder unit: Royal Free Hospital, Addenbrooke's Hospital, Queen Elizabeth Hospital Birmingham, Salford Royal Hospital, Cardiff and Vale University Health

Board, University College Hospital, Great Ormond Street Hospital, and Manchester Children's Hospital. At these centres, the local team informs participants about the existence of the study and hands over flyers with the study team contact details and the website IP address. PD patients and individuals that already know they carry a *GBA* variant are reached by advertising the study on online platforms, like Parkinson's UK (https://www.parkinsons.org.uk) and the UK Gaucher association (https://www.gaucher.org.uk).

After a participant is found to carry a *GBA* variant, the RAPSODI study team reaches out to ask to bring family members in.

Upon enrollment, all participants sign an online consent form (Table 3).

I agree to continue participating in the RAPSODI GD study.
I have read and understood the Participant Information Page for this study and have had the opportunity to ask questions (via email and telephone).
I understand and give permission for the RAPSODI GD research team to gain access to my patient notes to document any previous testing results for the glucocerebrosidase gene (*GBA*)
I understand that I may be required to complete a 'scratch and sniff' smell test, which will be posted to me 1-2 weeks after completing the surveys.
I understand I will be posted a saliva collection pot for genetic studies.
I agree to be contacted in the future for neurological examination or further questions related to the study which may be video recorded.
I understand that, if I give permission, my GP will be contacted to inform them of my participation in the study.
Yes, I give permission.
No, I do not give permission.
I understand that, if I give permission, my genetic results will be disclosed to my GP.
Yes, I give permission.
No, I do not give permission.
I understand that, if I give permission, the clinician who looks after my Parkinson's Disease will be contacted to inform them of my participation in the study.
Yes, I give permission.

No, I do not give permission.
I understand that, if I give permission, my genetic results will be disclosed to the clinician who looks after my Parkinson's Disease.
Yes, I give permission.
No, I do not give permission.
I understand that, if I give permission, the research site (for instance an NHS Hospital) that, introduced me to this study will be informed of my participation in the study.
Yes, I give permission.
No, I do not give permission.
I understand that, If I give permission, my genetic results will be disclosed to the research site (for instance an NHS Hospital) that introduced me to this study.
Yes, I give permission.
No, I do not give permission.
I understand that a member of the study team may contact me, to offer me the opportunity to participate in clinical trials or other research studies. If I give permission, my details, including information related to my Parkinson's disease, will be passed onto them.
I agree to the results of my tests being stored on a secure website for the duration of the study and after its completion.
I understand that any information collected will remain completely confidential.
I understand that the research team may ask me to provide samples of blood, urine or cerebrospinal fluid, however not consenting to give these sample will not affect my ability to participate in the study.
I understand that I may be offered the opportunity to participate in other research studies, either by email, phone or post.
I understand that my involvement is voluntary, and I am free to withdraw at any time, without providing a reason.
I understand that by taking part in this study I agree to be informed of my genetic test results, i.e. the research team will inform me of whether or not I am a carrier of a mutation in the GBA and LRRK2 genes.
I understand and, if it has not already been carried out, give permission for genetic test for the GBA gene to be carried out.
I understand and give permission for genetic testing for the LRRK2 gene to be carried out.
I understand that I may be approached to participate in other research studies based on the results of these genetic studies.

I understand that, if I give permission, any samples and associated data that I provide during my participation in the study may be transferred nationally or internationally to NON-COMMERCIAL collaborators. I understand that, if I give permission, confidentiality will be maintained at all times. Information that directly identifies you, such as your name, will be replaced with a ‘code’ or ‘ID number.’ Your name and other identifying information will not be shared with other researchers. The purpose of this sample and data transfer will advance future research and the potential clinical significance of the study results through the use of more advanced analysis techniques such as artificial intelligence and whole genome sequencing. [For a list of our collaborators and their work please visit www.pdfrontline.com]
Yes, I give permission.
No, I do not give permission.
I understand that, if I give permission, any samples and associated data that I provide during my participation in the study may be transferred nationally or internationally to COMMERCIAL collaborators. I understand that, if I give permission, confidentiality will be maintained at all times. . Information that directly identifies you, such as your name, will be replaced with a ‘code’ or ‘ID number.’ Your name and other identifying information will not be shared with other researchers. The purpose of this sample and data transfer will advance future research and the potential clinical significance of the study results through the use of more advanced analysis techniques such as artificial intelligence and whole genome sequencing. [For a list of our collaborators and their work please visit www.pdfrontline.com]
Yes, I give permission.
No, I do not give permission.

Table 3: Online consent form to take part in the RAPSODI study.

2.2.3. Definition of *GBA* carriers

GBA carriers in RAPSODI are defined as all individuals carrying a variant that was previously described in association with GD[138], or with a variant that is known to be a risk factor for PD but not causing GD (i.e. E365K and T408M)[77]. With regards to variants that were not previously described, they were considered pathogenic if they caused a change of amino acid in the *GBA* gene. A list of all *GBA* variants that were detected is reported in Table 22.

2.2.4. Collection of saliva samples

After completing the cognitive assessment, each participant receives a kit via Royal Mail, containing the saliva collection kit and the UPSIT kit. Participants are asked to complete the UPSIT test, collect a sample of saliva in the Oragene kit and then send the material back to the study team using the pre-labelled and pre-paid envelope provided.

Sequencing of GBA variants is carried out using genomic DNA extracted from saliva samples, following a procedure detailed in chapter 3.

2.2.5. Statistical analysis

R version 4.0.5 was used for the analysis. The tests used and the co-variables considered are reported within each sub-section, for simplicity. Appropriateness of a linear model was tested when linear regression was used (see Figure 4 for an example of the procedure used). Disease duration was highly collinear with age and as such it was not used as a covariate.

For each variable of interest, 3 tests were carried out:

- an exploratory comparison between the 5 groups, with adjustment for multiple comparisons;
- a comparison between PD patients (with and without *GBA* variants) and non-affected participants (without PD or GD);
- a comparison between *GBA*-positive and *GBA*-negative PD patients.

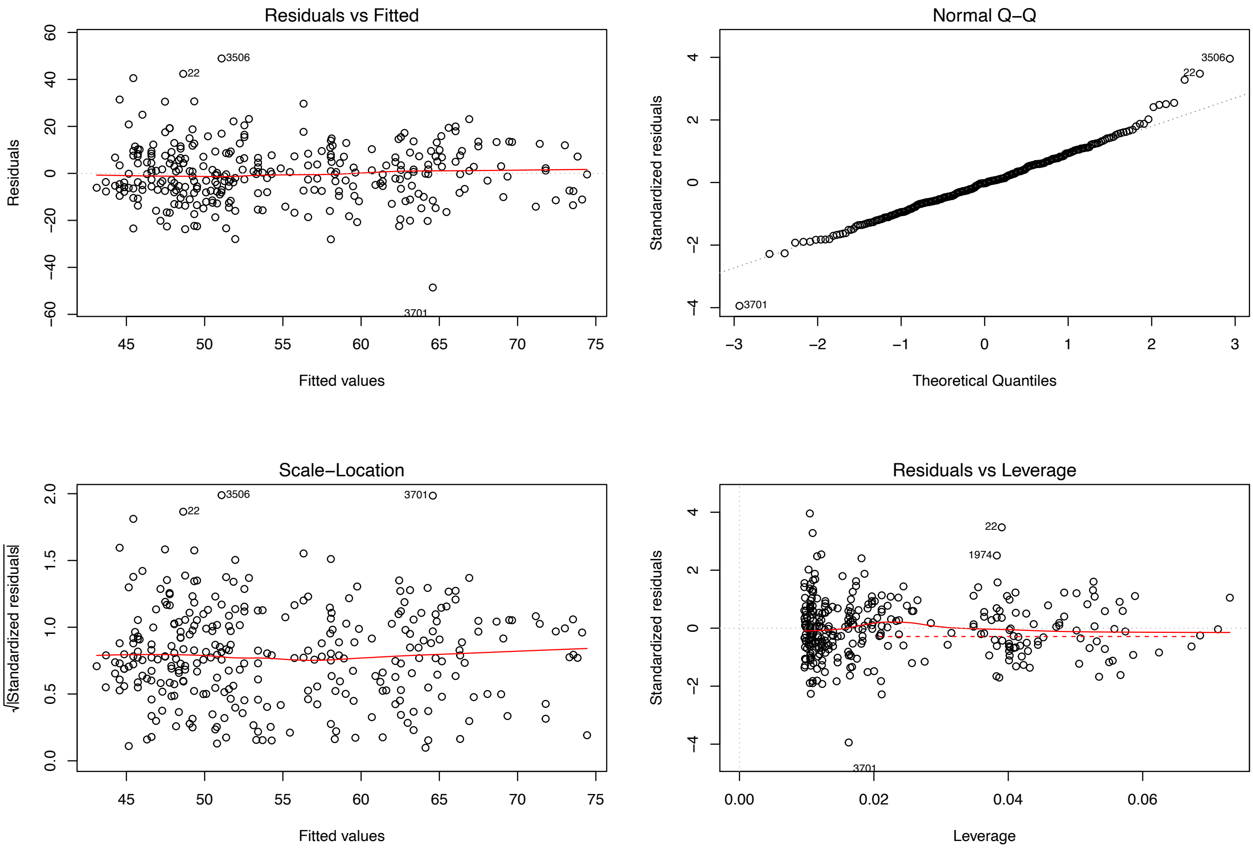


Figure 4: Charts used to visually check the appropriateness of a linear model.
Residuals vs fitted: used to check linearity assumption. A horizontal line without a clear pattern suggests a linear relationship.

Normal Q-Q: used to check normality of the residuals. An approximately straight line suggests normally distributed residuals.
Scale-location: An approximately horizontal line suggests with equally spread points suggests homogeneity of variance.
Residuals vs leverage: Used to detect outliers and high leverage points. The Cook's distance is represented by the red dotted line and allows for the detection of highly influential points.
Source: (http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/)

2.3. Results

The RAPSODI study will continue after the end of my PhD and new participants are recruited on an ongoing basis. The data reported in this thesis are a snapshot of the study as of the 1st March 2021. Only participants that completed the baseline assessment and for which the GBA gene was sequenced are included in the analysis. Only a small number of participants already completed follow-up assessment. For this reason, only baseline data are reported in this thesis.

2.3.1. Size and demographics of the cohort

A total of 808 participants enrolled into the study at the time of writing this thesis (Figure 32). Of them, 295 completed the baseline assessment and had their *GBA* gene sequenced. The remaining either did not complete the baseline assessment, or were not sequenced yet. Those were excluded from the preliminary analysis of results reported below.

Participants were divided into 5 disease status groups: non-affected *GBA* carriers, *GBA* negative controls, GD patients, *GBA*-positive PD patients and *GBA*-negative PD patients.

Number of participants, sex composition, age at baseline and duration of PD for each disease status group are reported in Table 4. A number of participants did not complete some parts of the study and as a result the number of observations for each assessment might be different from the total reported in Table 4. Age at baseline was significantly different between the groups (p-value < 0.001 - ANOVA) and multiple comparison revealed that the *GBA*-negative PD patients had a significantly higher age at baseline compared to *GBA*-negative controls and non-affected GBA carriers (adjusted p-values <0.001, Bonferroni correction). No significant difference in sex distribution was observed between the groups (Pearson's chi square).

When considering only the two PD groups, the *GBA*-negative PD patients were significantly older than the *GBA*-positive ones (mean difference 5.8 years, p-value 0.007 – t-test). They also showed a longer disease duration, although this difference was not significant (mean difference 1.4 years, p-value 0.45, t-test). The two groups were homogeneous for sex and years of education. Since age and disease duration were highly collinear (p-value 0.006 – linear regression), disease duration was not considered as a co-variate in the comparison between the groups.

Disease status	Number of participants	Males (percent of the total)	Age	**PD duration**
GBA-negative controls	72	26 (36%)	56.3 ± 13.2	NA
GBA-negative PD patients	155	83 (54%)	64.3 ± 8.8	2.9 ± 2.8
GBA-positive PD patients	31	17 (55%)	58.5 ± 9.0	1.6 ± 1.5
GD Patients	30	16 (53%)	58.6 ± 13.4	NA
Non-affected GBA carriers	42	15 (36%)	54.3 ± 16.9	NA

Table 4: Demographics of the RAPSODI cohort. Age and PD duration are reported as mean ± standard deviation

2.3.2. Constipation

Answers to the questions "Does opening your bowels require a lot of effort?", "Do you suffer from hard stools?" and "Do you ever use laxatives?" are reported in Table 5, Figure 5 and Figure 6. To test for differences between the groups, ordinal logistic regression (OLR) was used, with age and sex as co-variates.

When comparing all 5 groups, the two PD groups answered "yes" and "sometimes" more often compared to the non-PD groups to all 3 questions (all p-values <0.05) and the *GBA*-positive PD patients also showed a higher prevalence of "yes" and "sometimes" compared to the *GBA*-negative PD group for the first question (p-value

0.0042). When comparing PD patients with non-affected individuals (without PD or GD), the former answered “yes” and “sometimes” more often than the latter to all 3 questions (p-values < 0.001).

When comparing only the PD patients with and without *GBA* variants, the *GBA* positive PD patients answered “yes” or “sometimes” to questions 1 more often than the *GBA* negative PD patients (p-values < 0.004).

Does opening your bowels require a lot of effort			
	No	**Sometimes**	**Yes**
GBA-negative controls	46 (65.7%)	22 (31.4%)	2 (2.9%)
GBA-negative PD patients	56 (36.6%)	80 (52.3%)	17 (11.1%)
GBA-positive PD patients	6 (19.4%)	16 (51.6%)	9 (29%)
GD Patients	20 (69%)	6 (20.7%)	3 (10.3%)
Non-affected GBA carriers	28 (75.7%)	8 (21.6%)	1 (2.7%)
Do you suffer from hard stools			
	No	**Sometimes**	**Yes**
GBA-negative controls	41 (58.6%)	26 (37.1%)	3 (4.3%)
GBA-negative PD patients	67 (43.8%)	74 (48.4%)	12 (7.8%)
GBA-positive PD patients	12 (38.7%)	11 (35.5%)	8 (25.8%)
GD Patients	15 (51.7%)	11 (37.9%)	3 (10.3%)
Non-affected GBA carriers	25 (67.6%)	12 (32.4%)	0
Do you ever use laxatives			
	No	**Sometimes**	**Yes**
GBA-negative controls	61 (87.1%)	4 (5.7%)	5 (7.1%)
GBA-negative PD patients	99 (64.7%)	25 (16.3%)	29 (19%)
GBA-positive PD patients	18 (58.1%)	5 (16.1%)	8 (25.8%)
GD Patients	24 (82.8%)	3 (10.3%)	2 (6.9%)
Non-affected GBA carriers	32 (86.5%)	3 (8.1%)	2 (5.4%)

Table 5: Constipation among the 5 groups, according to the answers to the RAPSODI questionnaire.

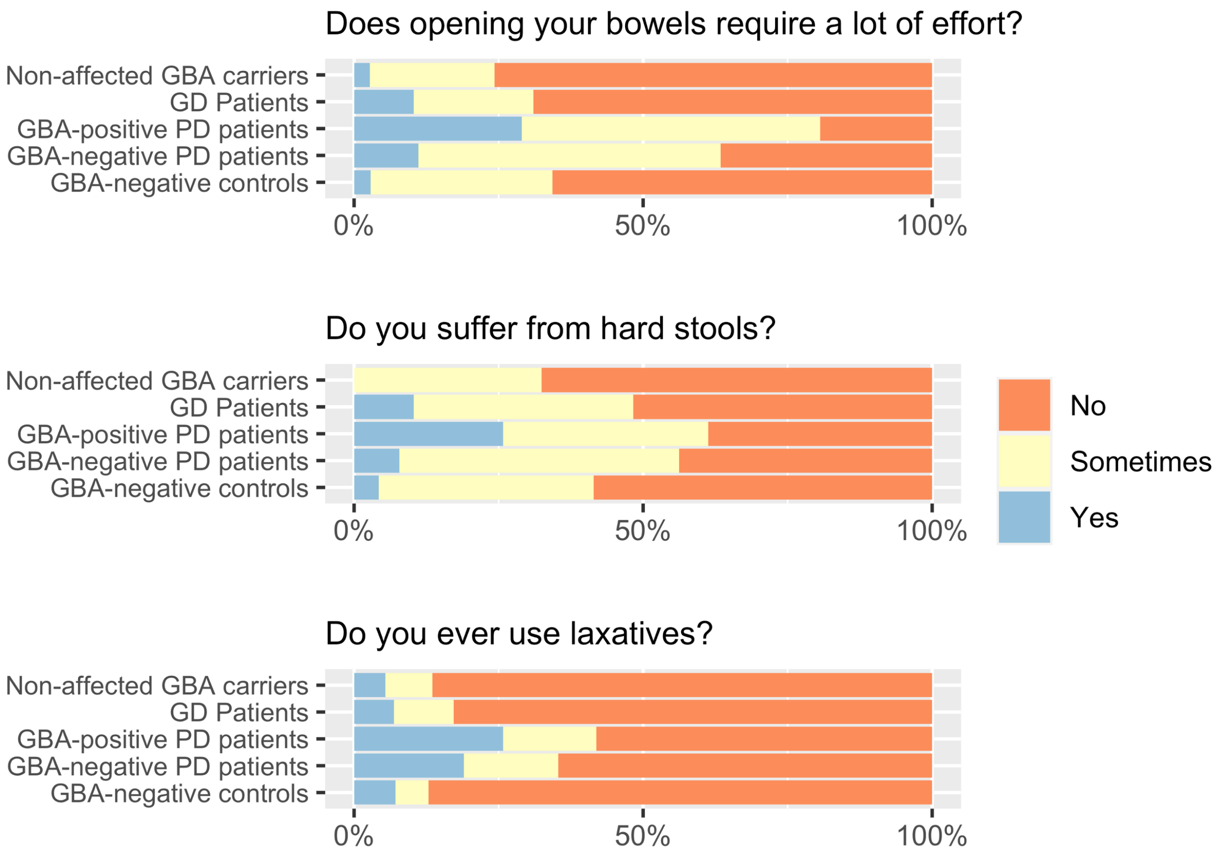


Figure 5: Constipation among the 5 groups, according to the answers to the RAPSODI questionnaire.

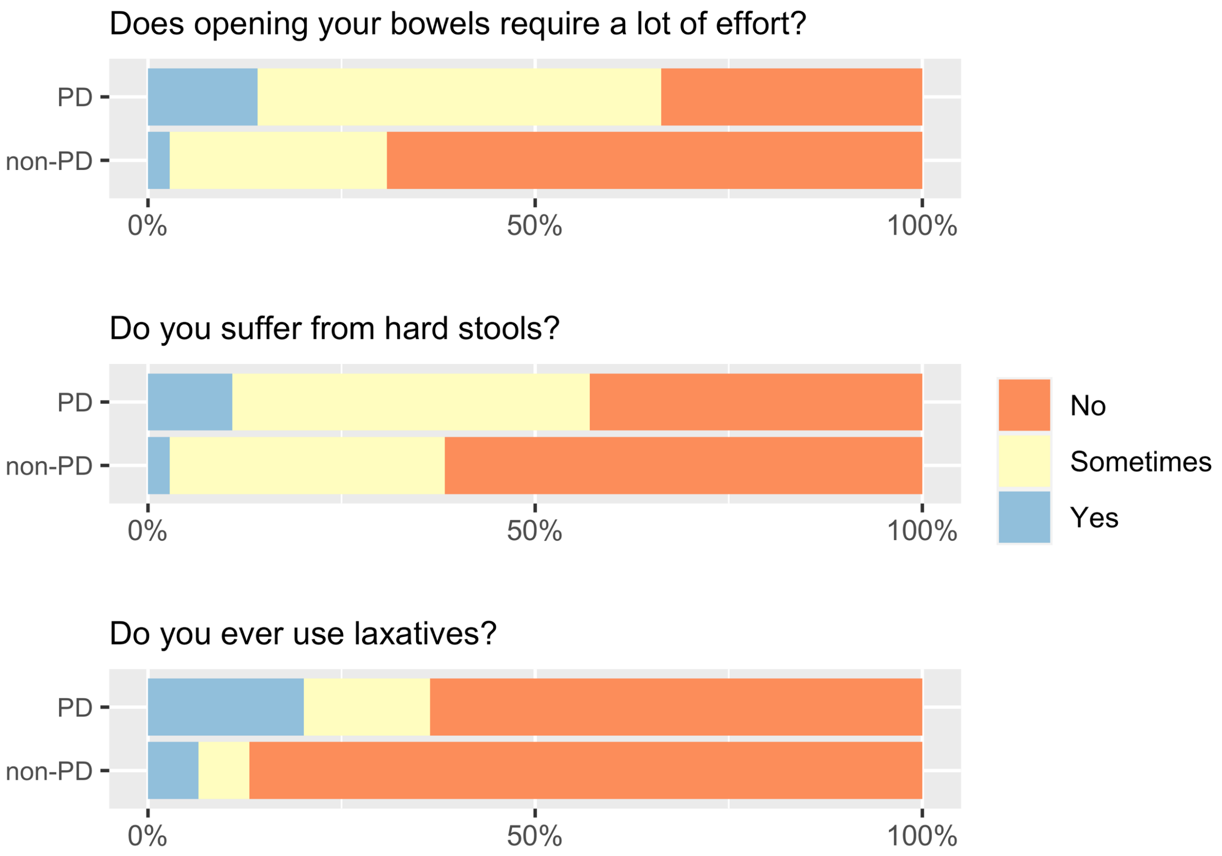


Figure 6: Constipation, PD vs non-PD participants.

2.3.3. Hospital Anxiety and Depression Scale

Numbers of participants, total HADS scores and depression and anxiety sub-scores for all groups are reported in Table 6 and Figure 7. Data were analysed using the cut-offs of the HADS scale (0-7 Normal, 8-10 Borderline and 11-21 Abnormal - Table 6, Figure 7 and Figure 7). To analyse inter-group differences and account for age and sex, I used ordinal logistic regression (OLR).

When comparing all 5 groups, the two PD groups showed a higher prevalence of depression and anxiety compared to all the non-PD groups (both p-values < 0.03). When comparing PD patients and non-affected participants, the former had a higher prevalence of both depression and anxiety (p-values < 0.002).

When comparing PD patients with and without *GBA* variants only, patients carrying *GBA* variants had a higher incidence of both depression and anxiety, but the difference was not significant.

Depression			
	Abnormal	**Borderline**	**Normal**
GBA-negative controls	0	3 (4.3%)	67 (95.7%)
GBA-negative PD patients	3 (2.1%)	17 (12.1%)	120 (85.7%)
GBA-positive PD patients	2 (8%)	4 (16%)	19 (76%)
GD Patients	1 (3.6%)	2 (7.1%)	25 (89.3%)
Non-affected GBA carriers	0	0	40 (100%)
Anxiety			
	Abnormal	**Borderline**	**Normal**
GBA-negative controls	1 (1.5%)	13 (19.1%)	54 (79.4%)
GBA-negative PD patients	16 (11.4%)	24 (17.1%)	100 (71.4%)
GBA-positive PD patients	3 (10%)	10 (33.3%)	17 (56.7%)
GD Patients	3 (10%)	0	27 (90%)
Non-affected GBA carriers	3 (7.7%)	3 (7.7%)	33 (84.6%)

Table 6: HADS depression and anxiety with cut-offs.

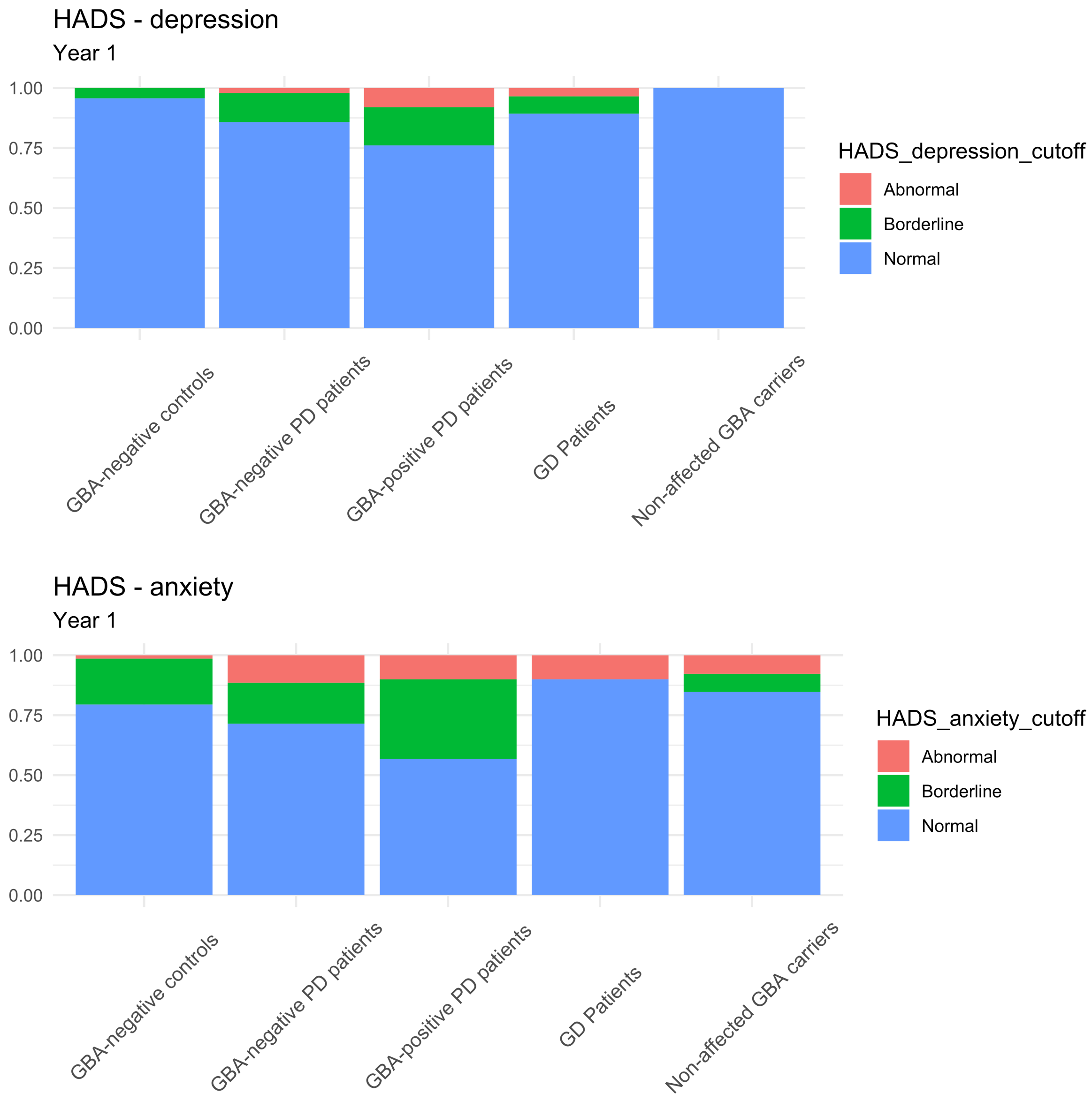


Figure 7: Prevalence of depression and anxiety according to the HADS scale.

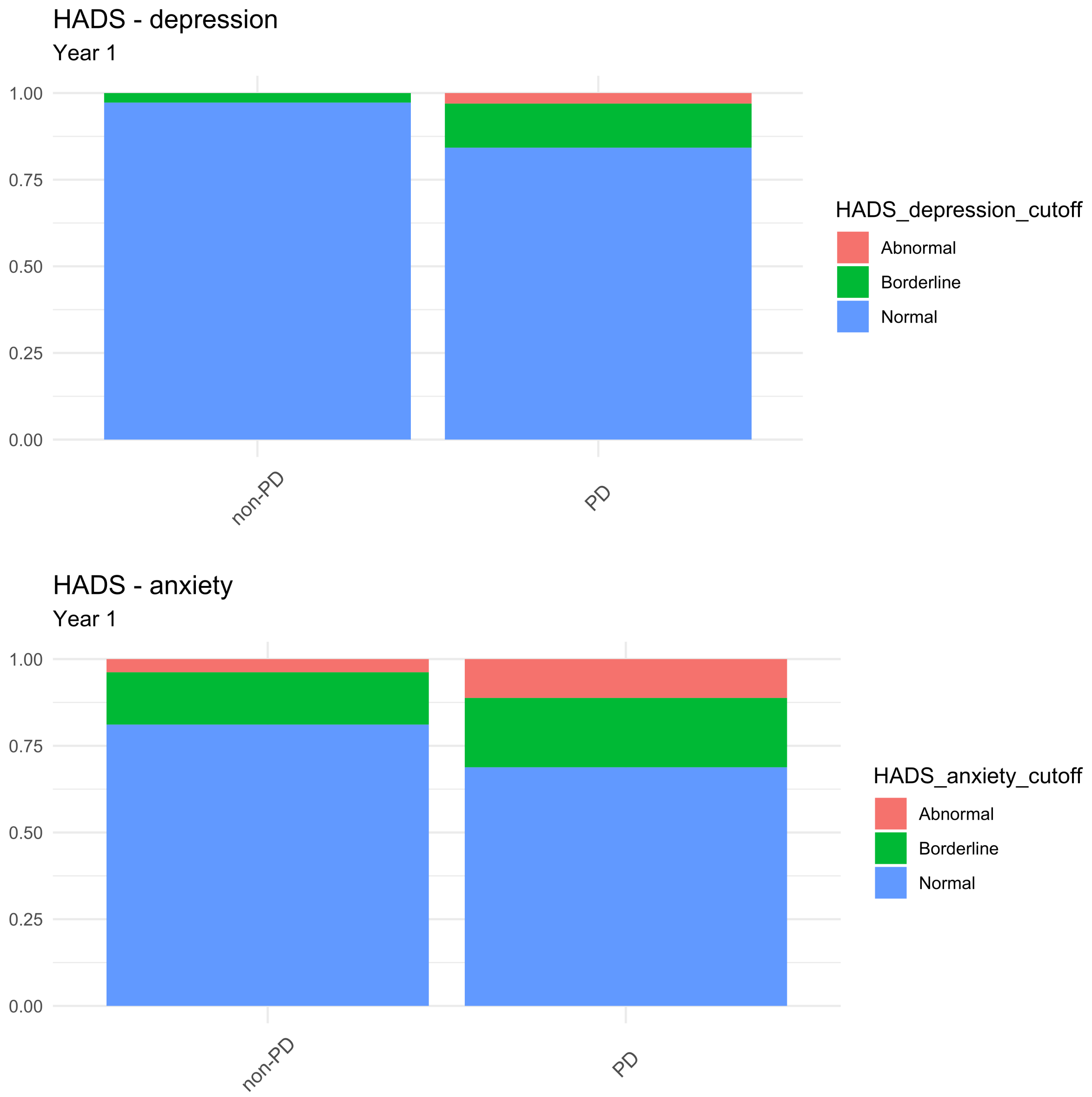


Figure 8: HADS results, PD vs non-PD participants.

2.3.4. REM Sleep Behaviour Disorder Screening Questionnaire

The RBDsq has been validated with a cut-off of 5 (scores equal or above 5 suggest a diagnosis of RBD). However, for PD patients question 10 ("I have / had a disease of the nervous system e.g. Stroke, head trauma, parkinsonism, restless leg syndrome, narcolepsy, depression, epilepsy, inflammatory disease of the brain") is meaningless, as they will all answer yes by default. For this reason, a cut-off of 6 can be used instead[139], or question 10 can be ignored and a cut-off of 5 used. Number of participants with RBD according to both cut-offs are reported in Table 7, Figure 9 and Figure 10. Logistic regression adjusted for age and sex was used for analysis.

When comparing all 5 groups, all 3 cut-offs showed a significantly higher prevalence of RBD in the two PD groups (p-values <0.001). When comparing PD patients and non-affected participants, the former had a higher prevalence of RBD using all 3 cut-offs. This difference was also significant when comparing these two PD groups alone and using the cut-off of 6 (p-value 0.045).

Cut-off 5		
	No	Yes
GBA-negative controls	61 (85.9%)	10 (14.1%)
GBA-negative PD patients	93 (60%)	62 (40%)
GBA-positive PD patients	14 (45.2%)	17 (54.8%)
GD Patients	26 (86.7%)	4 (13.3%)
Non-affected GBA carriers	36 (90.0%)	4 (10.0%)
Cut-off 6		
	No	Yes
GBA-negative controls	65 (91.5%)	6 (8.5%)
GBA-negative PD patients	107 (69%)	48 (31%)
GBA-positive PD patients	15 (48.4%)	16 (51.6%)
GD Patients	27 (90.0%)	3 (10.0%)
Non-affected GBA carriers	37 (92.5%)	3 (7.5%)
Cut-off 5, ignoring question 10		
	No	Yes
GBA-negative controls	63 (88.7%)	8 (11.3%)
GBA-negative PD patients	103 (66.5%)	52 (33.5%)
GBA-positive PD patients	15 (48.4%)	16 (51.6%)
GD Patients	27 (90.0%)	3 (10.0%)
Non-affected GBA carriers	36 (90.0%)	4 (10.0%)

Table 7: Prevalence of RBD (reported as number of participants and percentage) according to cut-off 5 and 6.

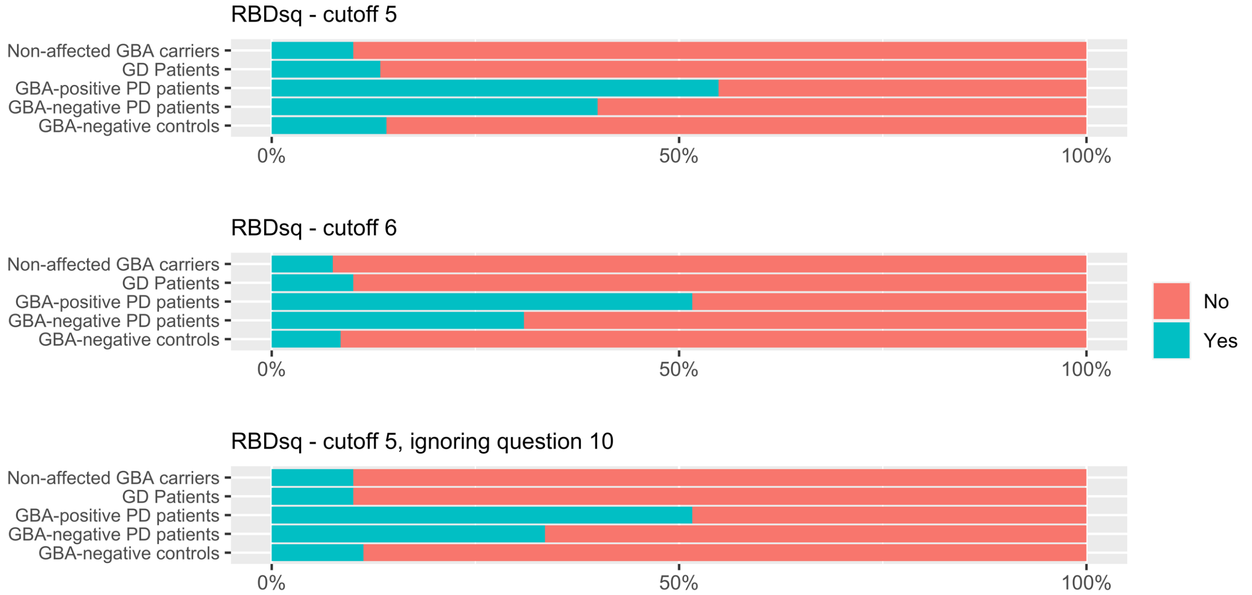


Figure 9: Proportion of subjects with RBD with cut-offs of 5 and 6.

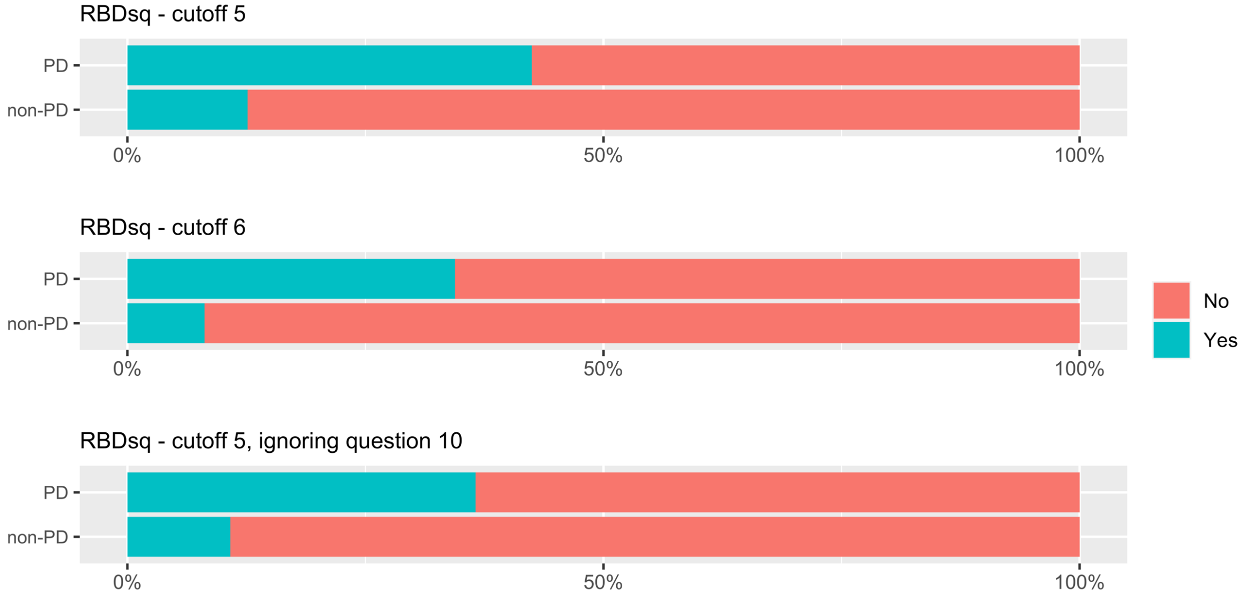


Figure 10: RBDsq, PD vs non-PD participants.

2.3.5. MDS-UPDRS part 2

Number of observations, means, SD and medians MDS-UPDRS part 2 for each group are reported in Table 8, Figure 11 and Figure 12. Given the semi-quantitative nature of the scale, linear regression was not the ideal approach. Values were divided into equal deciles and OLR was used instead. Inter-group comparisons after adjusting for age and sex revealed that the two PD groups had significantly higher scores compared to the non-PD groups (p-values < 0.001). When comparing PD patients and non-affected participants, the former had higher scores (p-value < 0.0001). No difference was observed between any of the non-PD groups or between PD with *GBA* and PD without *GBA*. It is important to mention that the MDS-UPDRS scale was designed for individuals with a diagnosis of PD and not for GD patients or the non-PD population. As such, comparison between different groups has limitations.

status	Count	Mean	Sd	Median
GBA-negative controls	71	0.9	2.5	0
GBA-negative PD patients	154	9.9	7.0	9
GBA-positive PD patients	31	10.9	6.4	11
GD Patients	23	2.1	2.8	1
Non-affected GBA carriers	33	1.0	1.5	0

Table 8: MDS-UPDRS part 2 results per each group.

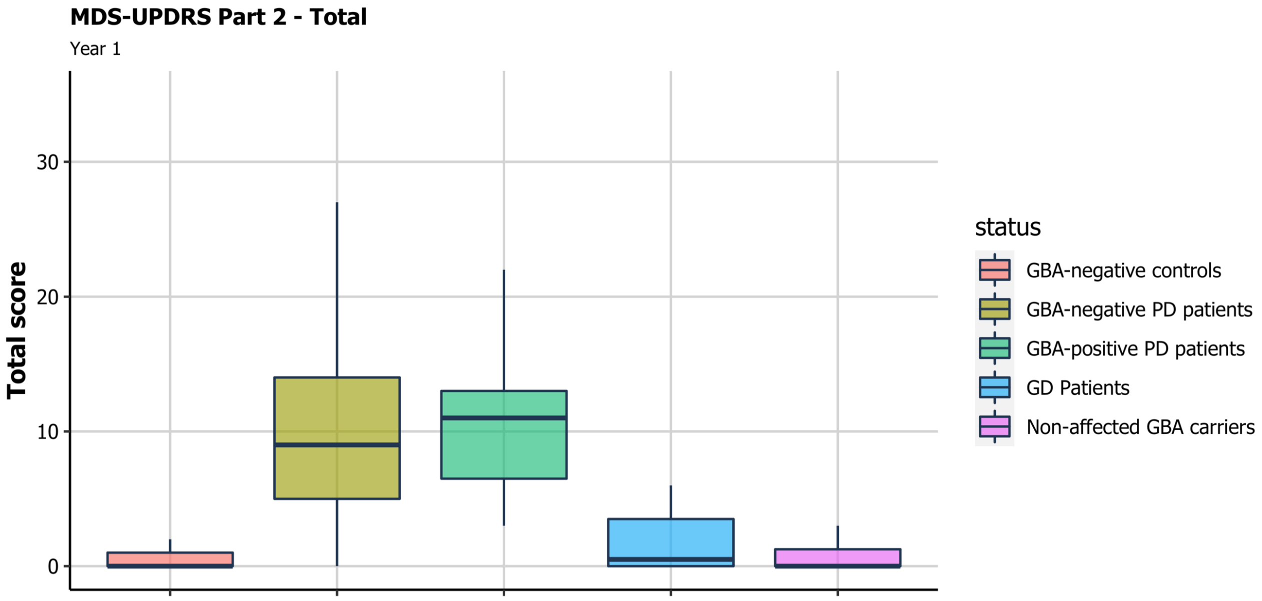


Figure 11: MDS-UPRS part 2 total scores across groups. The middle bar shows the median, the hinges the 25th and 75th percentiles and the whiskers include all values that are within 1.5 inter-quartile range from the hinges.

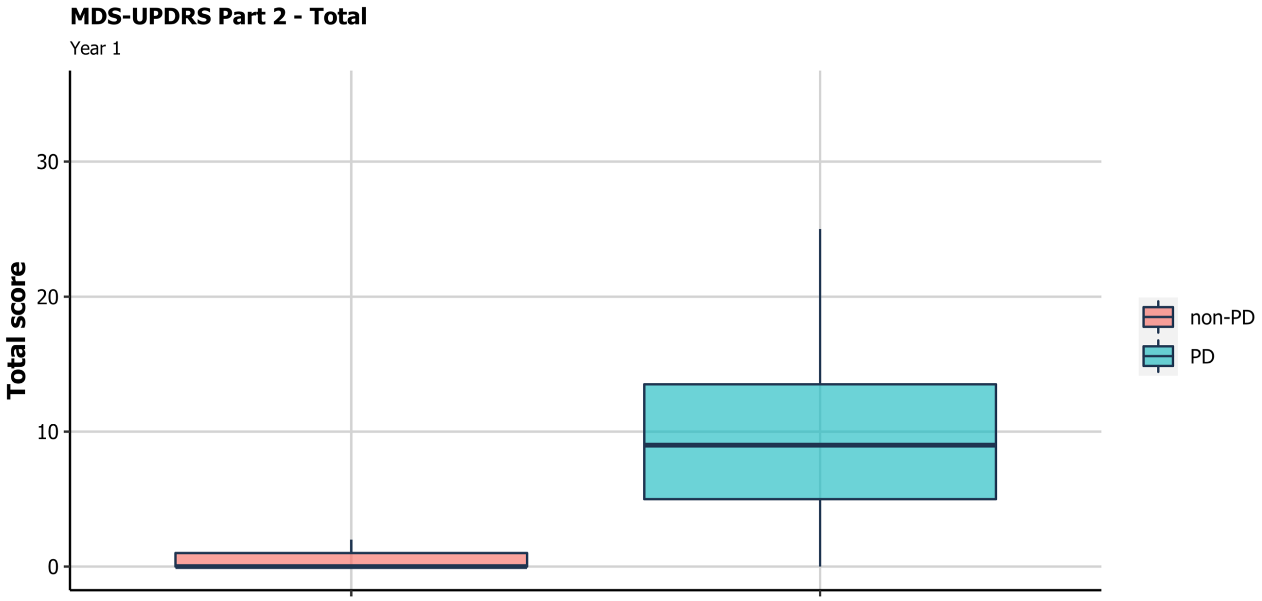


Figure 12: MDS-UPDRS part 2 total scores, PD vs non-PD participants. The middle bar shows the median, the hinges the 25th and 75th percentiles and the whiskers include all values that are within 1.5 inter-quartile range from the hinges.

2.3.6. UPSIT

Results of the baseline UPSIT scores are reported in Table 9 and in Figure 13. With linear regression, and using age and sex as covariates, the two PD groups showed a significantly lower score compared to the non-PD groups (p-values <0.0001). Interestingly, *GBA*-positive PD patients scored worse than *GBA*-negative PD patients (coefficient -3.05, p-value 0.016).

The cut-offs provided by the UPSIT manual identify different degrees of deficit: anosmia (0-18), severe microsmia (19-25), moderate microsmia (26-29 for males, 26-30 for females), mild microsmia (30-33 for males, 31-34 for females), normosmia (34-40 for males, 35-40 for females). Prevalence of the different degrees of hyposmia according to these cut-offs are reported in Table 11, Figure 14 and Figure 15. When comparing all 5 groups, OLR using age and sex as covariates revealed that the two PD groups had significantly worse sense of smell compared to the non-PD groups (p-values <0.0001) and that the *GBA*-positive PD patients had a worse performance compared to the *GBA*-negative PD patients (p-value 0.016). When comparing PD patients and non-affected participants, the former scored significantly worse (p-value < 0.0001). When comparing PD patients with and without *GBA* variants only, the former had worse olfaction (p-value 0.015).

status	Count	Mean	Sd	Median
GBA-negative controls	63	32.7936508	3.97637674	34
GBA-negative PD patients	145	21.8827586	7.67707939	22
GBA-positive PD patients	29	19.5517241	6.52241304	19
GD Patients	23	31.4347826	6.59769992	33
Non-affected GBA carriers	37	32.0810811	3.86852239	33

Table 9: Results of the UPSIT. Data are reported as number of observations, means, standard deviations (Sd) and medians.

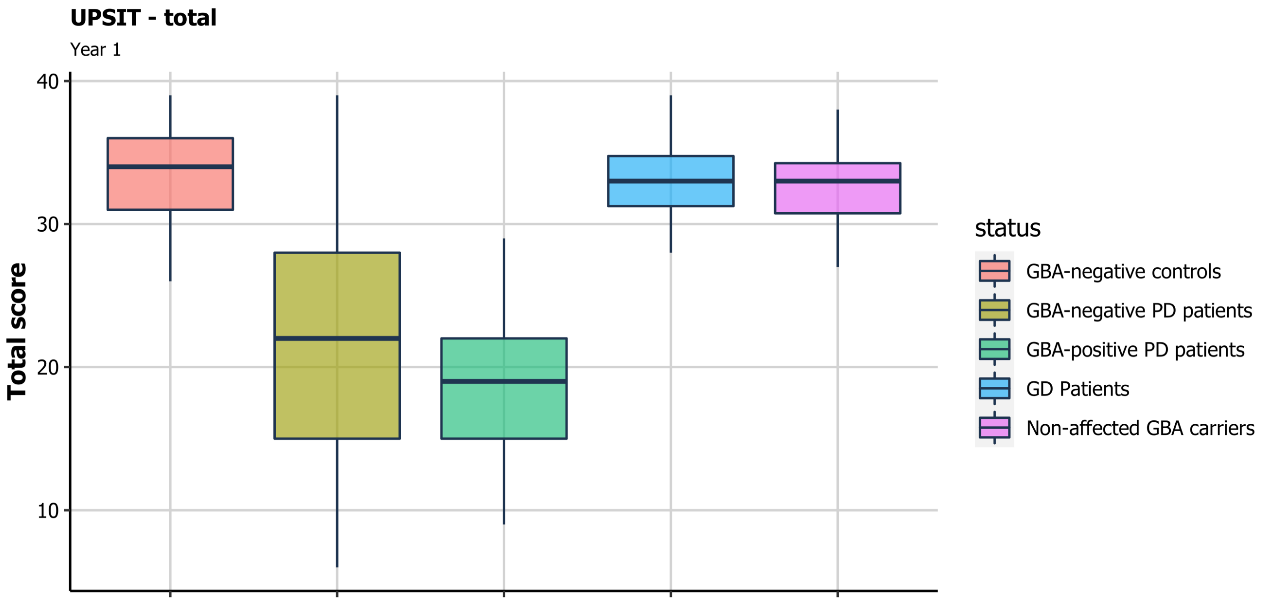


Figure 13: UPSIT scores at baseline. The middle bar shows the median, the hinges the 25th and 75th percentiles and the whiskers include all values that are within 1.5 inter-quartile range from the hinges.

	coefficient	p-value
GBA-negative PD patients-GBA-negative controls	-9.23	<0.0001
GBA-positive PD patients-GBA-negative controls	-12.29	<0.0001
GD Patients-GBA-negative controls	-0.85	0.5746
Non-affected GBA carriers-GBA-negative controls	-0.82	0.5211
GBA-positive PD patients-GBA-negative PD patients	-3.05	0.0164
GD Patients-GBA-negative PD patients	8.39	<0.0001
Non-affected GBA carriers-GBA-negative PD patients	8.41	<0.0001
GD Patients-GBA-positive PD patients	11.44	<0.0001
Non-affected GBA carriers-GBA-positive PD patients	11.47	<0.0001
Non-affected GBA carriers-GD Patients	0.03	0.9870

Table 10: Inter groups differences of UPSIT scores, estimated with linear regression with age and sex as covariates.

	Total anosmia	Severe microsmia	Moderate microsmia	Mild microsmia	Normosmia
GBA-negative controls	0	3 (4.8%)	10 (15.9%)	22 (34.9%)	28 (44.4%)
GBA-negative PD patients	52 (35.9%)	43 (29.7%)	25 (17.2%)	17 (11.7%)	8 (5.5%)
GBA-positive PD patients	14 (48.3%)	10 (34.5%)	3 (10.3%)	0	2 (6.9%)
GD Patients	2 (8.7%)	1 (4.3%)	2 (8.7%)	10 (43.5%)	8 (34.8%)
Non-affected GBA carriers	0	3 (8.1%)	5 (13.5%)	18 (48.6%)	11 (29.7%)

Table 11: Severity of hyposmia in the different groups at baseline, according to UPSIT scores.

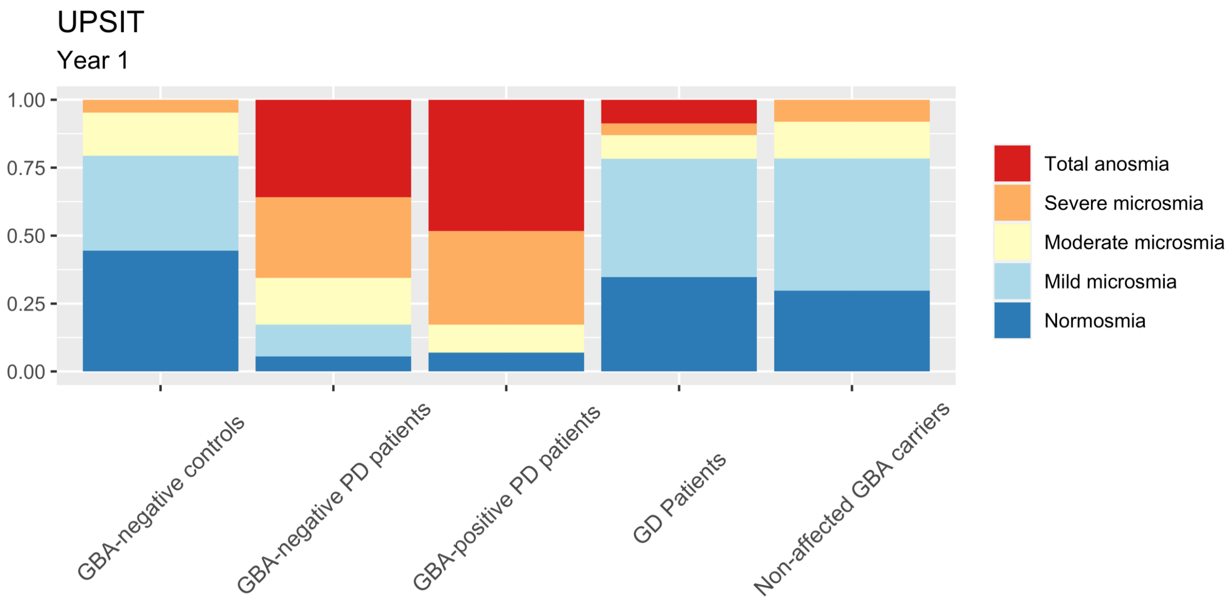


Figure 14: UPSIT results in the different groups at baseline.

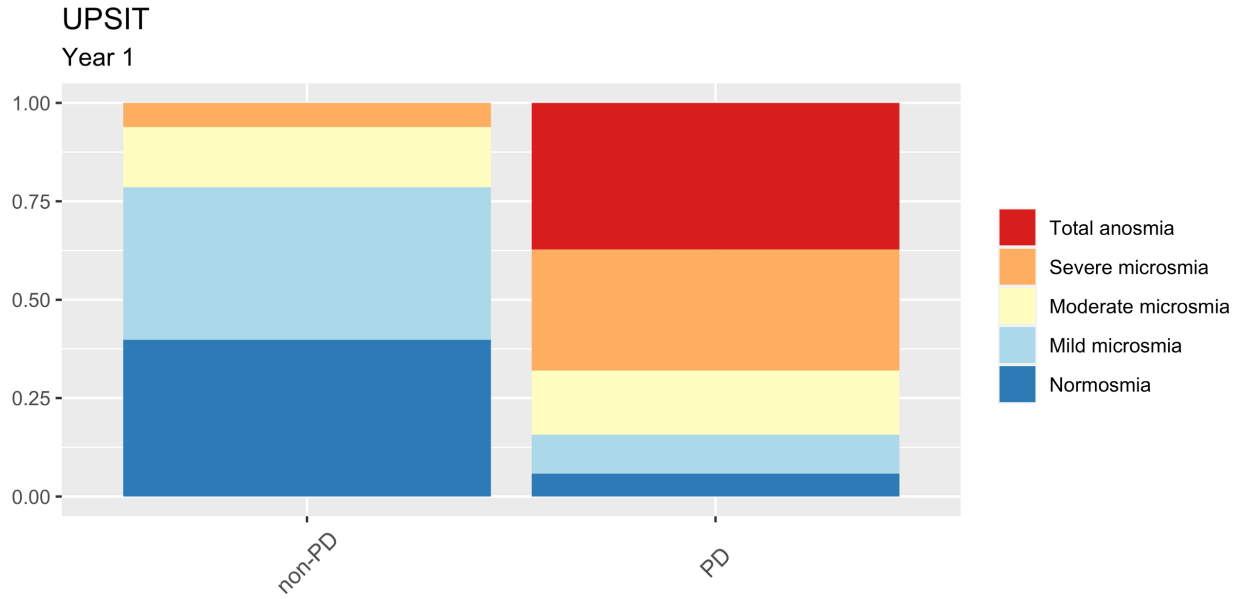


Figure 15: UPSIT results, PD vs non-PD participants.

2.3.7. Tap test

KS30 and AT30 scores for each group are reported in Table 12, Table 13, Figure 16 and Figure 15. Data are presented for dominant and non-dominant hand separately. Linear regression with age and sex as covariates was used for analysis. When comparing all 5 groups, the two PD groups had significantly worse scores compared to the non-PD groups for both dominant and non-dominant hand (p-values < 0.001). No differences were observed between the two PD groups (with and without *GBA* variants) or between any of the non-PD groups. When comparing all PD patients and non-affected participants, the former scored significantly worse (p-values < 0.0001). When comparing PD patients only, no differences were observed between *GBA* variants carriers and non-carriers.

Dominant hand				
status	Count	Mean	SD	Median
GBA-negative controls	70	62.7	13.1	64
GBA-negative PD patients	151	49.3	13.1	49
GBA-positive PD patients	28	48.6	13.8	46
GD Patients	21	64.3	12.7	62
Non-affected GBA carriers	31	63.7	11.7	64
Non-dominant hand				
status	Count	Mean	SD	Median
GBA-negative controls	69	54.9	9.8	56
GBA-negative PD patients	149	44.7	12.5	43
GBA-positive PD patients	29	41.7	11.9	44
GD Patients	21	56.2	12.8	56
Non-affected GBA carriers	31	56.9	10.5	60

Table 12: KS30 score for dominant and non-dominant hand. SD: standard deviation.

Dominant hand				
status	Count	Mean	SD	Median
GBA-negative controls	70	89.12	27.34	82.79
GBA-negative PD patients	151	103.65	39.81	94.62
GBA-positive PD patients	28	106.77	34.59	95.29
GD Patients	21	93.70	45.00	80.71
Non-affected GBA carriers	31	81.60	29.14	76.29
Non-dominant hand				
status	Count	Mean	SD	Median
GBA-negative controls	69	109.17	30.50	103.82
GBA-negative PD patients	149	130.62	42.98	122.07
GBA-positive PD patients	29	163.32	184.18	116.48
GD Patients	21	114.81	45.44	100.69
Non-affected GBA carriers	31	103.96	40.90	91.80

Table 13: AT30 score for dominant and non-dominant hand. SD: standard deviation.

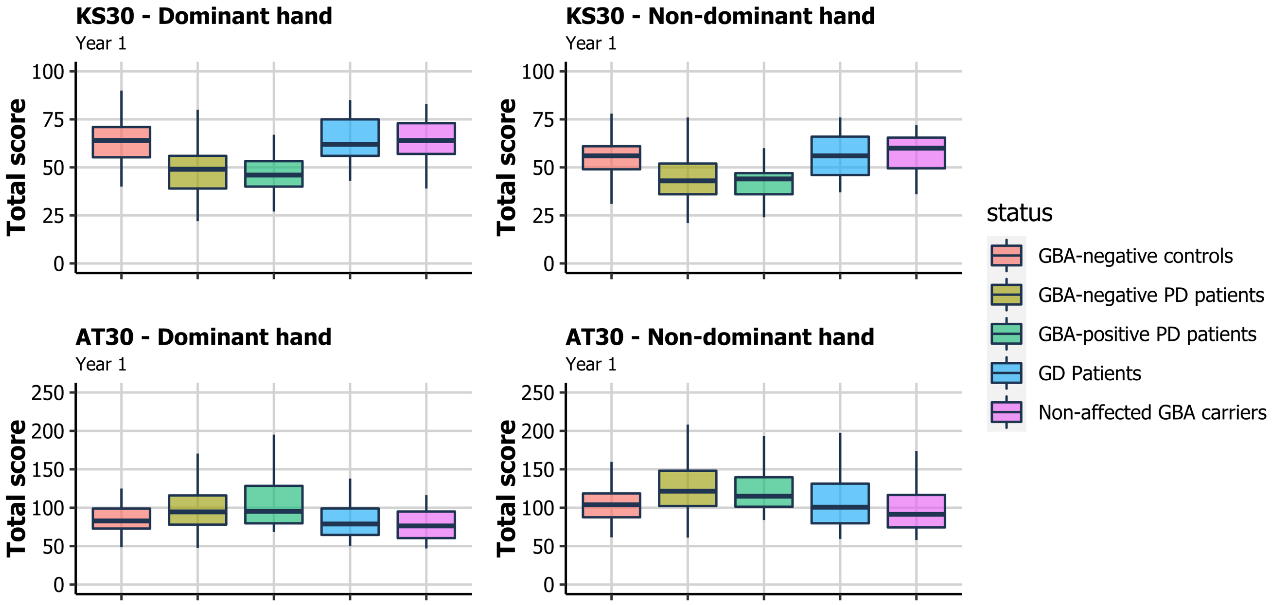


Figure 16: KS30 and AT30 scores. The middle bar shows the median, the hinges the 25th and 75th percentiles and the whiskers include all values that are within 1.5 inter-quartile range from the hinges.

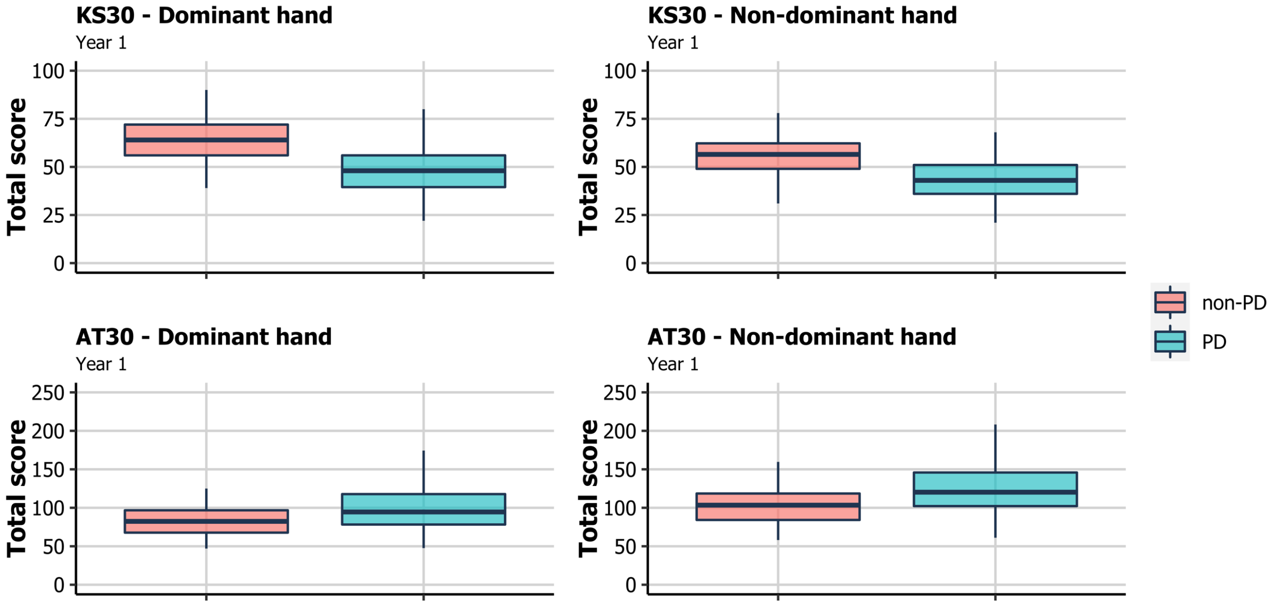


Figure 17: Tap-test results, PD patients and non-PD participants. The middle bar shows the median, the hinges the 25th and 75th percentiles and the whiskers include all values that are within 1.5 inter-quartile range from the hinges.

2.3.8. Cognitive assessment

2.3.8.1. Picture recognition test

DPICOACC and DPICNACC scores for each group are reported in Table 14, Figure 18 and Figure 19.

As they represent proportions of correct answers, both scores were analysed with quasibinomial regression, with sex, age and years of education as covariates. When comparing all 5 groups, DPICOACC was significantly lower in the *GBA*-positive PD patients compared to all the non-PD groups (p-values < 0.05) and in *GBA*-negative PD patients compared to *GBA*-negative controls (p-value 0.005). DPICNACC was significantly lower in *GBA*-positive PD patients compared to all the non-PD groups (p-values < 0.02) and in *GBA*-negative PD patients compared to non-affected *GBA* carriers (p-value 0.045). When comparing PD patients and non-affected participants, the former had worse scores for both DPICOACC and DPICNACC (p-values 0.002 and 0.028, respectively).

When considering the two PD groups only, both DPICOACC and DPICNACC were worse in *GBA*-positive PD patients compared to *GBA*-negative PD patients (p-values 0.044 and 0.033, respectively).

DPICOACC				
status	**Count**	**Mean (%)**	**SD**	**Median**
GBA-negative controls	69	91.8	7.8	95
GBA-negative PD patients	150	87.1	10.5	90
GBA-positive PD patients	30	82.8	16.6	90
GD Patients	24	89.6	10.2	90
Non-affected GBA carriers	34	88.8	10.9	90
DPICNACC				
status	**Count**	**Mean (%)**	**SD**	**Median**
GBA-negative controls	69	77.0	16.0	80
GBA-negative PD patients	150	72.5	15.6	75
GBA-positive PD patients	30	68.2	17.3	70
GD Patients	24	76.7	16.3	80
Non-affected GBA carriers	34	80.3	16.0	82.5

Table 14: Picture recognition test baseline scores for each group. SD: Standard deviation.

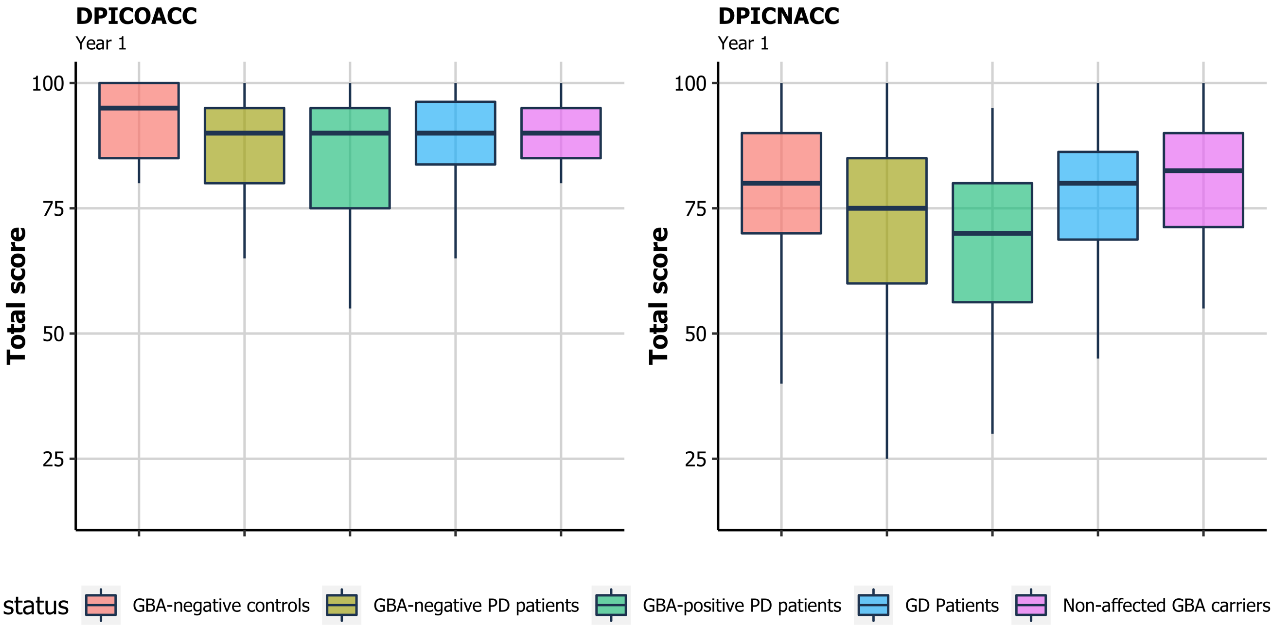


Figure 18: Picture recognition test scores. The middle bar shows the median, the hinges the 25th and 75th percentiles and the whiskers include all values that are within 1.5 inter-quartile range from the hinges.

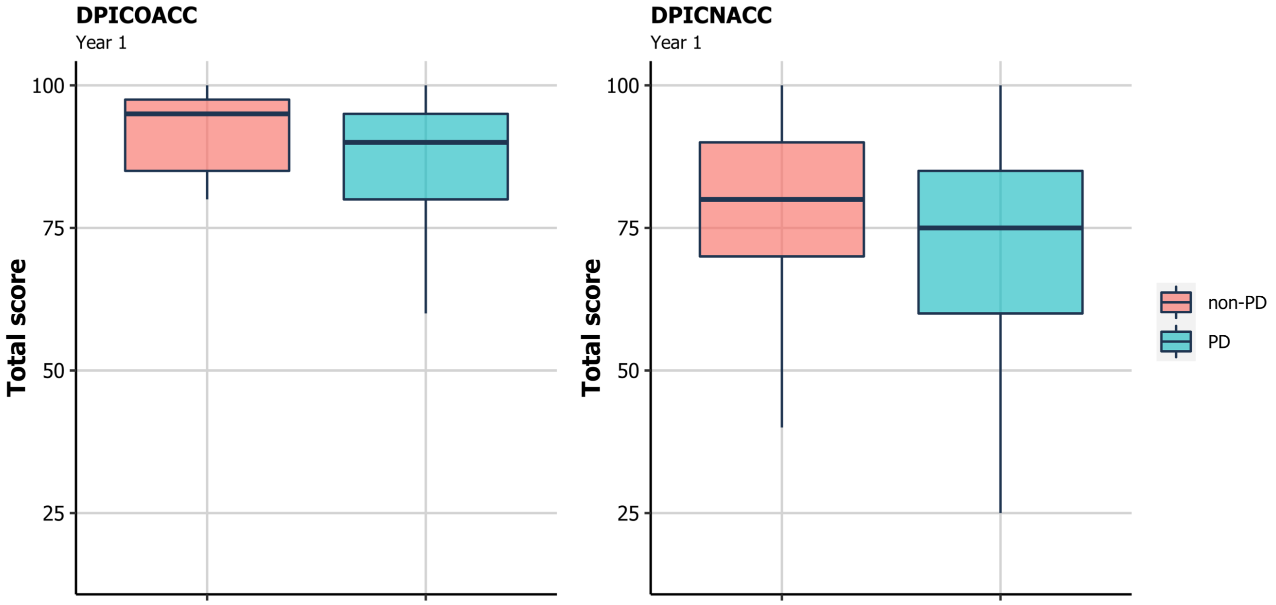


Figure 19: Picture recognition test, PD patients vs non-PD participants. The middle bar shows the median, the hinges the 25th and 75th percentiles and the whiskers include all values that are within 1.5 inter-quartile range from the hinges.

2.3.8.2. Simple reaction time

Results of the simple reaction time test are reported in Table 15, Figure 20 and Figure 21. Linear regression was used with sex, age and years of education as covariates. When comparing all 5 groups, *GBA*-positive PD patients scored significantly worse than *GBA*-negative controls and non-affected *GBA* carriers (p-values 0.0496 and 0.014, respectively), and *GBA*-negative PD patients scored worse than non-affected *GBA* carriers (p-value 0.016). When comparing PD patients and non-affected participants, the former scored significantly worse (p-value 0.007). When comparing PD patients only, no differences were observed between carriers and non-carriers.

status	Count	Mean	Sd	Median
GBA-negative controls	69	348.3	58.3	337.2
GBA-negative PD patients	150	382.7	108.2	363.7
GBA-positive PD patients	30	388.1	79.1	364.3
GD Patients	24	350.0	51.1	342.9
Non-affected GBA carriers	34	328.8	47.5	330.1

Table 15: Simple reaction time test results at baseline.

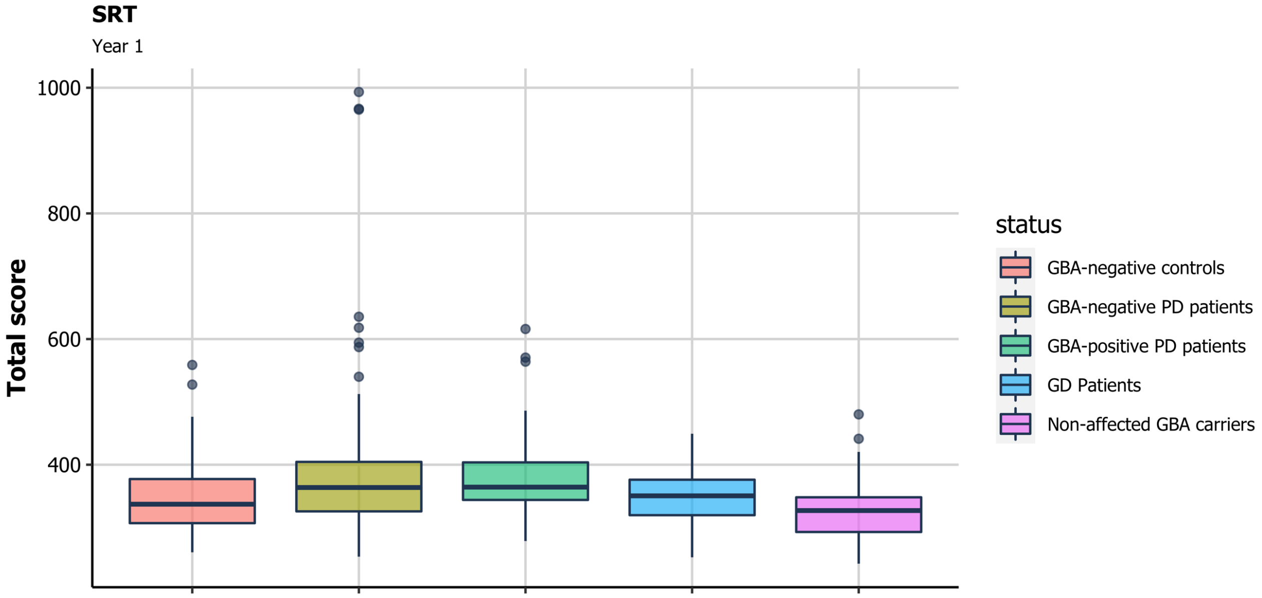


Figure 20: Simple reaction time test at baseline. The middle bar shows the median, the hinges the 25th and 75th percentiles and the whiskers include all values that are within 1.5 inter-quartile range from the hinges. Dots represent observation outside the 1.5 inter-quartile range from the hinges.

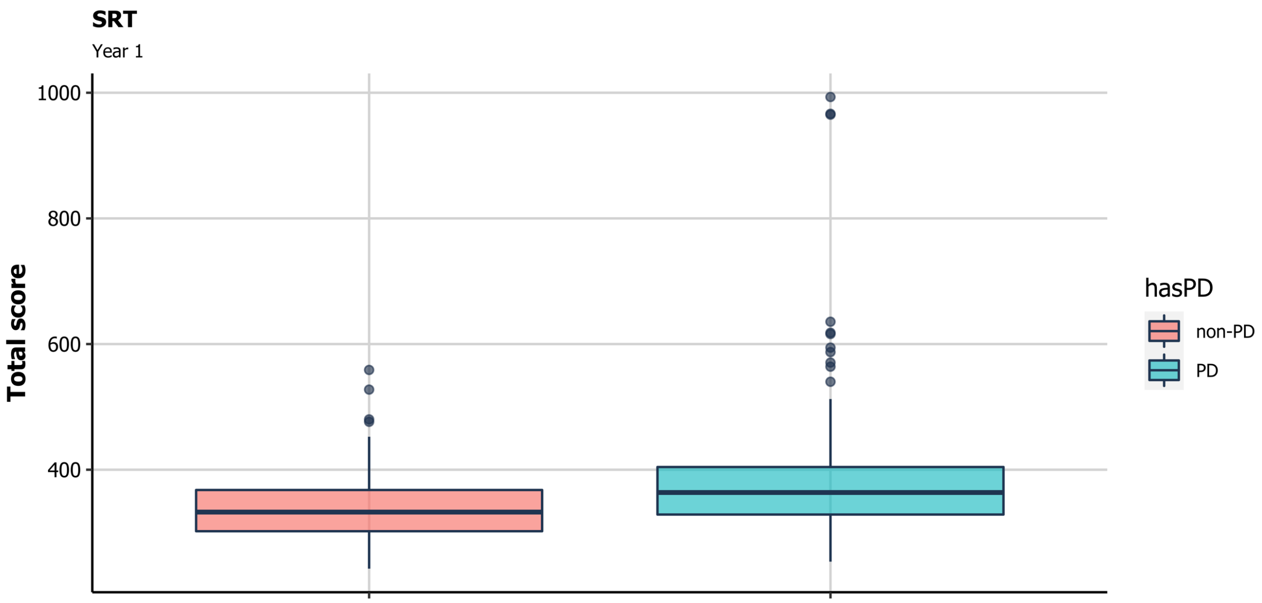


Figure 21: Simple reaction time test, PD patients vs non-affected participants. The middle bar shows the median, the hinges the 25th and 75th percentiles and the whiskers include all values that are within 1.5 inter-quartile range from the hinges. Dots represent observation outside the 1.5 inter-quartile range from the hinges.

2.3.8.3. Choice reaction time test

Results of the choice reaction time test are reported in Table 16, Figure 22 and Figure 23.

The CRTACC score was analysed with quasibinomial regression with age, sex and years of education as covariates, CRT was analysed with linear regression, with age, sex and years of education as covariates. When comparing all 5 groups, *GBA*-positive PD patients scored significantly worse than *GBA*-negative controls for CRTACC (p-value 0.006).

When comparing PD patients and non-affected participants, no statistical differences were observed. When considering only *GBA*-positive and *GBA*-negative PD patients, the former had worse CRT (p-value 0.006).

CRTACC				
status	Count	Mean	SD	Median
GBA-negative controls	69	97.2	2.4	98
GBA-negative PD patients	150	96.1	3.8	96
GBA-positive PD patients	30	94.7	7.1	97
GD Patients	24	96.3	2.7	97
Non-affected GBA carriers	34	96.6	3.1	98
CRT				
status	Count	Mean	SD	Median
GBA-negative controls	69	514.6	77.5	506.4
GBA-negative PD patients	150	547.6	127.9	516.1
GBA-positive PD patients	30	561.7	92.7	554.1
GD Patients	24	517.7	52.4	512.3
Non-affected GBA carriers	34	490.9	83.6	486.9

Table 16: Choice reaction time test results. SD: Standard deviation.

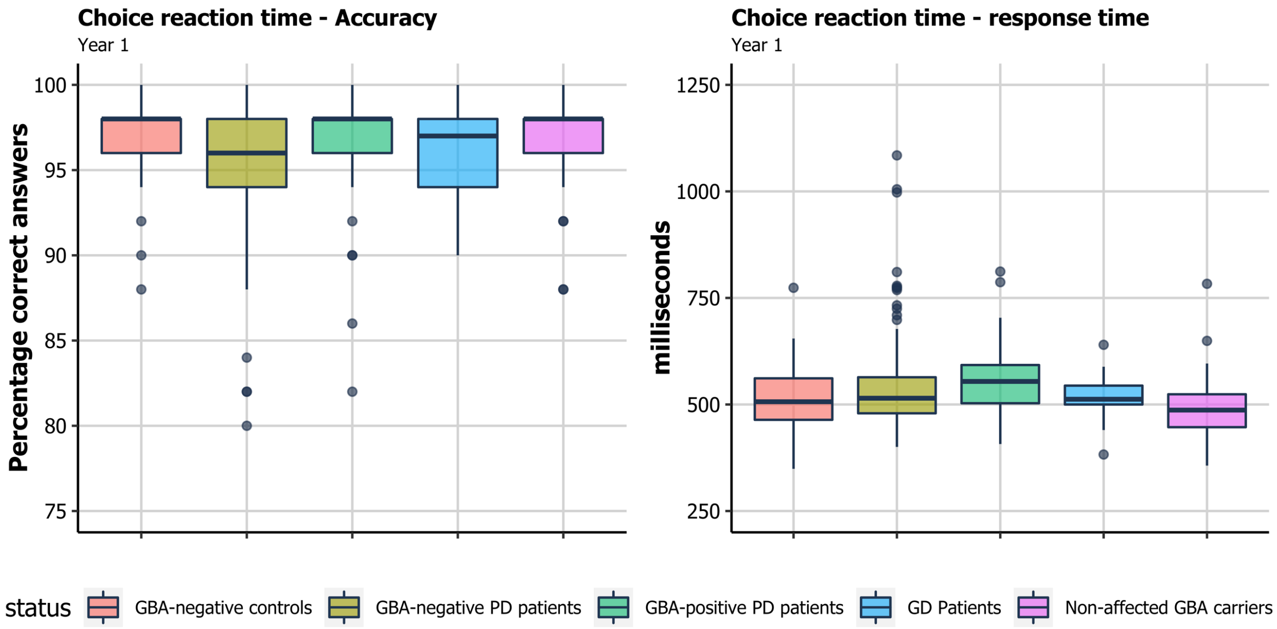


Figure 22: Results of the Choice reaction time test by groups. The middle bar shows the median, the hinges the 25th and 75th percentiles and the whiskers include all values that are within 1.5 inter-quartile range from the hinges. Dots represent observation outside the 1.5 inter-quartile range from the hinges.

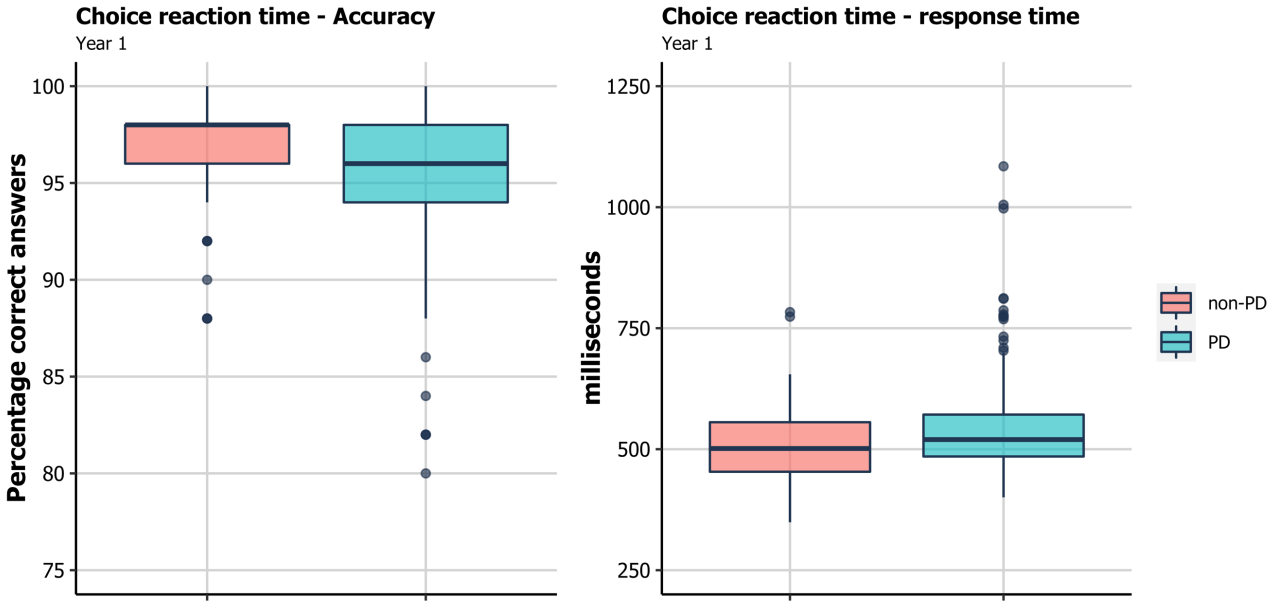


Figure 23: Choice reaction time test, PD patients vs non-affected participants. The middle bar shows the median, the hinges the 25th and 75th percentiles and the whiskers include all values that are within 1.5 inter-quartile range from the hinges. Dots represent observation outside the 1.5 inter-quartile range from the hinges.

2.3.8.4. Digit vigilance

Results of the digit vigilance test are reported in Table 17, Figure 24 and Figure 25. VIGACC was analysed with quasibinomial regression with age, sex and years of education as covariates. VIGRT was analysed with linear regression with age, sex and years of education as covariates.

No significant differences between the groups were observed, either before or after removing the outliers.

VIGACC				
status	Count	Mean	SD	Median
GBA-negative controls	69	96.5	5.4	97.8
GBA-negative PD patients	150	95.2	11.1	97.8
GBA-positive PD patients	30	95.3	5.0	96.6
GD Patients	24	97.0	5.2	100.0
Non-affected GBA carriers	34	97.8	3.5	100.0
VIGRT				
status	Count	Mean	SD	Median
GBA-negative controls	69	558.8	129.3	524.3
GBA-negative PD patients	150	591.3	261.4	533.1
GBA-positive PD patients	30	561.5	111.3	534.0
GD Patients	24	485.0	55.6	477.5
Non-affected GBA carriers	34	504.4	90.3	484.3

Table 17: Digit vigilance reaction time test results. SD: Standard deviation.

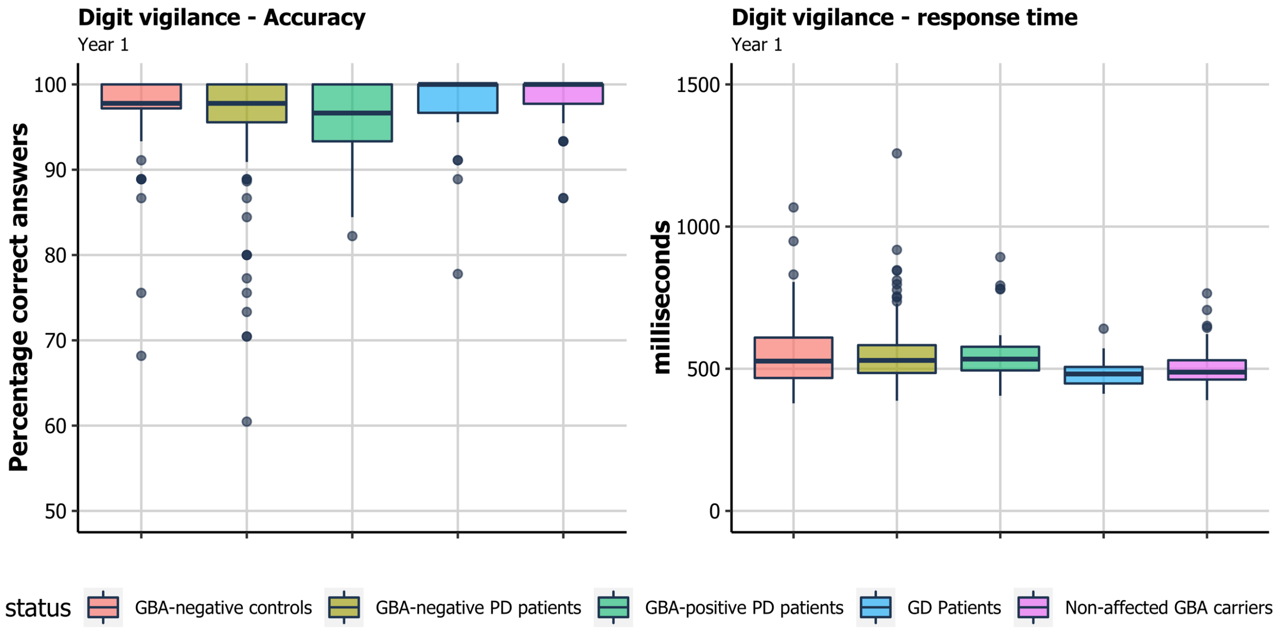


Figure 24: Digit vigilance test results. The middle bar shows the median, the hinges the 25th and 75th percentiles and the whiskers include all values that are within 1.5 inter-quartile range from the hinges. Dots represent observation outside the 1.5 inter-quartile range from the hinges.

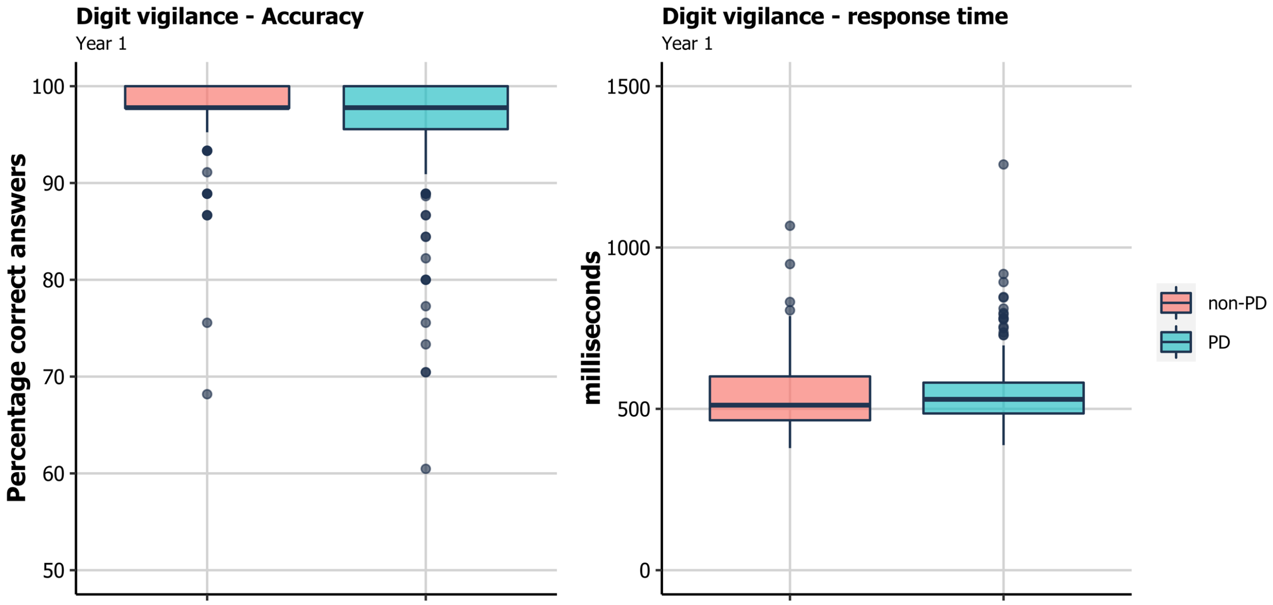


Figure 25: Digit vigilance test, PD patients vs non-affected participants. The middle bar shows the median, the hinges the 25th and 75th percentiles and the whiskers include all values that are within 1.5 inter-quartile range from the hinges. Dots represent observation outside the 1.5 inter-quartile range from the hinges.

2.3.8.5. Spatial working memory test

Results of the spatial working memory test are reported in Table 18, Figure 26 and Figure 27. SPMOACC and SPMNACC scores were analysed with quasibinomial regression with age, sex and years of education as covariates; SPMRT was analysed with linear regression, with age, sex and years of education as covariates.

When comparing all 5 groups, the only significant difference for SPMOACC was between *GBA*-positive PD patients and non-affected *GBA* carriers (p-value 0.047). For SPMNACC, *GBA*-positive PD patients showed a statistically significant lower score compared to all other non-PD groups (p-values < 0.04), while *GBA*-negative PD patients performed worse than *GBA*-negative controls and non-affected *GBA* carriers (p-values 0.047 and 0.49, respectively).

When comparing PD patients and non-affected participants, the former showed worse SPMNACC (p-value 0.004).

When comparing PD patients only, no statistically significant differences between *GBA* carriers and non-carriers were detected.

No differences in SRT were detected in any of the comparisons.

SPMOACC				
status	**Count**	**Mean**	**Sd**	**Median**
GBA-negative controls	68	92.9	9.5	93.8
GBA-negative PD patients	146	89.3	15.5	93.8
GBA-positive PD patients	30	89.0	15.1	93.8
GD Patients	23	93.2	11.3	100.0
Non-affected GBA carriers	31	95.4	6.6	100.0
SPMNACC				
status	**Count**	**Mean**	**Sd**	**Median**
GBA-negative controls	68	94.9	8.7	100.0
GBA-negative PD patients	146	87.6	19.5	95.0
GBA-positive PD patients	30	86.8	17.3	95.0
GD Patients	23	94.6	11.0	100.0
Non-affected GBA carriers	31	96.3	6.9	100.0
SPMRT				
status	**Count**	**Mean**	**Sd**	**Median**
GBA-negative controls	68	1027.2	285.1	935.1
GBA-negative PD patients	146	1153.3	410.5	1059.0
GBA-positive PD patients	30	1163.0	365.2	1038.3
GD Patients	23	1028.6	239.5	1019.4
Non-affected GBA carriers	31	1070.0	460.5	914.8

Table 18: Results of spatial working memory test. SD: standard deviation.

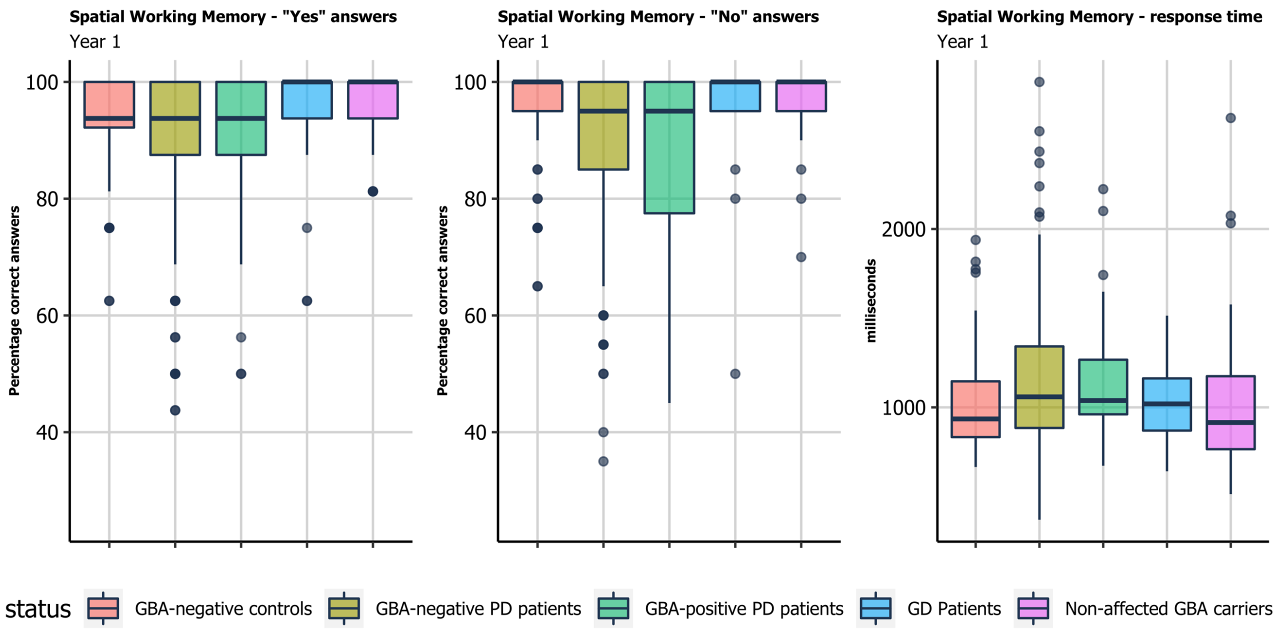


Figure 26: Spatial working memory test results. The middle bar shows the median, the hinges the 25th and 75th percentiles and the whiskers include all values that are within 1.5 inter-quartile range from the hinges. Dots represent observation outside the 1.5 inter-quartile range from the hinges.

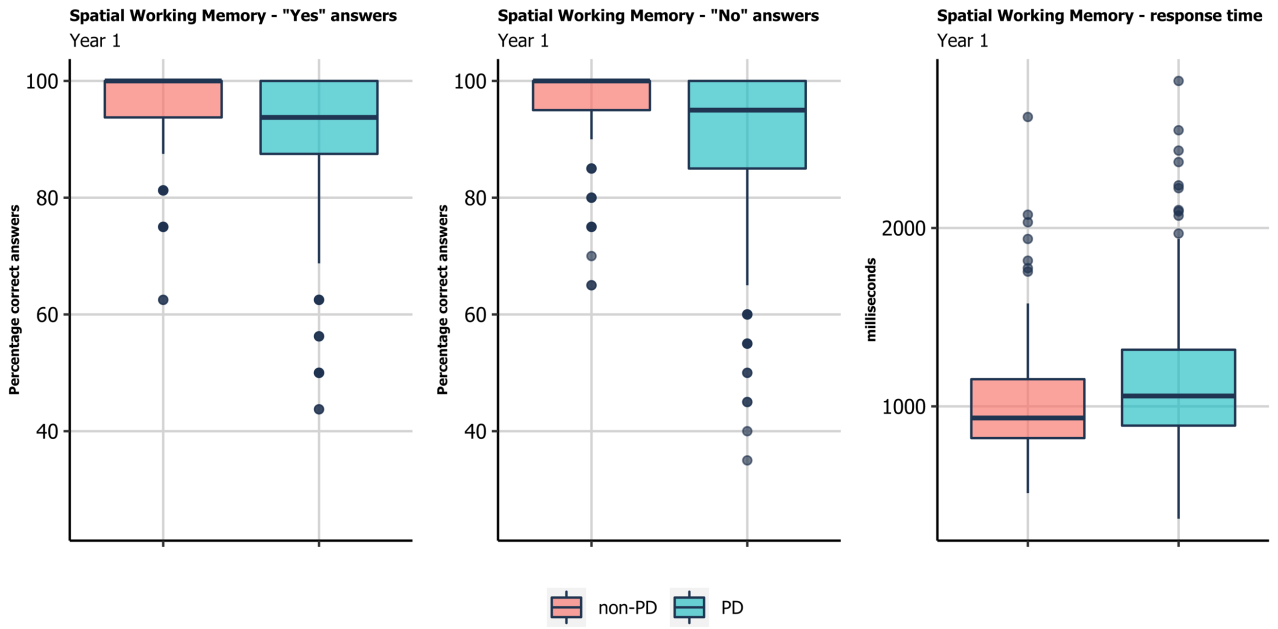


Figure 27: Spatial working memory test, PD patients vs non-affected participants. The middle bar shows the median, the hinges the 25th and 75th percentiles and the whiskers include all values that are within 1.5 inter-quartile range from the hinges. Dots represent observation outside the 1.5 inter-quartile range from the hinges.

2.3.8.6. Numeric working memory

Results of the numeric working memory test are reported in Table 19, Figure 28 and Figure 29. NWMOACC and NWMNACC scores were analysed with quasibinomial regression with age, sex and years of education as covariates. NMWRT was analysed with linear regression, with age, sex and years of education as covariates after removal of outliers (observations more than 1.5 interquartile range from the 75th percentile).

When comparing all 5 groups, for NWMOACC both PD groups scored significantly lower than non-affected *GBA* carriers (p-values <0.03) and the *GBA*-negative PD patients also scored lower than the *GBA*-negative controls (p-values 0.024). For NWMNACC, the *GBA*-positive PD patients scored significantly lower than all the non-PD groups (p-values < 0.04) and the *GBA*-negative PD patients scored significantly lower than the *GBA*-negative controls (p-value 0.03).

When comparing PD patients and non-affected participants, the former performed worse in both NWMOACC and NWMNACC (p-values 0.0035 and 0.0035, respectively).

When comparing PD patients only, no significant differences were observed between *GBA* carriers and non-carriers.

No statistically significant differences in NMWRT between the groups were detected in any of the comparisons.

NWMOACC				
status	**Count**	**Mean**	**SD**	**Median**
GBA-negative controls	68	93.7	9.0	96.7
GBA-negative PD patients	147	89.6	14.4	93.3
GBA-positive PD patients	30	89.6	10.2	93.3
GD Patients	23	93.3	6.4	93.3
Non-affected GBA carriers	31	96.1	5.9	100.0
NWMNACC				
status	**Count**	**Mean**	**SD**	**Median**
GBA-negative controls	68	97.5	7.6	100.0
GBA-negative PD patients	147	93.2	13.8	100.0
GBA-positive PD patients	30	89.6	13.6	93.3
GD Patients	23	96.8	5.6	100.0
Non-affected GBA carriers	31	97.2	5.1	100.0
NMWRT				
status	**Count**	**Mean**	**SD**	**Median**
GBA-negative controls	68	860.8	203.3	825.3
GBA-negative PD patients	147	952.8	255.4	897.1
GBA-positive PD patients	30	959.2	234.4	975.7
GD Patients	23	944.1	224.3	890.8
Non-affected GBA carriers	31	897.1	280.2	881.7

Table 19: Results of the numeric working memory test. SD: standard deviation.

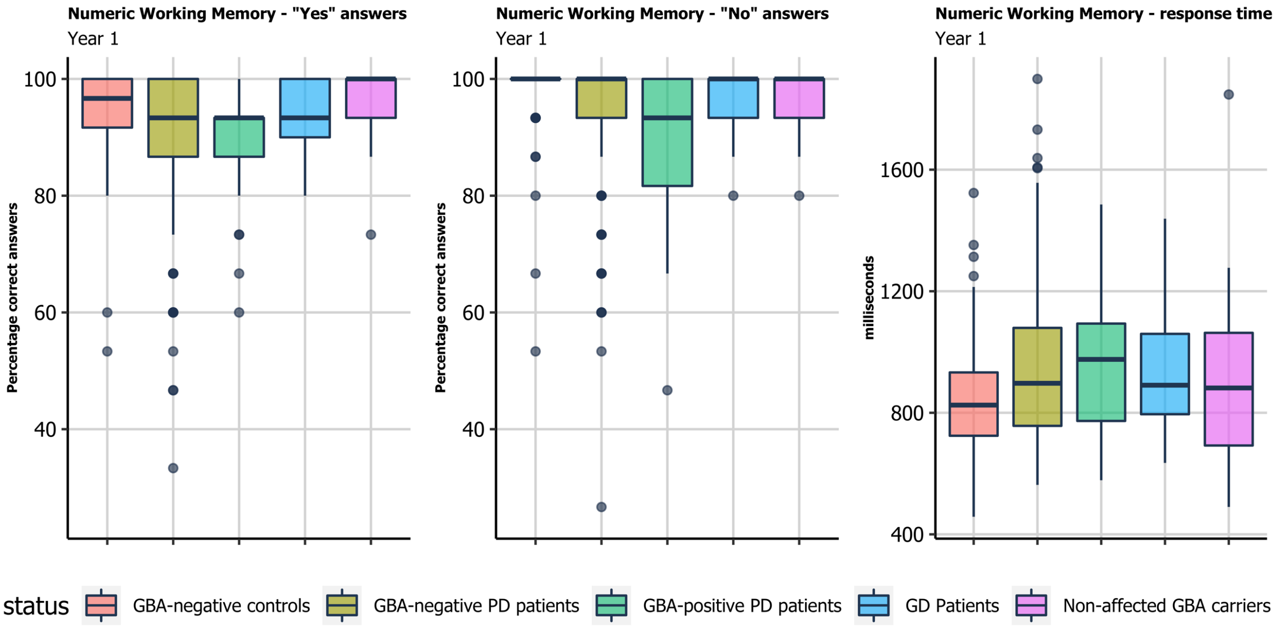


Figure 28: Results of the numeric working memory test. The middle bar shows the median, the hinges the 25th and 75th percentiles and the whiskers include all values that are within 1.5 inter-quartile range from the hinges. Dots represent observation outside the 1.5 inter-quartile range from the hinges.

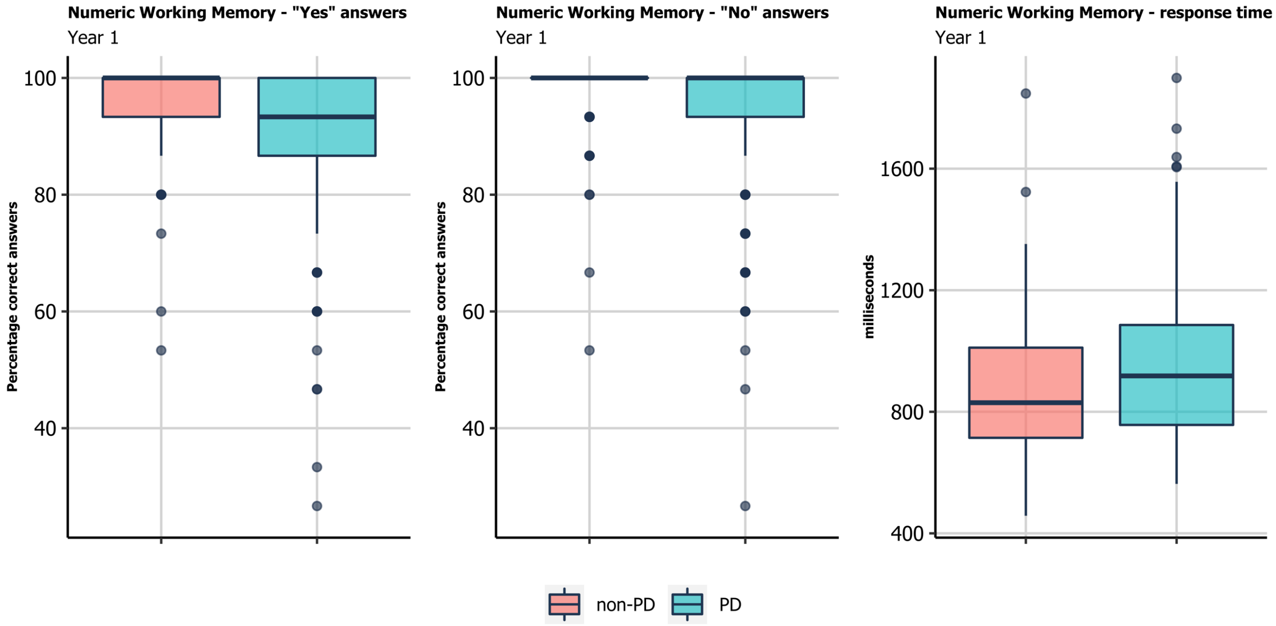


Figure 29: numeric working memory, PD patients vs non-affected carriers. The middle bar shows the median, the hinges the 25th and 75th percentiles and the whiskers include all values that are within 1.5 inter-quartile range from the hinges. Dots represent observation outside the 1.5 inter-quartile range from the hinges.

2.3.9. Estimating a risk score in GBA variants carriers

To identify those participants at a higher risk of PD among GBA variants carriers, a risk score was calculated. This is just an example of how the data could be used to predict the risk of PD, as validation will require longitudinal assessment. Points were assigned to each risk factor arbitrarily, according to the evidence in the literature supporting a link to the development of PD. Participants scored 1 point if they had a relative with PD, 2 points if they had hyposmia (or worse) at UPSIT, 2.5 points if they had RBD according to the RBDsq (with a cut-off of 5), 1 point if their KS30 scores and AT30 scores were in the lower 20% of the composite group including GBA-negative controls and non-affected GBA carriers, and 0.5 point if they answered "yes" to all 3 questions about constipation. Moreover, 1 or 0.5 points were assigned for each of the 6 cognitive tests where participants scored in the lower 20% of the composite group including GBA-negative controls and non-affected GBA carriers. The resulting score can vary from 0 to 11.5, with higher values representing a higher risk of PD. This risk score was analysed with OLR to evaluate whether higher values were associated with older age and no correlation was detected in GBA-negative controls and non-affected GBA carriers (p-value 0.26 - Figure 30).

	Points
RBD	2.5
hyposmia or worse	2
tap-test worse 20%	1
Picture recognition test worse 20%	1
Spatial working memory test worse 20%	1
Family history of PD	1
Depression	1
Simple reaction time test worse 20%	0.5
Digit vigilance test worse 20%	0.5
Numeric working memory test worse 20%	0.5
Constipation	0.5
Choice reaction time test worse 20%	0.5

Table 20: Points assigned to each risk factor to calculate a risk score of PD.

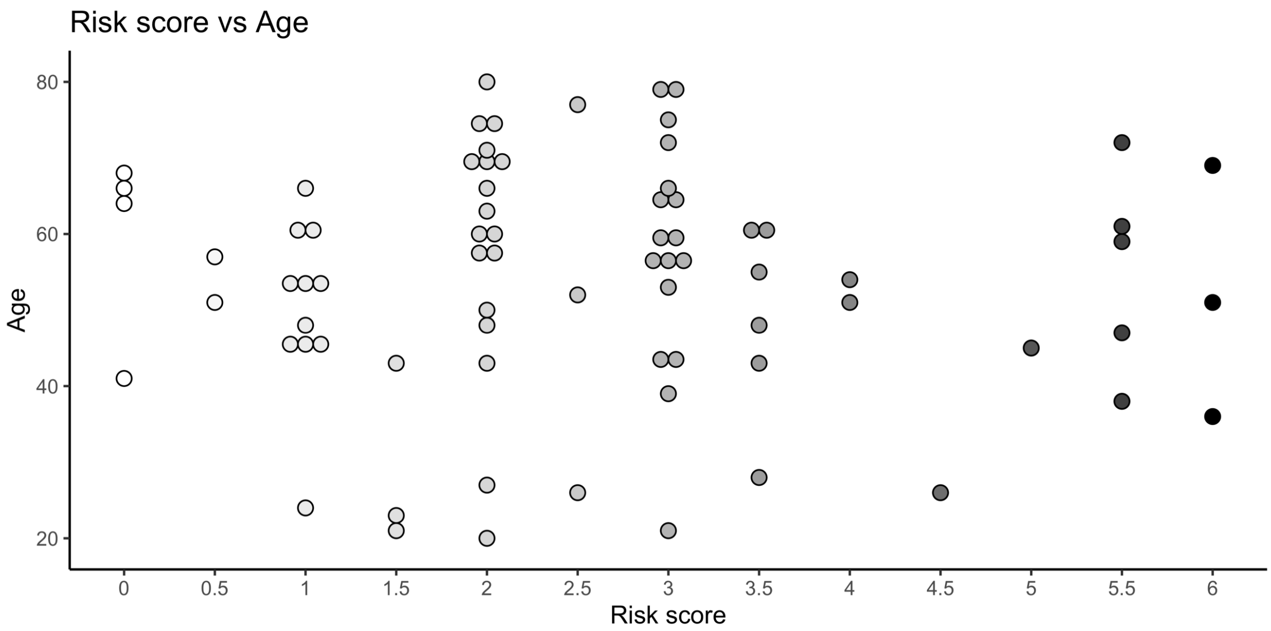
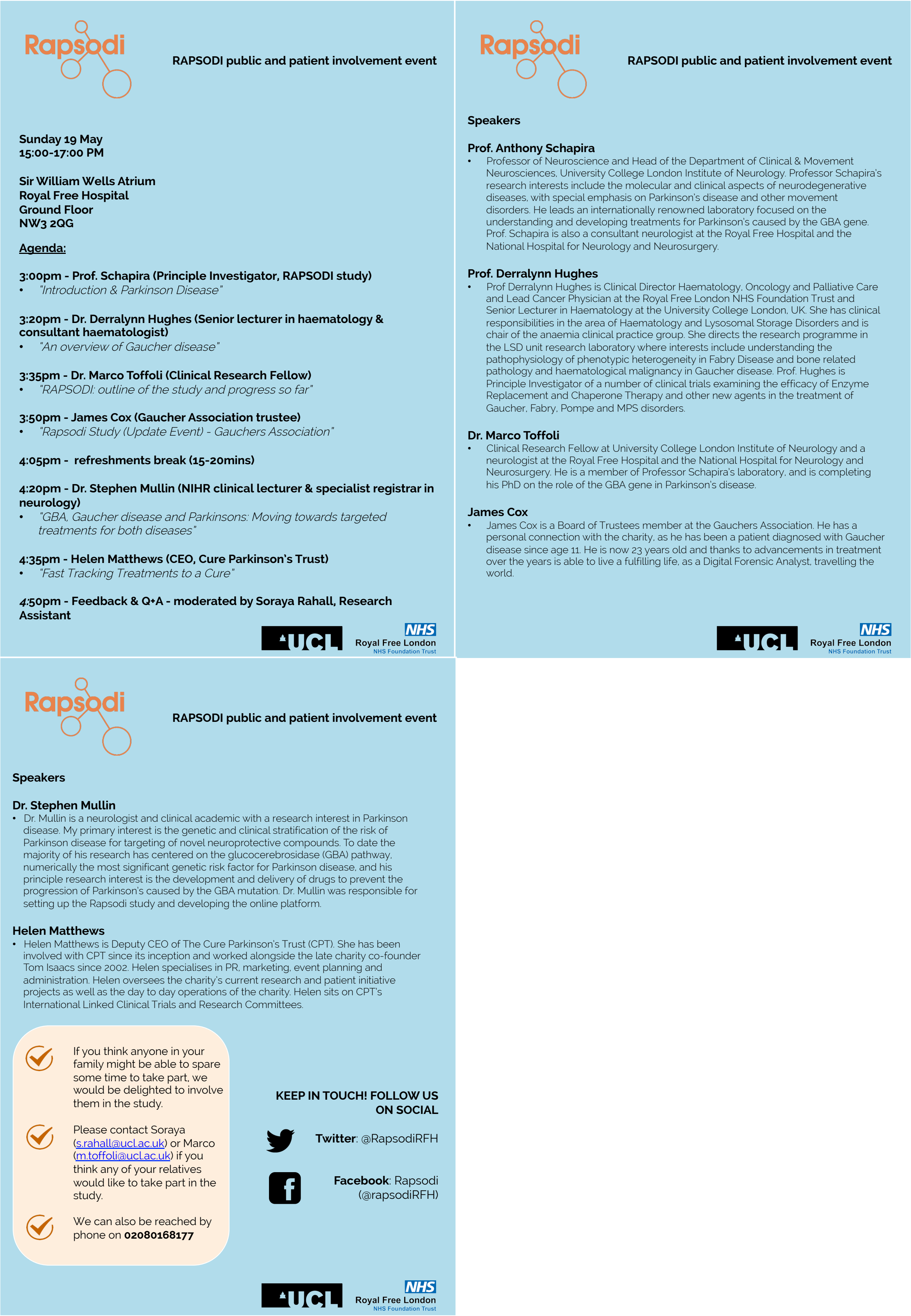


Figure 30: Risk score vs age in GBA-negative controls and non-affected GBA carriers.

2.3.10. Public engagement

We sent out an email bulletin to all RAPSODI participants every 3 months. This included relevant research findings and publications based on data collected through the portal. The study was advertised on the UK Gaucher Association website (www.gaucher.org.uk) and on their email bulletin and on the Parkinson's UK website (www.parkinsons.org.uk). A live event with RAPSODI participants was organised on the 29th May 2019. The flyer to advertise the event, including the agenda and list of speakers, is reported in Figure 31. Over 70 participants attended to the event and were able to ask questions to the speakers. A second edition of the live event was due to take place in 2020, but due to the new challenges posed by the COVID-19 pandemic, a webinar was organised instead. This took place on the 18th June 2020 and was attended by 31 participants. The agenda of the webinar is reported in Table 21. The webinar was attended by 31 participants. Videos of both the 2019 event and 2020 webinar were published on the Youtube page of the study (https://www.youtube.com/channel/UCdsOHwDsn_yaunfAxAnkM7A) to increase exposure and collected a total of 914 views.

These forms of participants and public engagement triggered a successful response from family members of attendees, ultimately boosting the recruitment of new participants and the compliance with the yearly follow-up assessment.
Figure 32 shows how some significant events affected recruitment.



RAPSODI public and patient involvement event

Sunday 19 May
15:00-17:00 PM

Sir William Wells Atrium
Royal Free Hospital
Ground Floor
NW3 2QG

Agenda:

3:00pm - Prof. Schapira (Principle Investigator, RAPSODI study)

- *"Introduction & Parkinson Disease"*

3:20pm - Dr. Derralynn Hughes (Senior lecturer in haematology & consultant haematologist)

- *"An overview of Gaucher disease"*

3:35pm - Dr. Marco Toffoli (Clinical Research Fellow)

- *"RAPSODI: outline of the study and progress so far"*

3:50pm - James Cox (Gaucher Association trustee)

- *"Rapsodi Study (Update Event) - Gauchers Association"*

4:05pm - refreshments break (15-20mins)

4:20pm - Dr. Stephen Mullin (NIHR clinical lecturer & specialist registrar in neurology)

- *"GBA, Gaucher disease and Parkinsons: Moving towards targeted treatments for both diseases"*

4:35pm - Helen Matthews (CEO, Cure Parkinson's Trust)

- *"Fast Tracking Treatments to a Cure"*

4:50pm - Feedback & Q+A - moderated by Soraya Rahall, Research Assistant

RAPSODI public and patient involvement event

Speakers

Prof. Anthony Schapira

- Professor of Neuroscience and Head of the Department of Clinical & Movement Neurosciences, University College London Institute of Neurology. Professor Schapira's research interests include the molecular and clinical aspects of neurodegenerative diseases, with special emphasis on Parkinson's disease and other movement disorders. He leads an internationally renowned laboratory focused on the understanding and developing treatments for Parkinson's caused by the GBA gene. Prof. Schapira is also a consultant neurologist at the Royal Free Hospital and the National Hospital for Neurology and Neurosurgery.

Prof. Derralynn Hughes

- Prof Derralynn Hughes is Clinical Director Haematology, Oncology and Palliative Care and Lead Cancer Physician at the Royal Free London NHS Foundation Trust and Senior Lecturer in Haematology at the University College London, UK. She has clinical responsibilities in the area of Haematology and Lysosomal Storage Disorders and is chair of the anaemia clinical practice group. She directs the research programme in the LSD unit research laboratory where interests include understanding the pathophysiology of phenotypic heterogeneity in Fabry Disease and bone related pathology and haematological malignancy in Gaucher disease. Prof. Hughes is Principle Investigator of a number of clinical trials examining the efficacy of Enzyme Replacement and Chaperone Therapy and other new agents in the treatment of Gaucher, Fabry, Pompe and MPS disorders.

Dr. Marco Toffoli

- Clinical Research Fellow at University College London Institute of Neurology and a neurologist at the Royal Free Hospital and the National Hospital for Neurology and Neurosurgery. He is a member of Professor Schapira's laboratory, and is completing his PhD on the role of the GBA gene in Parkinson's disease.

James Cox

- James Cox is a Board of Trustees member at the Gauchers Association. He has a personal connection with the charity, as he has been a patient diagnosed with Gaucher disease since age 11. He is now 23 years old and thanks to advancements in treatment over the years is able to live a fulfilling life, as a Digital Forensic Analyst, travelling the world.

RAPSODI public and patient involvement event

Speakers

Dr. Stephen Mullin

- Dr. Mullin is a neurologist and clinical academic with a research interest in Parkinson disease. My primary interest is the genetic and clinical stratification of the risk of Parkinson disease for targeting of novel neuroprotective compounds. To date the majority of his research has centered on the glucocerebrosidase (GBA) pathway, numerically the most significant genetic risk factor for Parkinson disease, and his principle research interest is the development and delivery of drugs to prevent the progression of Parkinson's caused by the GBA mutation. Dr. Mullin was responsible for setting up the Rapsodi study and developing the online platform.

Helen Matthews

- Helen Matthews is Deputy CEO of The Cure Parkinson's Trust (CPT). She has been involved with CPT since its inception and worked alongside the late charity co-founder Tom Isaacs since 2002. Helen specialises in PR, marketing, event planning and administration. Helen oversees the charity's current research and patient initiative projects as well as the day to day operations of the charity. Helen sits on CPT's International Linked Clinical Trials and Research Committees.

✓ If you think anyone in your family might be able to spare some time to take part, we would be delighted to involve them in the study.

✓ Please contact Soraya (s.rahall@ucl.ac.uk) or Marco (m.toffoli@ucl.ac.uk) if you think any of your relatives would like to take part in the study.

✓ We can also be reached by phone on **02080168177**

KEEP IN TOUCH! FOLLOW US ON SOCIAL

Twitter: @RapsodiRFH

Facebook: Rapsodi (@rapsodiRFH)

Figure 31: Flyer of the 19th May 2019 RAPSODI live event.

Speaker	Topic
Marco Toffoli	This introductory statement will outline the theme of the webinar, “the link between Gaucher Disease and Parkinson Disease”. Dr Toffoli will inform the audience that they will hear from clinicians and researchers over the course of the hour, with an opportunity for Q&A at the end. The Q&A function for the platform will be explained.
Professor Derralynn Hughes	Professor Hughes will provide an update on the latest therapeutic strategies for managing Gaucher Disease.
Professor Anthony Schapira	The link between Parkinson disease and *GBA* variants
Abigail Higgins	This talk will focus on the research being undertaken by the Rapsodi study at UCL. The link between *GBA* mutations and Parkinson disease will be outlined, as to provide the rationale for Rapsodi. The aims of the study, to work with individual’s with *GBA* mutations to develop preventative strategies for Parkinson disease, will be discussed.
Chiao Lee	Chiao will discuss how both individuals affected by Gaucher Disease, and healthy carriers of *GBA* mutations share a slight increased risk for Parkinson Disease, and why this means that family participation in Rapsodi is so important. Chiao will explain the effects of *GBA* mutations on the activity of the enzyme glucocerebrosidase, that the *GBA* gene encodes for.

Table 21: agenda of the 18th June 2020 RAPSODI webinar.

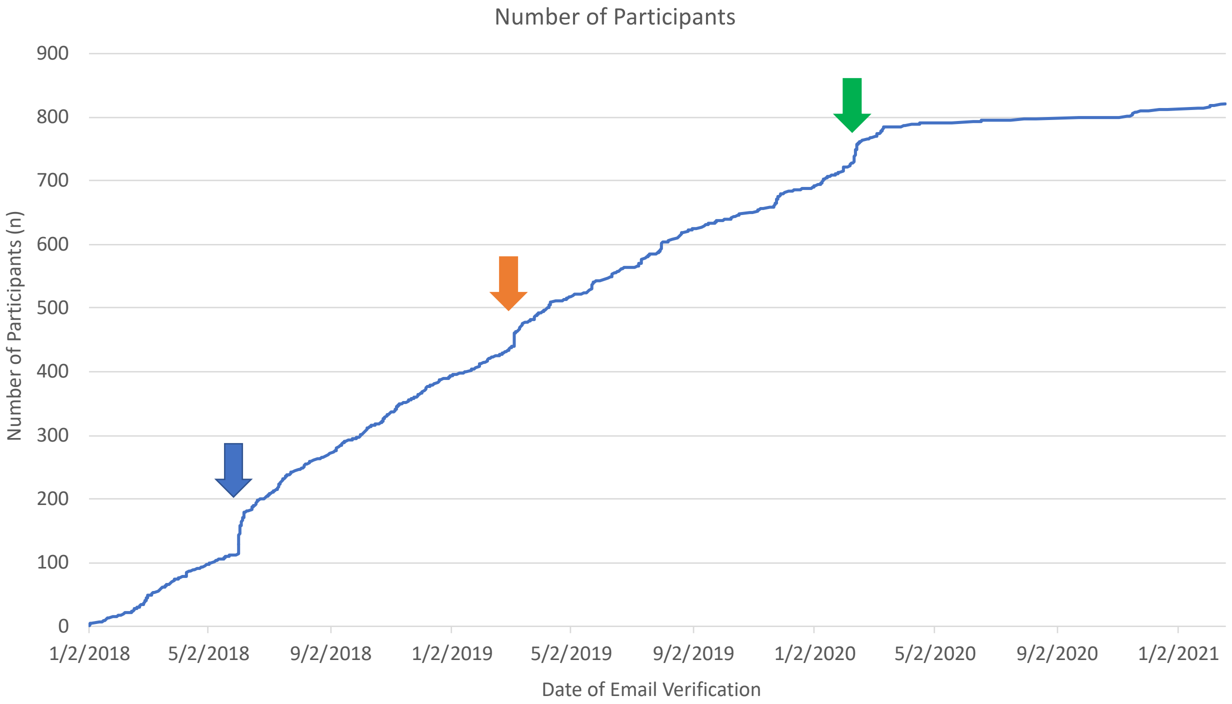


Figure 32: Participants recruited in the RAPSODI study over time. The blue arrow shows the date when the study was advertised on the Parkinson’s UK and Cure Parkinson’s trust. The Orange marks shows the date of the RAPSODI live

event in May 2019. The green arrow corresponds to the release of a Gaucher disease UK association bulletin advertising RAPSODI as well as a newsletter sent to all RAPSODI participants requesting to bring family members into the study.

2.4. Discussion

GBA variants are a key risk factor for PD, and carriers are the ideal candidates for disease modifying therapies. However, it is challenging to estimate the exact risk of PD in *GBA* carriers.

Between 2010 and 2016, we carried out a longitudinal cohort study on *GBA* carriers, to understand whether it was possible to predict who, among them, would develop PD [140–142]. This study included periodic face to face assessments and provided important insight, but also showed the limitations of the approach. In particular, it was hard to retain participants over the very long observation time required for carriers to convert to PD, as well as reaching a large enough number of participants for sufficient statistical power. The RAPSODI study was created to overcome these issues. In RAPSODI, recruitment and assessment of individuals are carried out remotely, through the online portal, allowing to reach a larger audience. Moreover, participants do not have to come to hospitals for face to face assessments and this increases compliance with follow-ups. In the very first part of my PhD, I was able to help setup the RAPSODI study, start recruiting participants and analyse some preliminary baseline data. The experience gathered so far will be important for future development of RAPSODI and other similar studies.

The design process and first years of life of the portal highlighted some important organisational aspects. First, selecting appropriate recruitment strategies is paramount. Meeting patients in clinics, asking them to enrol into RAPSODI, and then reaching out to their family members remains a viable path, but it does not exploit the full potential of the online design of the study. Indeed, alternative ways of recruiting participants, including advertising the study on patients' groups online and being active on social media boosted recruitment significantly. For example, we noticed a surge of recruitment of PD participants after advertising the study on the Parkinson's UK (https://www.parkinsons.org.uk/) and Cure Parkinson's (https://cureparkinsons.org.uk/) websites in June 2018 (Figure 32). Organising interactive events where participants and people interested in the study can ask questions has also proven beneficial for recruitment and compliance with follow-up. This is particularly true for *GBA* carriers, as they are identified mainly as family members of people with GD, a very rare condition. When participants attended to the 2019 RAPSODI event, they brought their family members with them, which were then offered to take part in the study, leading to another significant rise in number of participants right after the event (Figure 32). Interestingly, a third boost in recruitment was observed in the week following advertisement of the study on the Gaucher UK association bulletin and a newsletter to all RAPSODI participants with a report of recent publications related to RAPSODI (Figure 32).

It is paramount that participants carry out the assessment every year. To this end, reminder emails and periodic bulletins proved invaluable. We had a direct proof of the importance of this aspect in 2019, when we identified a bug in the system preventing the firing of most emails reminding participants to complete the second year of the study. As a result, a significant number of participants did not complete the year 2 assessment. This was a complex issue and the RAPSODI web development team managed to completely solve it only in late 2020. After the reminder emails system started to function correctly, we noted that most participants completed the yearly follow-up promptly. In 2019, we also started to send out periodic email bulletins to all RAPSODI participants, outlining the progress and achievements of the study. This also showed a beneficial effect on retention of participants at follow-up.

Assessing participants remotely has limitations. Most notably, all tests are unsupervised and there is no way to make sure that participants are not "cheating". In the tap test, they might use both hands at the same time, or they might seek external help to perform the cognitive tasks and the UPSIT. Speed of internet connection, potential crashes of the web browser and external distractions are also a factor, most notably in the cognitive tests where speed of response is the main outcome. To partially overcome these limitations, we removed extreme outliers from the tap test and used median response time (less influenced by occasional long response times) as outcome in the cognitive tests.

Relying on a web portal to assess participants also produces an intrinsic selection bias, as participants that do not know how to use a computer or do not own one are automatically excluded from the study. However, the main goal of RAPSODI is to assess people that carry *GBA* variants but do not have PD yet (i.e. are not very old) and one of the exclusion criteria is the absence of dementia. For these reasons, I believe that informatic literacy did not bias recruitment significantly. Access to a computer on the other hand was a limitation, as many participants only owned a tablet, not compatible with the tap test.

All in all, we believe that the advantages of remote assessment in the RAPSODI study outweigh limitations.

The main outcome of RAPSODI is to identify *GBA* carriers that develop PD and then look at their baseline assessment to identify features that would help predict this conversion. Meeting this goal will require a long follow-up period, not available at the time of writing this thesis.

However, some important observations can be made from the preliminary data gathered at the baseline assessment of participants recruited up to the 1st March 2021.

First, we were able to differentiate reliably between participants with and without PD. This is not surprising, as PD patients are expected to have worse motor performance as well as a higher prevalence of non-motor symptoms of PD. Nonetheless, this proves that the assessment tools we applied are appropriate for studying people with clinically evident PD. We know that non-motor symptoms can sometimes predate the occurrence of motor symptoms, a condition sometimes referred to as "prodromal PD". As our assessment showed differences in PD patients compared to controls, we can hope it has the potential to also highlight features of prodromal PD in *GBA* carriers.

The tap test was already validated in people with PD[134], and in this work we provided further evidence of its validity. We stratified tap test data into dominant and non-dominant hand, but noticed that for people with PD it might be more appropriate to use most and least affected hand instead. Unfortunately, we did not initially capture which side of the body is most affected in PD patients, but added this question to the questionnaire so that it will be possible to do so in future years. The CogTrack cognitive assessment tool was already used successfully on a number of conditions[143,144], but not in people with PD. Our data show that CogTrack can also reliably detect cognitive deficits in people with PD, in particular with regards to working memory and picture recognition. It is well known that working memory and executive functions are the most affected domains in PD[145] and in this work we were able to provide evidence in support of the use of CogTrack in people with PD. Interestingly, our preliminary analysis also showed that PD patients carrying *GBA* variants had worse non-motor symptoms when compared to *GBA*-negative PD patients. Indeed, carriers showed a higher prevalence of RBD and constipation and performed worse at the UPSIT, picture recognition and choice reaction time tests. *GBA*-associated PD has already been showed to have more prominent non-motor features compared to sporadic PD[146], and our findings support and provide more evidence to this. Of particular interest is the worse performance of *GBA*-positive PD patients in the picture recognition test. This reinforces the concept that *GBA* associated PD has a distinctive cognitive pattern, with the posterior cortex more affected than in sporadic PD[147,148].

Important limitations of the statistical analysis are the poor matching of age across groups and the lack of adjustment for PD duration (due to collinearity with age).

These limitations were mitigated as age was accounted for in the multivariate analysis, and PD duration was not statistically different between the two PD groups. As this was an exploratory analysis, no adjustment for multiple comparisons was carried out to account for the many tests carried out, increasing the risk of type I error.

These findings are promising and suggest that expansion and longer observation of the cohort will provide further insight into the role of *GBA* variants in PD. Moreover, perhaps the most important achievement of the first years of RAPSODI was the collection of a large number of DNA samples from carriers of *GBA* variants, which allowed the development of the sequencing technique discussed in the following chapters.

3. Optimisation of a pipeline for sequencing the *GBA* gene with Oxford Nanopore Technology

Content of this chapter has been adapted from a recent publication I was co-author of[149], with permission of the publisher. Additional data contained in this chapter will be reported in a separate publication, which is currently under submission.

3.1. Overview and Rationale

Sequencing of the *GBA* gene is challenging. Sequencing of all 11 exons with Sanger reactions is expensive and time consuming, while “SNV specific approaches” only allow the detection of the most common GBA SNVs, which is limiting considering that there are at least 495 disease causing variants in *GBA*[107,150]. Even short read sequencing of WGS data is limited by the presence of the highly homologous *GBAP1*, resulting in a high number of false positives and negatives[110,111].

Dr Christos Proukakis previously developed a long-read approach for sequencing *GBA* using Oxford Nanopore Technologies (ONT) and successfully sequenced 9 DNA samples. In this section of my PhD, I improved the pipeline to process a higher number of samples and optimised it for DNA extracted from saliva. The initial 95 samples that I sequenced were included in the methodology manuscript published in 2019[149].

3.1.1. Oxford Nanopore technology, an overview

The core sequencing component of ONT is a membrane with hundreds of embedded nanopores. A complex system of proteins allow the concentration of the double stranded DNA molecules near the membrane, separation of the dsDNA into single strands and subsequent passage of these strands through the pores, from the 5’ to the 3’ extremities. As they pass through the pores, each group of 5 nucleotides

(5-mer) generates a current that is specific for that molecule and can translated into the real sequence of nucleotides with downstream *in silico* analysis[151–153]. A schematic representation of the funcitoning of Oxford Nanopore Technology is provided in *Figure* 33.

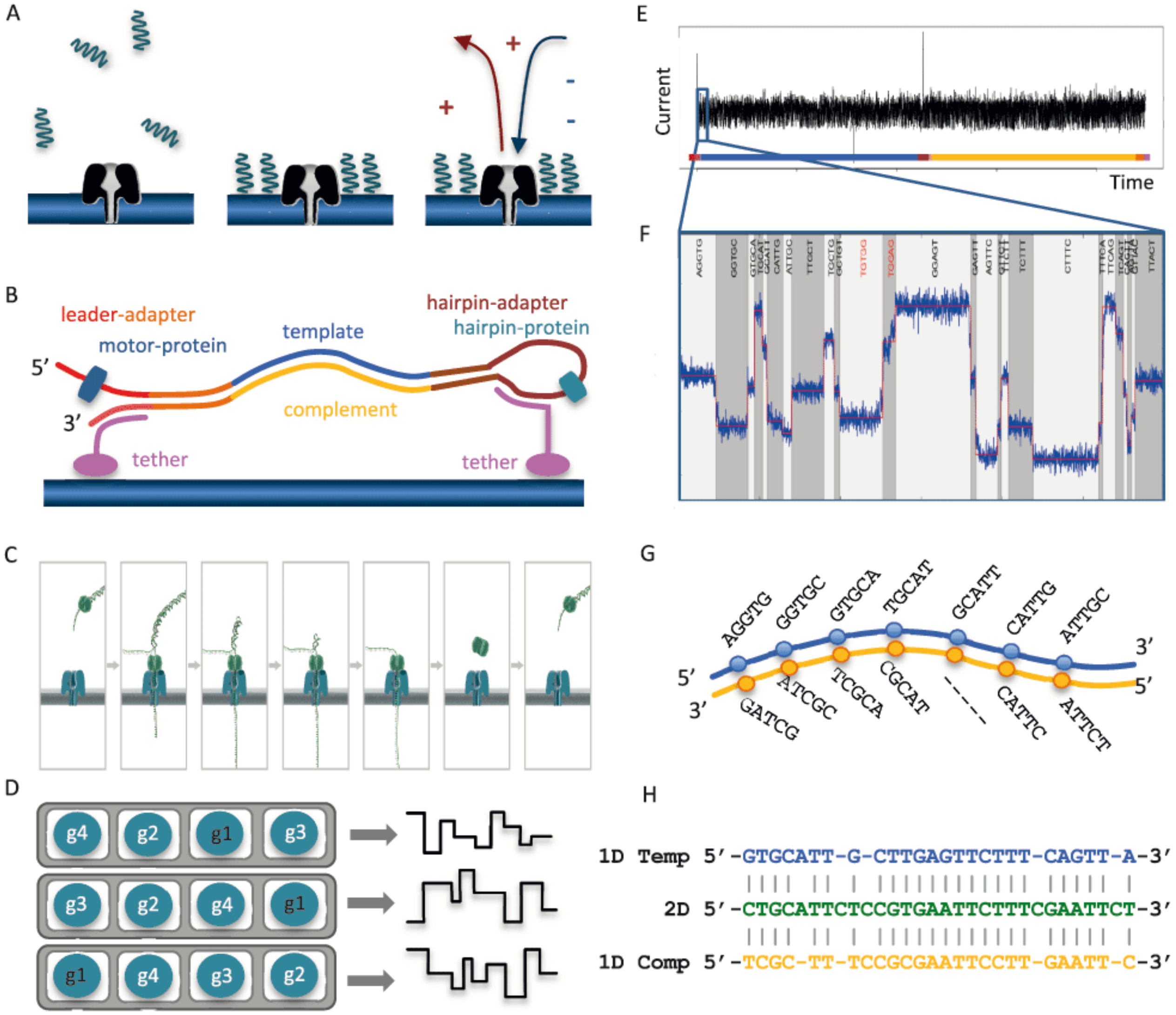


Figure 33: schematic representation of functioning of Oxford Nanopore Technology.

(A) DNA molegules are directed through nanopores through a current. (B) Specific proteins prepare the DNA for being sequenced through the nanopores. (C) DNA molecule passes through the nanopores. (D) A flow cell is composed of multiple nanopores. Each of them can sequence one DNA molecule at a time, in parallel with the other pores. (E) The current generated by the DNA passing through the pores is measured. (F) The data is analysed and 5-mers are deducted with a statistical model. (G) Template and complement molecules are inferred separately. (H) Data is aligned to the reference genome.

Image reproduced from Ip CLC et all[151], under Creative Commons Attribution License.

3.2.Methods

3.2.1. Extraction of DNA from Saliva samples

We collected saliva samples from all RAPSODI participants (see chapter 2) using the Oragene DNA OG-500 kit from DNA genotek (https://dnagenotek.com). Upon receiving the samples, we extracted DNA following the protocol provided by the kit manufacter. Briefly, this includes incubation of the saliva at 50°C overnight, then lysis of cellular and nuclear membranes with a proprietary reagent (PT-L2P), washes with 100% and 70% ethanol and resuspension in TE buffer (10mM Tris-HCL, 1 mM EDTA, pH 8.0). I introduced an additional step: samples are left with lid open for 15 minutes after the second ethanol wash, to allow evaporation of residual ethanol. This is to avoid carryover of ethanol, which might cause failure of downstream PCRs. The extracted DNA was divided into 3 aliquots, one stored at +4°C and the remainings stored at -20°C.

QC of the extracted DNA was carried out with spectrophotometric assay (Nanodrop). Sample were considered of acceptable quality if both the 260/230 and 260/280 values were above 1.7.

3.2.2. PCR amplification of the *GBA* gene

The lab protocol to amplify the *GBA* gene was designed prior to my PhD, using PCR with primers designed previously[154]. These were modified to carry the ONT barcode adapters. The product is an amplicon of 8.9 Kb covering the 11 exons encoding for the mature GCase protein and the 3’ and 5’UTR regions (chr1:155232524-155241392), but not the exon coding for the 39-aa leader peptide[154].

Optimisation of the PCR conditions was required for amplifying samples derived from saliva, often showing inferior quality compared to DNA extracted from blood. For this reason, I tested different conditions, altering the annealing temperature and the concentration of primers, template DNA and magnesium. I also tested a different enzyme from that originally used in the lab, the LongAmp® Taq 2X Master Mix (New England BioLabs, catalog number M0287L). Quality of the PCR product was assessed with electrophoresis on a 0.8% agarose gel and concentration determined with fluorometry (QUBIT, Broad Range kit; ThermoFisher scientific, catalog number Q32850). Throughout the text, this step is referred to as “amplifying PCR”. More details on the parameters and optimisation of this step in the results section and in Table 23.

3.2.3. Barcoding of DNA samples

A second PCR reaction was carried out to attach the barcodes to the amplicons(https://nanoporetech.com/community). PCR Barcoding Expansion Kit 1 (up to 12 samples, catalog number SQK-PBK004) and 96 (up to 96 samples, catalog number PBC096) were used, following the protocol provided with the kit. Throughout the text, this step is referred to as “barcoding PCR”. More details on the parameters and optimisation of this step in the results section.

3.2.4. Purification of PCR products

Purification of the PCR product after each PCR step was carried out with spin columns (QIAquick PCR purification kit - QIAGEN)[155] or with magnetic beads (Ampure XP – Beckman Coulter)[156]. With spin columns, the PCR product solution is added to the column, where the DNA binds to the silica membrane in the presence of the binding buffer (a high-salt solution). The column is then washed and finally, the DNA is eluted in water and collected in a LoBind tube. To purify with beads, the PCR product solution is mixed with the magnetic beads solution and the DNA binds to the beads, which are then pelleted on a magnet. The pellet is washed two times with 85% ethanol and the DNA is then eluted in water. The beads are pelleted again on the magnet and the eluate is stored in a new LoBind tube.

More details on the optimisation of this step are provided in the results section.

3.2.5. Library preparation and sequencing with MinION

Library preparation was carried out according to the ONT protocol "PCR barcoding (96) amplicons (SQK-LSK109)" found at https://community.nanoporetech.com/protocols. The library preparation kit "SQK-LSK109" was used. Briefly, barcoded amplicons from all samples were pooled together. Subsequently, a DNA repair step was carried out with NEBNext FFPE DNA Repair Mix (M6630 – New England BioLabs) and NEBNext Ultra II End repair / dA-tailing Module (E7546 – New England BioLabs), then the product was purified with magnetic beads (Agencourt AMPure XP beads – Beckman Coulter). The ONT sequencing adapters were ligated with NEBNext Quick Ligation Module (E6056 – New England BioLabs) and the product was purified with AMPure XP beads at the concentration suggested in the manufacturer's protocol. The Long Fragment Buffer (LFB, part of the SQK-LSK109 kit) was used for additional size selection of DNA fragments longer than 3 Kb.

Finally, the prepared library was loaded into 9.4 flow-cells following the manufacturer's protocol.

3.2.6. Bioinformatics

All bioinformatics analysis were carried out on a unix machine running Ubuntu 18 (16GB RAM, 2Tb Ssd, Pentium quad core, 8 threads), with the exception of basecalling with Guppy, which was run on the UCL computing platform "Myriad". The most updated version of each software was used at all time. The code described in the subsections below is reported in Appendix – Code.

3.2.6.1. Sequencing

Raw sequencing data (fast5 files) were acquired on a MinION device using MinKNOW software (ONT, versions 20.10.3), downloaded from https://community.nanoporetech.com/downloads. The output is in .fast5 format.

3.2.6.2. Basecalling and demultiplexing

Basecalling of the fast5 files was carried out with Albacore[149] or with Guppy (version 4.2.2), downloaded from https://community.nanoporetech.com/downloads (ONT, Oxford, UK). Guppy uses local acceleration and requires a graphics processing unit to function. For this reason, Basecalling was carried out on one of UCL cloud computing platform Myriad (2 GPU, 12 threads, 5Gb RAM per CPU). The output of barcoding and demultiplexing is in .fastq format, with .fastq files organised in subfolders, one subfolder for each barcode.

3.2.6.3. Alignment

Alignment was carried out with NGMLR (version 0.2.7)[157], downloaded from https://github.com/philres/ngmlr[157]. The output was in .bam format.

3.2.6.4. Variants-calling

Nanopolish and Clair (version 2.1.1)[158] were evaluated for calling of SNVs. The output was in .vcf format.

Clair is based on machine learning and uses pre-trained models that are available at (https://github.com/HKU-BAL/Clair#pretrained-models)[158]. The most updated pre-trained modules are trained with a set depth of coverage, and might not be reliable with higher depth of coverage. For this reason, before running Clair, all .bam files were “downsampled”, which means that a set number of reads were randomly selected for each sample to achieve the desired depth of coverage using samtools view.

3.2.6.5. Differentiating true and false positive calls

The main issue with variants calling was the high number of false positive calls, expecially single nucleotide insertions and deletions. This made interpretation of the results confusing when the false positives were in coding regions of *GBA.* Moreover, this interfered with the phasing process described later. To address this problem, I tested different approaches, as detailed in the next paragraphs. To identify the superior method, precision (positive predicting value) and recall (sensitivity) were estimated. True and false positive calls were defined by visual assessment on IGV (section 3.2.6.5.1). False negatives were identified when a true positive variant that was called by one method was missed with another. Only variants in exons, inlcuding 10 flanking bases, were considered for this analysis.

3.2.6.5.1. Visual evaluation of all variants called

A first approach consisted in the manual inspection of all the variants called, to decide whether they were true or false positives. The software “Integrative Genomics Viewer” (IGV, version 2.8.9 http://software.broadinstitute.org/software/igv/) was used for this purpose. Some false positives were easily detectable as they were present in all samples with a frequency of the alternative allele/reference allele of roughly 25/75%. In other instances however it was hard to determine whether calls were true or false. More details on this are given in the results section.

3.2.6.5.2. Nanopolish quality score

Nanopolish provides a quality score (QS) for each variant called, which is the log-likelihood ratio was calculated by the software's probabilistic model[159]. Nanopolish QS is directly proportional to the number of reads covering each variant, meaning that it is not possible to set a threshold that can be applied to all variants. To address this issue, I normalised it for the number of reads per position. I called the resulting metric the "adjusted QS", which is calculated by dividing the Nanopolish QS by the total number of reads covering that position. As discussed in the results section, the adjusted QS is independent of the coverage and thus a threshold can be set that is valid for all variants and all samples.

3.2.6.5.3. Strand filtering

Reads produced by ONT can be grouped according to the DNA strand of origin, conventionally referred to as the positive and negative strands. Since the two DNA strands are complementary, true positive SNVs are equally represented in both the positive and negative strands. On the other hand, some false positive calls show a propriety called strand bias, meaning that the alternative allele is over represented in one strand compared to the other.

To filter for strands, first I divided the aligned reads (.bam files) into positive and negative strands using *samtools view*, then the variant caller is run on both strands. Only calls that are present in the positive strand, negative strand and total reads were considered true positive.

3.2.6.5.4. Clair quality score

Clair provides a QS for each SNV called. I downsampled the .bam files before running Clair, as the trained models for Clair are based on a coverage of 550.

I was then able to select a threshold that was valid across different MinION runs and for different SNVs and samples. Different QS thresholds were tested.

3.2.6.6. Phasing

Whatshap[160] was used to phase variants. Only base substitutions were phased, while insertions and duplications were not. The software was downloaded from https://whatshap.readthedocs.io/en. The output was in .vcf format.

3.2.6.7. Creation of scripts

Bash scripts were created using Emacs (version 27.1) and Python scripts in visual studio code (version 1.53.1).

3.2.6.8. Files manipulation and Miscellaneous

For manipulation of files the following software was used: Samtools (http://www.htslib.org), BCFtools (http://www.htslib.org), bedtools (https://bedtools.readthedocs.io/en/latest/). Downsampling of .bam files was achieved with the Samtools view command.
IGV (http://software.broadinstitute.org/software/igv/) was used for visually inspect the alignment files. f

3.2.7. Confirmation of positive and negative results

Positive results were confirmed by Sanger sequencing of the exon of interest, carried out at the Royal Devon and Exeter Hospital.
To confirm negative results (i.e. samples where no variant was detected), I carried out a PCR reaction to amplify a region of DNA containing exons 8-11 of GBA with primers and PCR conditions obtained from a previous publication[106]. Subsequently,

2 Sanger reactions, forward and reverse, were carried out to sequence exons 9, 10 and 11, where the vast majority of *GBA* coding SNVs are located. This was performed through a commercial provider of Sanger sequencing services (https://www.sourcebioscience.com).

3.3. Results

3.3.1. Samples analysed

A total of 381 samples were successfully sequenced in 7 MinION runs. Of them, 269 did not carry any coding variant in *GBA*, 78 were heterozygous and 34 were homozygous or compound heterozygous carriers of *GBA* variants. The most common variant was p.N409S (60 alleles, allele frequency - AF 7.9%), p.L483P (24 alleles, AF 3.2%) p.E365K (17 alleles, AF 2.2%), T408M (7 alleles, AF 0.9%). It is interesting to note that a number of intronic variants likely to affect splicing (IVS9+1, IVS6-2, IVS2-1) were detected in 5 participants. Two of these variants (IVS6-1 and IVS2-1) were never described before and are likely pathogenic. Moreover, 7 participants carried a complex, pathogenic reciprocal recombinant, as discussed in more details in chapter 4. A full list of all variants detected is reported in Table 22. The clinical phenotype of participants sequenced was reported in capter 2.

Variant	Genomic coordinates (GRCh38)	Number of alleles	Allele frequency
WT	NA	616	80.8%
N409S	Chr1: 155235843T>G	60	7.9%
L483P	Chr1: 155235252A>G	24	3.1%
E365K	Chr1: 155236376C>T	17	2.2%
Rec*	*	7	0.9%
T408M	Chr1: 155236246G>A	5	0.7%
R502C	Chr1: 155235196G>A	4	0.5%
IVS9+1	Chr1: 155235680C>T	3	0.4%
84GG	Chr1: 155240661dup	2	0.3%
R301H	Chr1: 155237439G>C	2	0.3%
V433L	Chr1: 155235772C>A	2	0.3%
IVS6-2	Chr1: 155238308T>C	2	0.3%
A357D	Chr1: 155236399G>T	1	0.1%
A495P Val499= L483P	Chr1: 155235217C>G + 155235203 C>G + 155235252A>G	1	0.1%
c413del	Chr1: 155239661del	1	0.1%
D354H	Chr1: 155236409C>G	1	0.1%
D419N	Chr1: 155235814C>T	1	0.1%
G241R	Chr1: 155238174C>T	1	0.1%
G289V	Chr1: 155237474C>A	1	0.1%
K13R	Chr1: 155240707T>C	1	0.1%
L144R	Chr1: 155239639A>C	1	0.1%
L519P	Chr1: 155235050A>G	1	0.1%
P211T	Chr1: 155238234G>T	1	0.1%
R209P	Chr1: 155238269C>G	1	0.1%
R301G	Chr1: 155237438C>T	1	0.1%
R398ter	Chr1: 155236277G>A	1	0.1%
T270I	Chr1: 155237531G>C	1	0.1%
T408M + T104=	Chr1: 155236246G>A + 155239758C>T	1	0.1%
T408M + W432ter	Chr1: 155236246G>A + 155235773C>T	1	0.1%
V486E	Chr1: 155235243A>T	1	0.1%

Table 22: All pathogenic GBA variants detected with Nanopore sequencing.

**Complex recombinants, containing the RecNciI variants A495P Val499= L483P, detected with additional analysis as described in chapter 4.*

3.3.2. Confirmation of positive and negative results

Sixtyfive positive samples were confirmed by Sanger sequencing of the exon where the coding SNV was detected. To confirm negative results, Sanger sequencing was carried out on 35 samples where no coding SNV was detected with ONT. All results were consistent with the ones detected with ONT.

3.3.3. Optimisation of the amplifying PCR

Initial PCR conditions for amplifying the *GBA* gene were reported in our 2019 paper[149]. These were based on the use of Kapa Hi-Fi Polymerase (Kapa Biosystems). However, when processing a higher number of samples, a relatively high number of them were failing or producing an insufficient amplification (Figure 34). Moreover, the Kapa enzyme is relatively expensive and has a limited shelf-life, so I decided to test the LongAmp® Taq 2X Master Mix (New England BioLabs). Starting PCR conditions were obtained from the manufacturer protocol, and initial annealing temperature for the primers was estimated to be 65°C with an online calculator (https://tmcalculator.neb.com). Additional annealing temperatures were tested and 65°C was confirmed to be the optimal condition (Figure 35). Next, different amounts of template DNA were tested (50ng, 100ng, 150ng and 200ng) and 200ng was found to be the optimal concentration. Optimal conditions of the amplifying PCR are reported in Table 23.

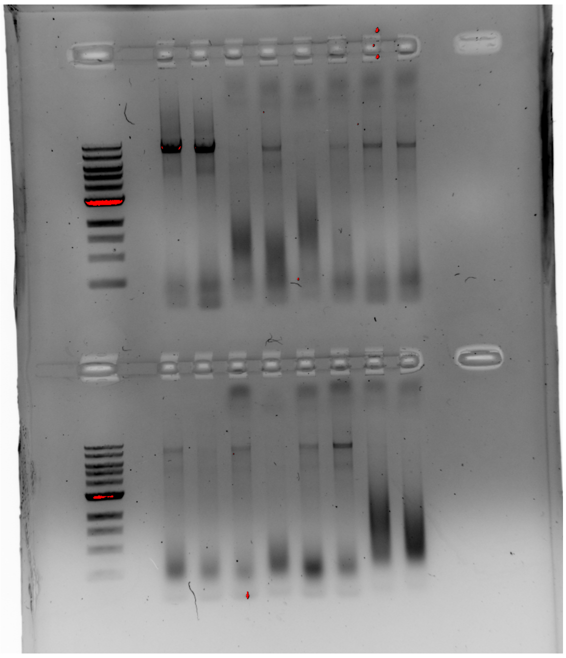


Figure 34: Example of an unsuccessful amplifying PCR. Most samples did not produce a clear band at the expected at the 8.9Kb mark, and an a-specific smear is visible.

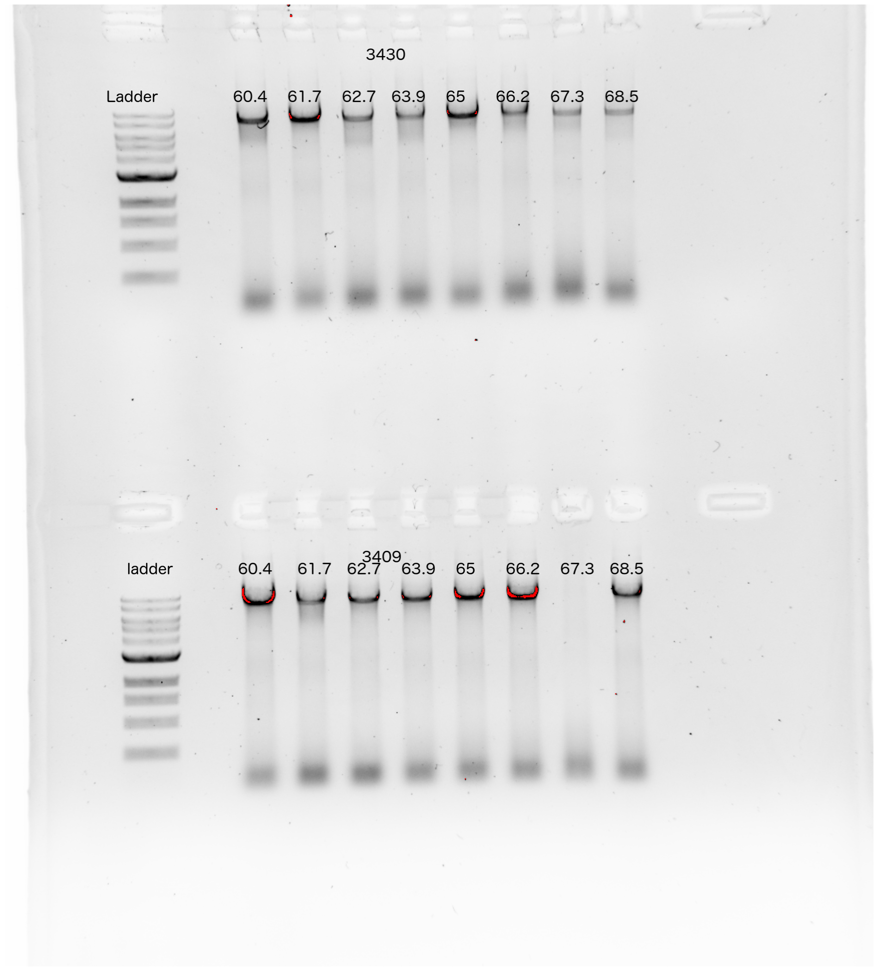


Figure 35: Testing different annealing temperatures for the amplifying PCR with LongAmp enzyme. Two samples were tested and 65°C was confirmed as the optimal condition.

Reagents				
Template DNA	200ng			
Nuclease free water	To 19μL			
Mg^{2+}	2μL			
Forward primer	2μL			
Reverse primer	2μL			
Polymerase mix	25μL			
Conditions	Temperature	time	cycles	
Initial denaturation	94	30 seconds	1	
denaturation	94	15 seconds	35	
annealing	65	15 seconds		
extension	65	6 minutes	1	
Final extension	65	10 minutes	1	
Primer pair A	Forward	5’- TTTCTGTTGGTGCTGATATTGCTCCTAAAGTTGTCACCCATACATG -3’		
	Reverse	5’- ACTTGCCTGTCGCTCTATCTTCCCAACCTTTCTTCCTTCTTCTCAA -3’		

Table 23: Optimal conditions for PCR to amplify the GBA gene.

**primers contain the ONT barcode adaptor sequence.*

3.3.4. Optimisation of the barcoding PCR

The manufacturer protocol provides details for the barcoding PCR. However, we wanted to optimise the reaction further, in order to maximise yield and quality of the product. The parameters that underwent optimisations are: the number of cycles (12 vs 15), the amount of template DNA (140ng vs 100ng), the volume of primers mix (1µL vs 0.5µL). The conditions tested are displayed in Figure 36 and the optimal PCR conditions are reported in Table 24.

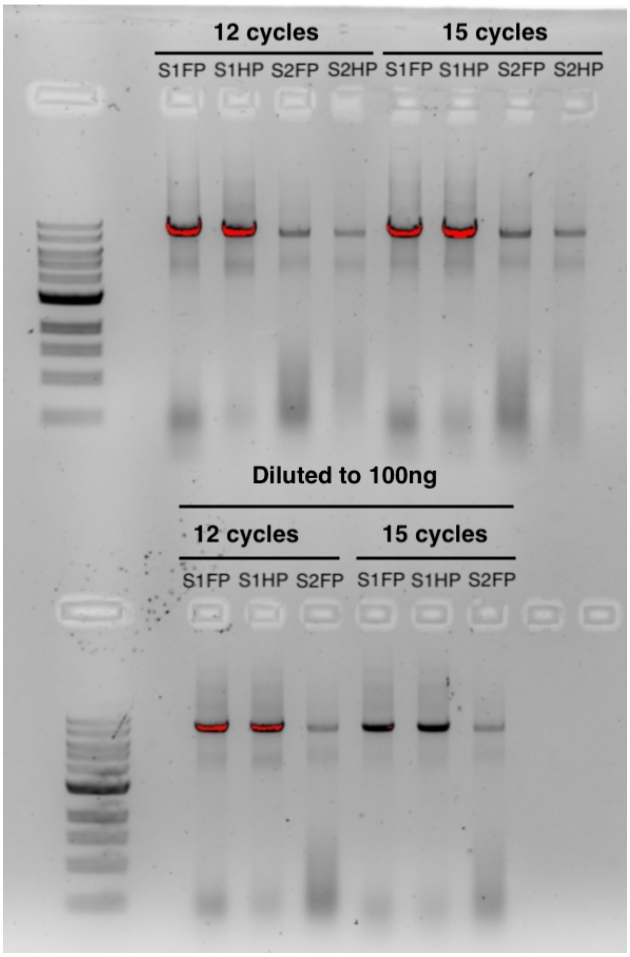

12 cycles
15 cycles
S1FP S1HP S2FP S2HP S1FP S1HP S2FP S2HP
Diluted to 100ng
12 cycles
15 cycles
S1FP S1HP S2FP S1FP S1HP S2FP

Figure 36: Barcoding PCR optimisation. Two samples are showed in this 0.8% agarose gel (S1 and S2). FP: 1µL primers mix, HP: 0.5µL primers mix. Top row shows experiment with 140ng of template DNA, bottom row 100ng of template DNA.

Reagents				
Template DNA	100ng			
Nuclease free water	To 24.5µL			
Primers mix	0.5µL			
Polymerase mix	25µL			
Conditions	Temperature	time	cycles	
Initial denaturation	95	3 minutes	1	
denaturation	95	15 seconds	12	
annealing	62	15 seconds	1	
extension	65	10 minutes	1	
Final extension	65	10 minutes	1	

Table 24: Optimal conditions for Barcoding PCR.

3.3.5. Purification of PCR product with magnetic beads produces better results

Initially, purification of products of both the amplifying PCR and the barcoding PCR were carried out with spin columns[155]. However, the barcoding PCR was failing in a relatively high number of samples. Looking at possible causes for this suboptimal performance of the PCR step, we noticed that most samples displayed an intense smear in the electrophoresis gel after the amplification PCR, straddling the 500 bp band on the ladder. We interpreted this band as low molecular weight (LMW) DNA molecules, as the molecular weight was too high for them to be primer dimers.
A high concentration of these LMW DNA fragments can interfere with the downstream PCR for different reasons. First, low fragments can compete with the longer *GBA* DNA molecules during the PCR reaction. Second, the DNA concentration measured by QUBIT fluorometry accounts for the entirety of the DNA molecules in the solution, including these DNA fragments. This means that the concentration of full *GBA* DNA amplicons is actually lower than that calculated with fluorometry concentration readings. To investigate whether this might have caused the failing of the downstream barcoding PCR, we analysed a batch of samples that succeeded and one of samples that failed amplification with PCR with automated electrophoresis (Tapestation, Agilent).

Results showed that samples that succeeded had a better genomic DNA quality and a better ratio of long/short DNA fragments (Figure 37). For this reason, I decided to modify our protocol by replacing spin columns with magnetic beads. In brief, magnetic beads work by reversibly binding to DNA molecules in a solution. Beads are then pelleted with a magnet and the remainder of the sample can be washed away with ethanol. Finally, the DNA is detached from the beads by adjusting buffer condition. Longer DNA molecules bind more efficiently to the beads, so by reducing the concentration of beads it is possible to remove from the solution the LMW DNA fragments[161]. I found that a volume ratio of beads/sample of 0.4x successfully removed the LMW band on the electrophoresis gel, ultimately improving results of the barcoding PCR (Figure 38).

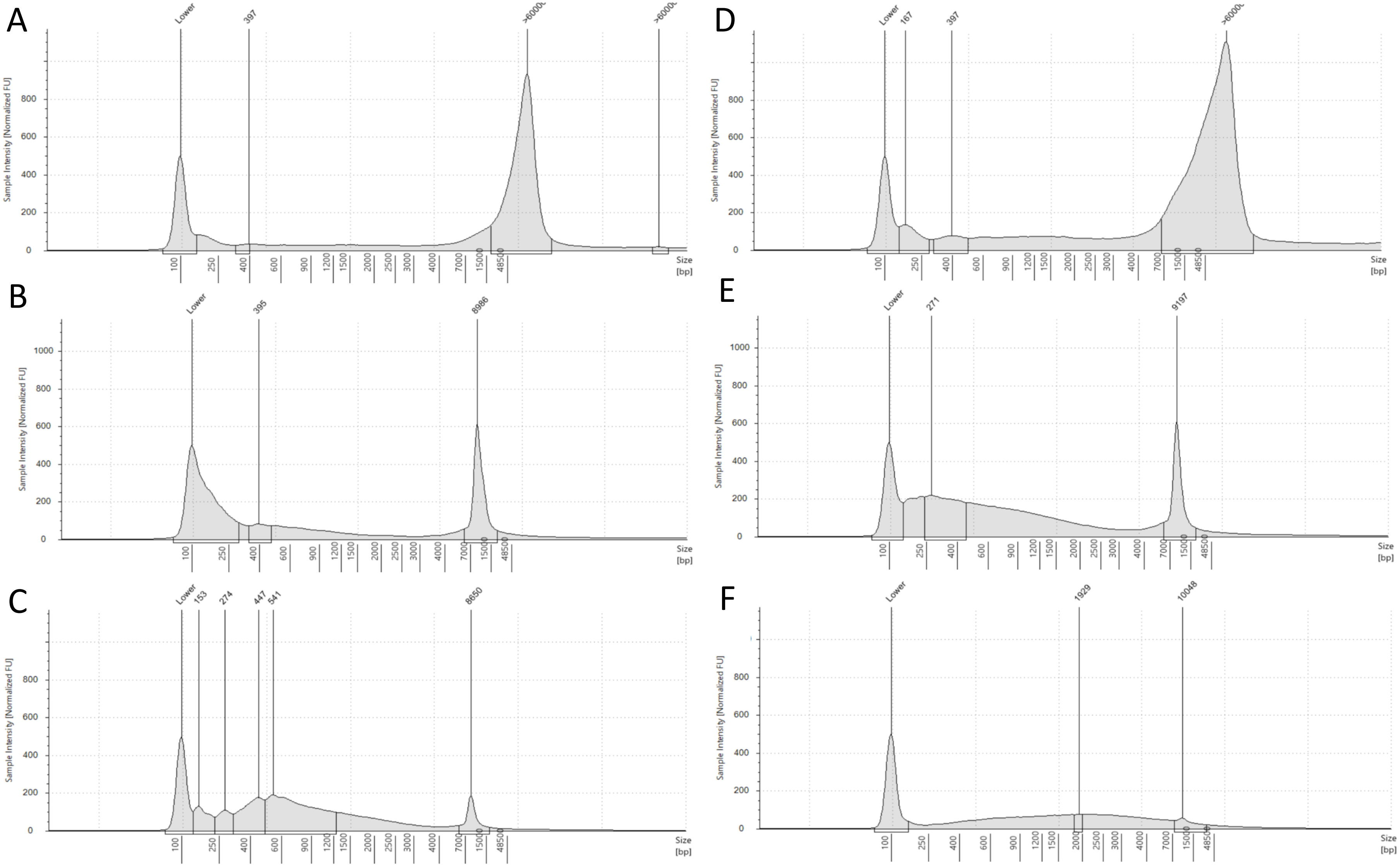


Figure 37: TapeStation (Agilent) results of one sample which succeeded barcoding PCR (A, B and C) and one sample which failed barcoding PCR (D, E and F). A: genomic DNA of successful sample. B: amplicon from amplifying PCR of successful sample. C: amplicon of barcoding PCR of successful sample. D: genomic DNA of failing sample. E: amplicon from amplifying PCR of failing sample. F: amplicon from barcoding PCR of failing sample.

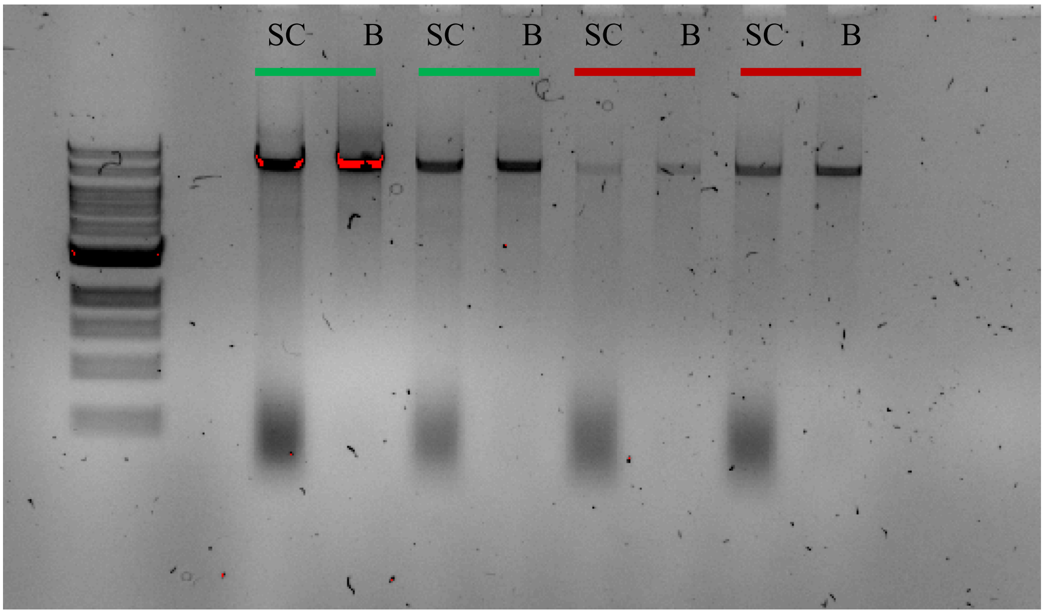


Figure 38: Comparison of purification methods on 0.8% agarose gel. Four samples, products of the amplifying PCR, were purified with either spin columns (SC) or 0.4x magnetic beads (B). The ~500bp smear at the bottom, corresponding to the LMW DNA fragments, is present with SC purification but not with beads purification. The intensity of the 8.9Kb band, indicating the amplicon, is unchanged.

3.3.6. Albacore vs Guppy

The original pipeline for analysing ONT data included basecalling with Albacore[162]. However, Albacore's development and support has been discontinued by ONT in 2019. For this reason, I switched to the newer ONT proprietary basecaller Guppy. A total of 137 samples that had already been analysed with Albacore were also basecalled with Guppy, and downstream genotyping results were 100% concordant.

3.3.7. The adjusted quality score improves Nanopolish precision

The original pipeline used Nanopolish to call variants in *GBA*[149]. However, Nanopolish produces a high number of false positive calls, even after excluding INDELS (Table 25). To improve the precision (positive predicting value) and recall (sensitivity) of the pipeline, we tried to filter results for the QS Nanopolish provides for each variant called, calculated using a Hidden Marcov Model (HMM). HMM can estimate the probability of an observed sequence over all possible sequences[163]. One limitation of Nanopolish QS is that it increases linearly with the depth of coverage at the position the variant is called (Figure 39), and since each sample has a different depth of coverage, it is impossible to set a fixed threshold for accepting or rejecting variants. To overcome this limitation, I decided to normalise the Nanopolish QS with the depth of coverage of the variant called, getting a "adjusted quality score" (aQS). Since the aQS is not affected by depth of coverage (Figure 40), it is possible to set a threshold for accepting calls. Upon analysis of the first 92 samples sequenced with nanopore, we decided to fix this threshold at 1.8. This approach was published in our original methods paper[149]. When comparing results of unfiltered Nanopolish with filtering with aQS in 7 samples (one 84GG / wt, one Arg502Cys / wt, one L483P/N409S, three wt / wt), precision increased form 0.78 to 0.92 when considering SNVs only and from 0.50 to 0.84 when considering all variants (including indels – Table 25). Recall did not change as no false negative variants were detected. Although the value of this estimate is limited as no Sanger sequencing was carried out on these samples, Sanger sequencing on other samples showed that no variants were being missed by nanopore sequencing (see section 3.3.2).

A similar approach to achieve the same result would have been to down-sample all samples to the same depth of coverage and then use the Nanopolish QS to filter variants. Advantages and disadvantages of both approaches are addressed in the discussion. Despite the improvement of introducing the aQS, Nanopolish was still occasionally producing false positive results, in particular one SNS in exon 11 (hg38.chr1: 155235011). Moreover, Nanopolish was occasionally calling homozygous variants as heterozygous (Table 25).

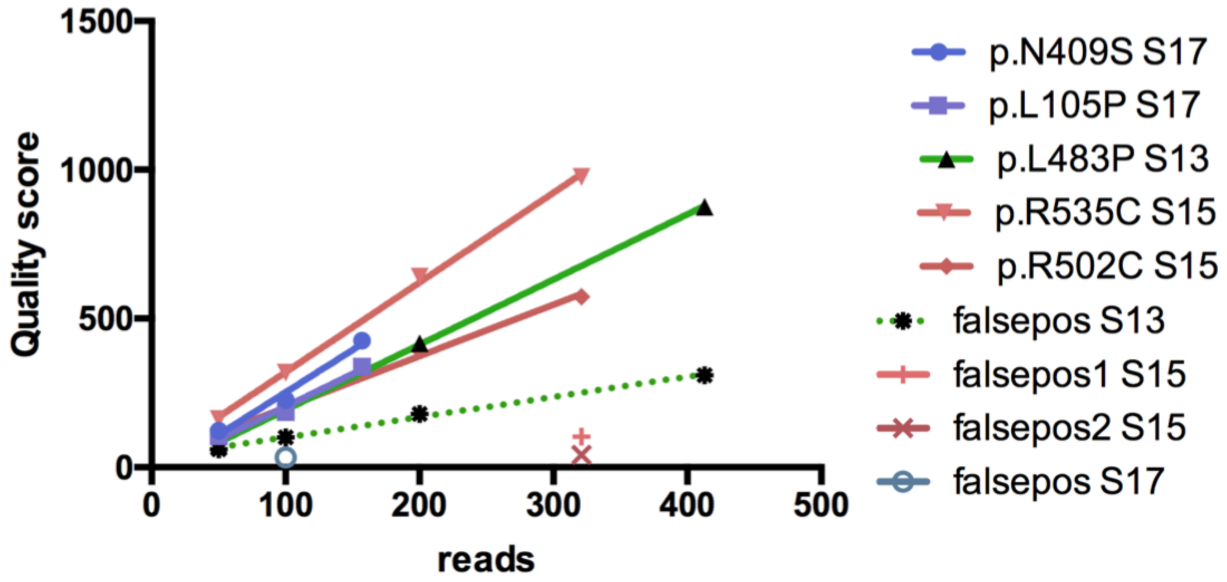


Figure 39: Nanopolish Quality score increases linearly with depth of coverage (reads). A total of 3 samples (S17, S13 and S15) were analysed after down-sampling to different depths of coverage and the QS of variants were reported. (Figure reproduced from Salazar et al[149], under Creative Commons CC BY license).

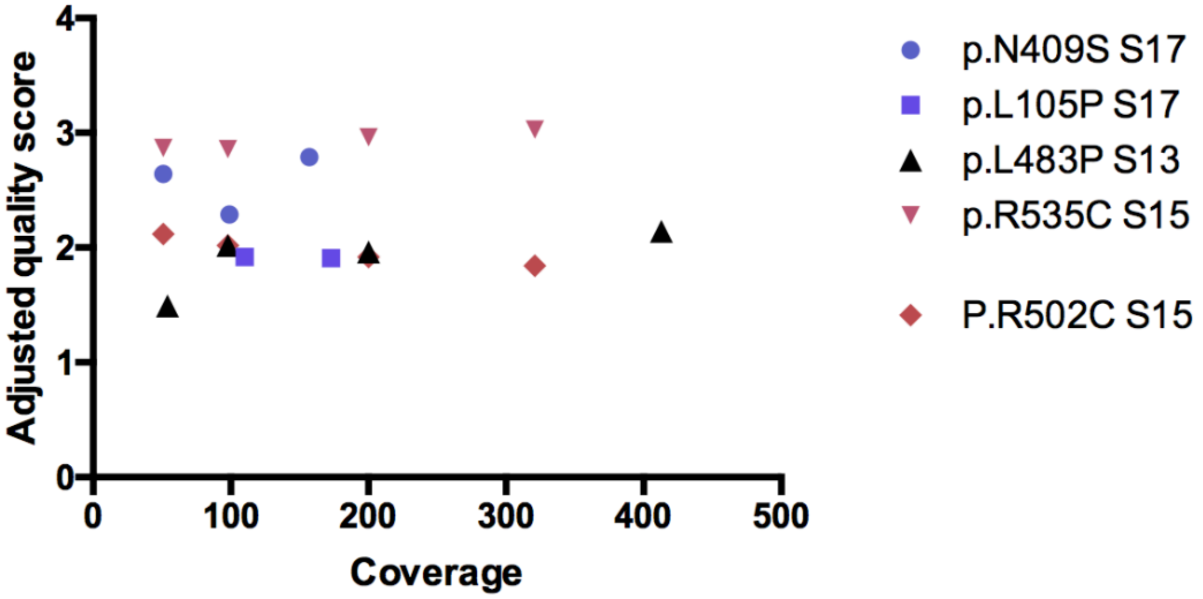


Figure 40: Nanopolish adjusted quality score does not change with depth of coverage. A total of 3 samples (S17, S13 and S15) were analysed after down-sampling to different depths of coverage and the QS of variants were reported. (Figure reproduced from Salazar et al[149], under Creative Commons CC BY license).

3.3.8. Strand filtering

The position hg38.chr1: 155235011 is particularly challenging for Nanopolish. This is due to an excess of G called in that position (Figure 41). Nanopolish sometimes calls a variant in this position, even though this is clearly a false positive, and since it is located in an exon, this is a problem. I noticed that the excess G is usually only present on the forward strand (Figure 42) and decided to run Nanopolish on both strands separately and if the variant is called only on one of them (strand mismatch), then I can assume that the variant is a false positive. To do this, I created an ad hoc script that 1) creates a .bam file with alignment of each strand separately with Samtools, 2) run Nanopolish on both strands and on the original .bam file containing both strands, separately and 3) merges the 3 individual .vcf files into one which contains only variants that were called on both strands and on the original .bam file, with Samtools. The script can be found in Appendix – Code.

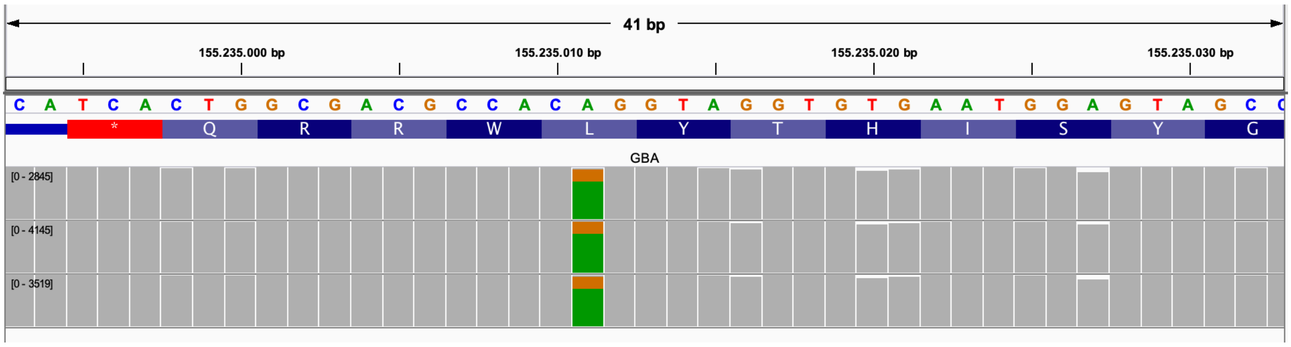

41 bp
155.235.000 bp
155.235.010 bp
155.235.020 bp
155.235.030 bp
C A T C A C T G G C G A C G C C A C A G G T A G G T G T G A A T G G A G T A G C C
* Q R R W L Y T H I S Y G
GBA
[0 - 2845]
[0 - 4145]
[0 - 3519]

Figure 41: Excess of G in position hg38.chr1:155235011 in 3 randomly selected samples. From top to bottom, the image shows genomic coordinates in Chromosome 1, genomic DNA sequence, aminoacid sequence, and sequencing results of three different samples. The green and ochre colors represent the persentage of A and G, respectively. Images exported from IGV.

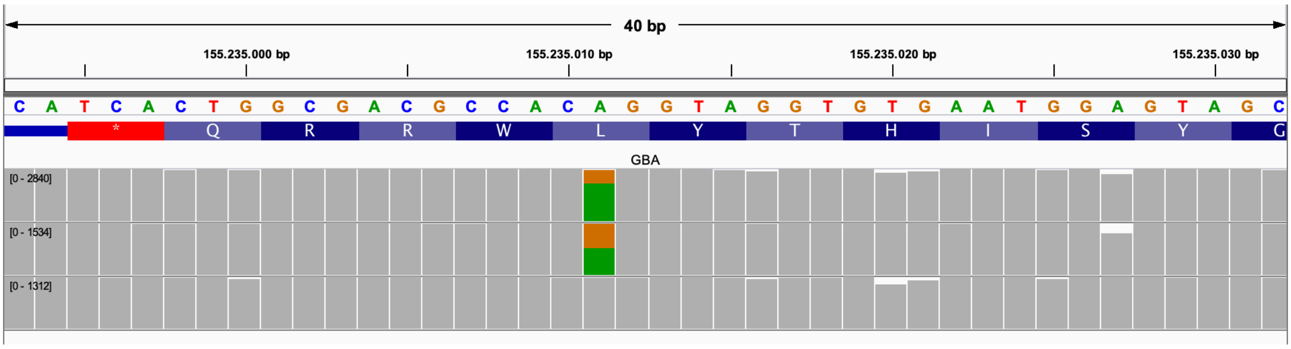

40 bp
155.235.000 bp
155.235.010 bp
155.235.020 bp
155.235.030 bp
C A T C A C T G G C G A C G C C A C A G G T A G G T G T G A A T G G A G T A G C
* Q R R W L Y T H I S Y G
GBA
[0 - 2840]
[0 - 1534]
[0 - 1312]

Figure 42: The excess G in position hg38.chr1:155235011 on the alignment (top row) is present on the forward strand (middle row) but not on the reverse strand (bottom row). From top to bottom, the image shows genomic coordinates in Chromosome 1, genomic DNA sequence, aminoacid sequence, and sequencing results of three different samples. The green and ochre colors represent the persentage of A and G, respectively. Images exported from IGV.

3.3.9. Clair is superior to Nanopolish in our dataset

Despite the aQS significantly improved Nanopolish performance, occasional false positive SNVs were still being called (in particular the exonic hg38.chr1: 155235011), as well as a consistent number of false positive indels. Moreover, Nanopolish was sporadically classifying as heterozygous variants that looked homozygous on IGV (Table 25). For these reasons, we decided to test another variant caller based on a machine learning algorithm, Clair[158]. At the time of the first testing, the models available were trained with a depth of coverage up to 70x (referred to as "Clair 70"). For this reason, all samples had to be down-sampled to a similar depth of coverage before being analysed. Recently, updated models with depth of coverage up to 550x were released (referred to as "Clair 550"). I tested both models and results are reported in Table 25. Since all samples had to be down-sampled to a fixed quantity, a threshold could be used to filter the variants called. I arbitrarily selected a threshold of 500 after looking at the data we generated and at the information available at https://github.com/HKU-BAL/Clair#pretrained-models. This is a conservative threshold which is unlikely to miss any variant but might produce some false positive calls. While Clair 70 showed similar results to Nanopolish with aQS and even missed one variant in one sample (false negative), Clair 550 was superior to Nanopolish, showing a precision of 1 when considering SNVs only and 0.87 when considering all variants (Table 25). The rate of false positive indels called is similar to that of Nanopolish with aQS. For this reason, I used a QS threshold of 700 for indels with Clair 550, reducing the rate of false positive calls to 0 in our dataset.

		nanopolish unfiltered	**nanopolish adjusted quality score**	**clair coverage 70x**	**clair coverage 550x**
S 1	variants called	14	8	8	8
	SNVs	10	8	8	8
	indels	4	0	1	0

	true positives (SNVs)	7	7	7	8
	false positives (SNVs)	2	0	0	0
	false negatives (SNVs)	0	0	1	0
	wrong GT (SNVs)	1	1	0	0
	false positive (indels)	4	0	1	0
	true positives (indels)	0	0	0	0
S 2	variants called	17	11	8	12
	SNVs	12	10	8	10
	indels	5	1	0	2
	true positives (SNVs)	9	9	8	10
	false positives (SNVs)	2	0	0	0
	false negatives (SNVs)	0	0	2	0
	wrong GT (SNVs)	1	1	0	0
	false positive (indels)	5	1	0	2
	true positives (indels)	0	0	0	0
S 3	variants called	16	9	9	9
	SNVs	10	8	8	8
	indels	6	1	1	1
	true positives (SNVs)	8	8	8	8
	false positives (SNVs)	2	0	0	0
	false negatives (SNVs)	0	0	0	0
	wrong GT (SNVs)	0	0	0	0
	false positive (indels)	6	1	1	1
	true positives (indels)	0	0	0	0
S 4	variants called	8	2	2	3
	SNVs	3	2	2	2
	indels	5	0	0	1
	true positives (SNVs)	2	2	2	2
	false positives (SNVs)	1	0	0	0

	false negatives (SNVs)	0	0	0	0
	wrong GT (SNVs)	0	0	0	0
	false positive (indels)	5	0	0	1
	true positives (indels)	0	0	0	0
S 5	variants called	20	13	11	14
	SNVs	14	12	10	12
	indels	6	1	1	2
	true positives (SNVs)	11	11	10	12
	false positives (SNVs)	2	0	0	0
	false negatives (SNVs)	0	0	2	0
	wrong GT (SNVs)	1	1	0	0
	false positive (indels)	6	1	1	2
	true positives (indels)	0	0	0	0
S 6	variants called	16	11	10	11
	SNVs	11	10	10	10
	indels	5	1	0	1
	true positives (SNVs)	9	9	8	10
	false positives (SNVs)	1	0	0	0
	false negatives (SNVs)	0	0	2	0
	wrong GT (SNVs)	1	1		0
	false positive (indels)	5	1	0	1
	true positives (indels)	0	0	0	0
S 7	variants called	24	14	13	13
	SNVs	12	11	10	10
	indels	12	3	3	3
	true positives (SNVs)	10	10	10	10
	false positives (SNVs)	2	1	0	0
	false negatives (SNVs)	0	0	0	0
	wrong GT (SNVs)	0	0	0	0

	false positive (indels)	11	2	2	2
	true positives (indels)	1	1	1	1
To t	Precision (SNVs only)	0.777777778	0.918032787	0.946428571	1
	Recall (SNVs only)	1	1	0.883333333	1
	Precision (all variants)	0.495652174	0.838235294	0.885245902	0.871428571
	Recall (all variants)	1	1	0.885245902	1

Table 25: Results of comparison of different variants-calling approaches. SNV: single nucleotide variants (intended here as single base substitutions). The "all variants" rows indicate SNVs and indels.
Wrong GT: the variant is called as heterozygous instead of homozygous.

3.3.1. Detection of suspicious coverage drops in homopolymers

Upon receiving results of *GBA* sequencing through the RAPSODI portal, one participant got in touch with the RAPSODI team reporting a mismatch between these results and previous results he obtained independently. Specifically, our ONT analysis did not detect any variant in coding regions of *GBA*, while previous independent genetic testing detected a deletion in chr1:155239657 (NM_000157.3:c.413del, p.Pro138Leufs*62). I confirmed the presence of the deletion by Sanger sequencing (Figure 43).

This variant was located within a homopolymer (poly-G). *GBA* contains two homopolymer stretches in exonic regions (chr1:155239990-155239995 and chr1:155239657-155239661). ONT has limited resolution in the presence of these homopolymers and usually erroneously produces a drop of coverage at the 3' end of the homopolymer, as shown in Figure 44. For this reason, variants callers usually discard deletions in these positions, making it challenging to detect any deletion within a homopolymer. However, from visual inspection it was evident that the sample from this participant showed a greater drop in coverage than that observed in other samples (Figure 44).

This suggested it might still be possible to detect deletions within homopolymers from ONT data by looking at the level of coverage drop at these positions.

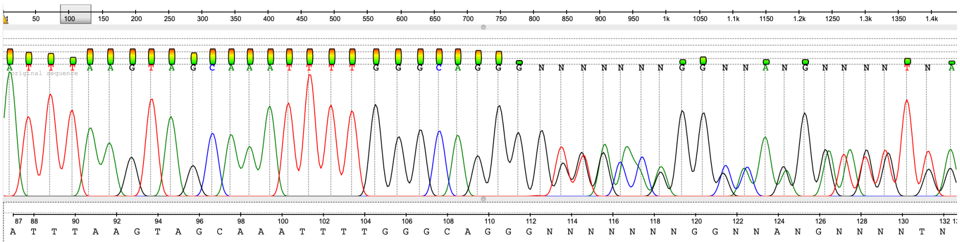
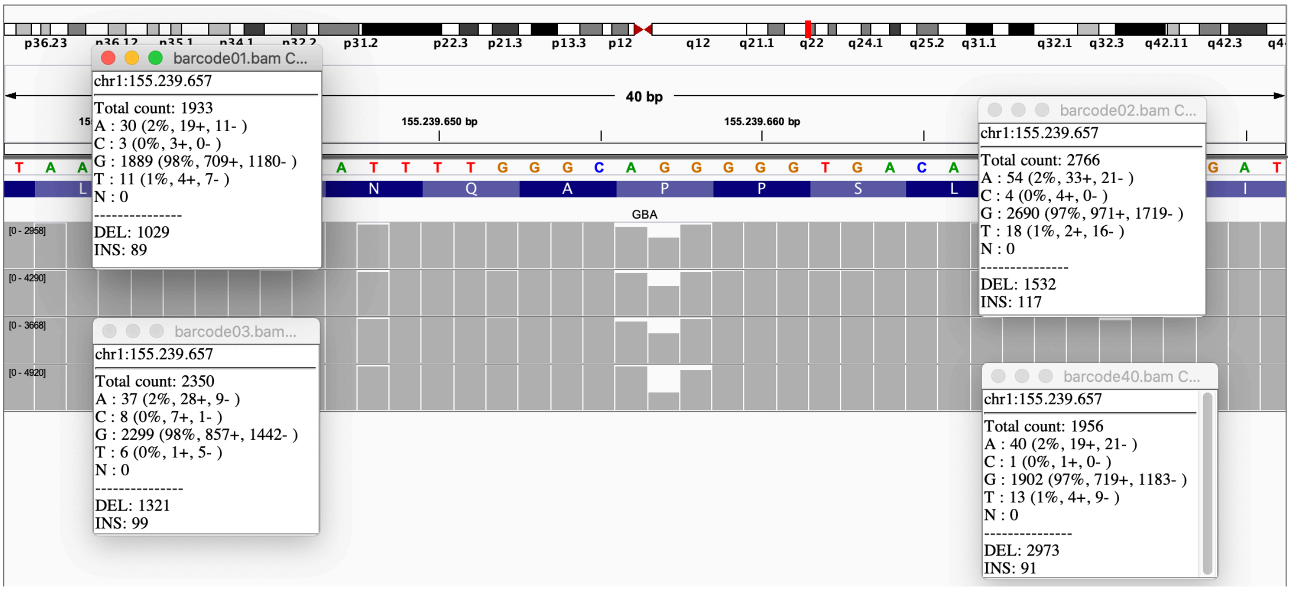


Figure 43: Chromatogram of Sanger sequencing of the single base deletion that was initially missed by ONT (chr1:155239657). At position 113 (corresponding to chr1:155239657), the chromatogram suddenly turns to double sequence, indicating a single base deletion.



barcode01.bam C...

chr1:155.239.657

Total count: 1933
A : 30 (2%, 19+, 11-)
C : 3 (0%, 3+, 0-)
G : 1889 (98%, 709+, 1180-)
T : 11 (1%, 4+, 7-)
N : 0

DEL: 1029
INS: 89

barcode02.bam C...

chr1:155.239.657

Total count: 2766
A : 54 (2%, 33+, 21-)
C : 4 (0%, 4+, 0-)
G : 2690 (97%, 971+, 1719-)
T : 18 (1%, 2+, 16-)
N : 0

DEL: 1532
INS: 117

barcode03.bam...

chr1:155.239.657

Total count: 2350
A : 37 (2%, 28+, 9-)
C : 8 (0%, 7+, 1-)
G : 2299 (98%, 857+, 1442-)
T : 6 (0%, 1+, 5-)
N : 0

DEL: 1321
INS: 99

barcode40.bam C...

chr1:155.239.657

Total count: 1956
A : 40 (2%, 19+, 21-)
C : 1 (0%, 1+, 0-)
G : 1902 (97%, 719+, 1183-)
T : 13 (1%, 4+, 9-)
N : 0

DEL: 2973
INS: 91

Figure 44: Drop in coverage within the poly-G sequence in 3 wild-type samples (barcodes 01, 02 and 03, upper 3 samples) and in the sample suspected to be a false negative (barcode 40, bottom sample). From top to bottom, the image shows genomic coordinates in Chromosome 1, genomic DNA sequence, aminoacid sequence, and the gray bars represent the depth of coverage of the 4 samples at each position. The side windows show proportion of A, C, G, T, deletions and insertions at position chr1:155239657 for the 4 samples. Images exported from IGV.

To detect abnormal coverage drops in these regions, that could signify an underlying single base deletion not detectable with ONT sequencing, I developed a novel approach. This consists in estimating a new parameter, the “adjusted depth” (AD), which corresponds to the depth of coverage at a given position divided by the mean depth of coverage at the 100 flanking positions (50 bases to the 5’ side and 50 bases to the 3’ side).

To achieve this I: 1) created a BASH script that calls “Samtools depth” to produce a csv file with depth of coverage at each position from a .bam file and 2) created a python file called fINDEL that estimates the AD for each of these positions.

The code is reported in Appendix – Code.

I tested this approach on the exonic regions of *GBA* in 92 samples sequenced with the ONT pipeline. These included the sample from the participant with the undetected deletion. With a set AD threshold of 0.5, the false negative deletion was correctly detected, and 3 additional variants were reported in some samples: chr1:155234775, chr1:155234800 and chr1:155239990 (Table 26). These additional variants were interpreted as false positives, as they were present in a high number of samples. This assumption was supported by visual inspection on IGV.

It was evident that, by setting a fixed threshold, there were some false positive results. To overcome this problem, I decided to compare the AD at each position with the AD at the same position in all the other samples. I created a R script that flags all instances where a sample has a AD that is more that 5 median absolute deviations (MAD) from the mean of all the other samples at that same position (Figure 47). With this approach, the indel that was missed by standard ONT analysis was detected, and no false positives were reported. The R script can be found in Appendix – Code.

Interestingly, this analysis also showed that one sample had a significantly higher depth compared to all the others at position chr1:155239990. Upon visual inspection, I observed that the coverage drop was indeed less than expected, and this was caused by the presence of a higher than expected (30%) representation of the base T, potentially due to a base substitution at chr1:155239989 (C>T) (Figure 45 and Figure 47). Sanger sequencing of exon 3 of *GBA* confirmed the presence of a variant at chr1:155239989 (C>T) (Figure 46).

Variants detected	N° of samples variant was detected in	AD*
chr1:155239657	1	0.41
chr1:155234775	92	0.42
chr1:155234800	92	0.40
chr1:155239990	91	0.47

Table 26: Results of analysis of 92 samples with AD. Chr1:155239657 is a true positive, while chr1:155234775, chr1:155234800 and chr1:155239990 are false positives.

**AD: adjusted depth, see text for more details.*

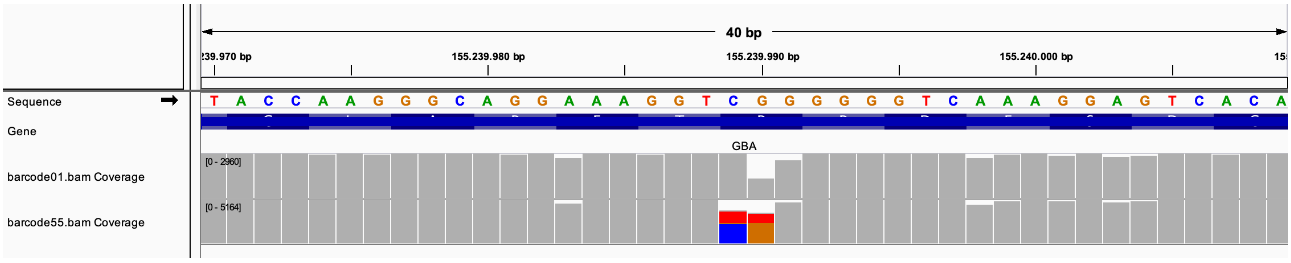


Figure 45: Possible single base substitution in poly-G sequence. At the top is a wild type sample for reference, at the bottom the sample with a higher coverage at position chr1:155239990. From top to bottom, the image shows genomic coordinates in Chromosome 1, genomic DNA sequence, aminoacid sequence, and sequencing results of the 2 samples. The red, blue and ochre colors represent the proportion of T, C and G at positions of interest. Images exported from IGV.

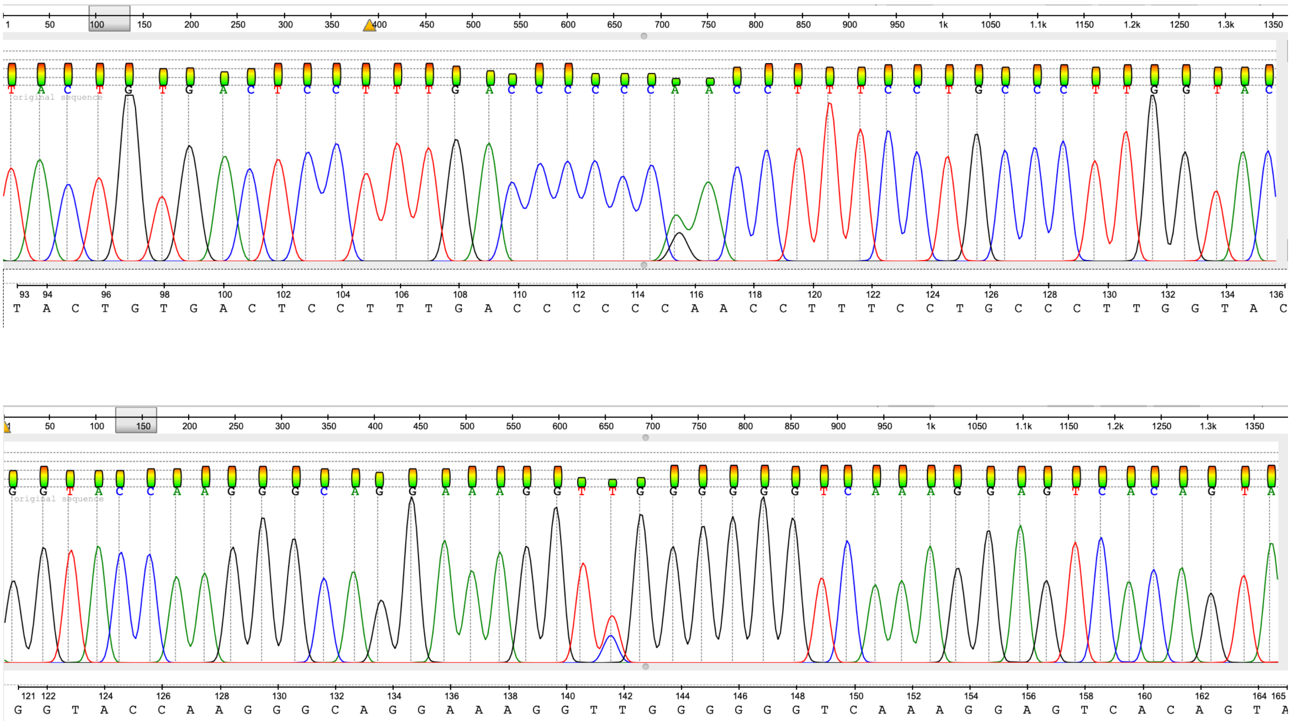


Figure 46: Chromatogram of Sanger confirmation of variant nearby homopolymer. Top row: forward strand. Bottom row: reverse strand. At position 115 (forward strand) and 141 (reverse strand) a heterozygous single base substitution can be seen, corresponding to position chr1:155239990.

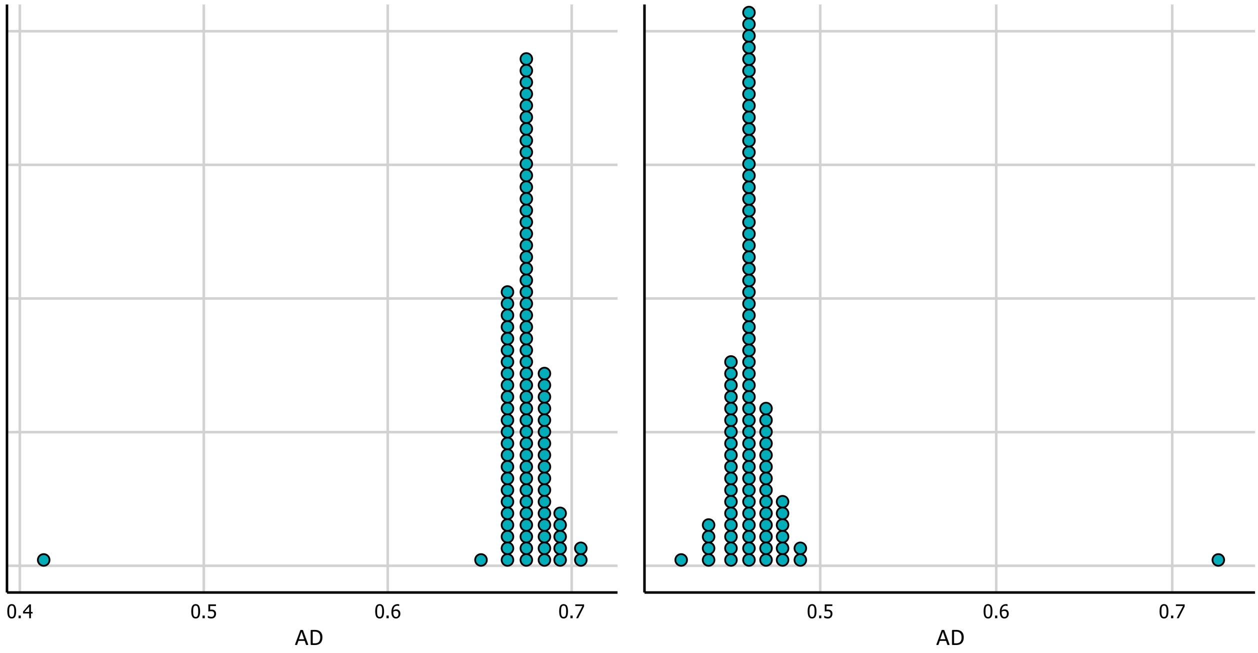


Figure 47: Analysis of AD highlights two variants within homopolymers. On the left, a single base deletion at chr1:155239657, with AD that is more than 5 MAD lower than the mean of the other samples at that position. On the right, a single base substitutionat chr1:155239990, causing the the AD to be more than 5 MAD higher than the other samples at that position.

3.4. Discussion

Being able to reliably sequence the *GBA* gene is paramount to understand its role in the development of PD.

However, sequencing *GBA* is a challenging task. Short read approaches are intrinsically problematic, due to the presence of *GBAP1*, and many exome and WGS protocols failed to reliably detect some *GBA* variants[110,111,164].

The work of dr. Proukakis to apply ONT to sequence *GBA* suggested ONT might have been the best solution to my problem, and by the time I started my PhD, a few samples were already successfully sequenced in our UCL laboratories. However, the ONT method needed to be validated and adapted for work on DNA extracted from saliva and for higher throughput. Some of the improvements to the method that are described in this chapter were included in the validation paper published in 2019[149] and others will be reported in an additional manuscript which is currently under submission[165].

In this chapter, I describe improvements to both the samples preparation protocol and the bioinformatics pipeline (summarised in Figure 48).

The samples preparation improvements were needed as DNA extracted from saliva has a lower quality, and required revisiting of the PCR conditions and the technique for purification of the PCR product. In particular, the introduction of magnetic beads for DNA purification and major changes to the PCR protocol dramatically increased the rate of success and quality of the sequencing.

PCR enrichment ensures a high coverage and the possibility to sequence up to 96 samples with one single MinION run. However, PCR has limitations, such as the inability to amplify some specific recombinants, as discussed in chapter 4. Future development might focus on the application of alternative means of enrichment, like Cas9[166], which could be used to enrich the region including both *GBA* and its pseudogene *GBAP1*, revealing the complexity of its structural variants. Of note, the laboratory protocol I developed has been adopted by Exeter Genomics Laboratory at the Royal Devon & Exeter Hospital and implemented into their automated pipeline to streamline sequencing of *GBA* for our research purposes.

The bioinformatics pipeline has also been updated. A different base-caller (Guppy) was preferred as the original one (Albacore) was discontinued. As Guppy is the base-caller supported by ONT, and it is likely it will be the default caller in the future. For this reason, including it into the pipeline is an important update. Guppy improved the quality of the *in silico* analysis, but requires higher computing resources, including a powerful GPU. This can be a limiting factor in many settings; we solved the problem by moving the basecalling step on the UCL computing platform Myriad. The alignment step is particularly delicate for *GBA* sequencing, due to the presence of *GBAP1,* which can cause misalignment as previously discussed. Our pipeline uses NGMLR, which searches for the best fit of each k-mers in the whole reference and is thus able to improve alignement in complex regions[157], such as the *GBA* locus. A different aligner, Minimap2, is widely used for long read sequencing data[167]. I did some comparisons between the two, but not to an extent that was sufficient to draw any conclusions, so these are not presented in this thesis. However, NGMLR has proved reliable in our analysis and we never felt the need to improve the alignment step further. Future research might focus on the comparison of different aligners, including Minimap2.

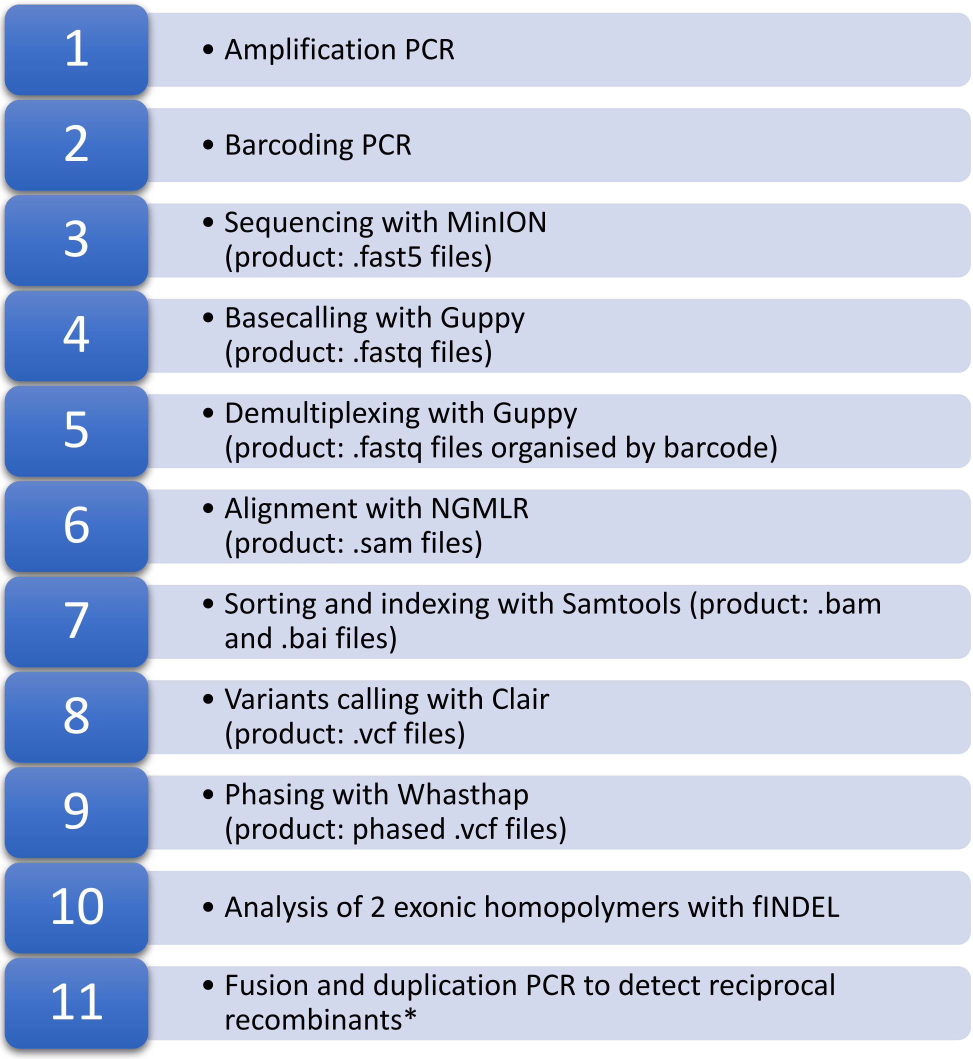
A range of variants-callers were tested and we finally settled for the machine learning based caller Clair. Clair produced less false positive calls and improved the overall accuracy of the variant-calling step. However, Nanopolish recently added a function to report the support fraction for each allele by strand and it would be interesting to investigate whether this improvement solved the issues the software was showing in our dataset.

As the field of bioinformatic analysis of long read sequencing is rapidly evolving, further optimisation will likely be needed in the future.

Possibly the most relevant achievement of this part of my PhD is the development of a method to reliably detect variants within homopolymers in the *GBA* gene. Homopolymers are problematic for ONT, as the technology struggles to determine the length of stretches of the same base[168], and GBA has two homopolymer stretches in exonic regions. The approach described here is novel and was able to successfully detect two variants, one single base deletion and one synonymous SNV, within *GBA* homopolymers. The method consists in comparing the AD of each base within the homopolymer between all samples in the same run. This allows to spot samples that give a signal that is significantly different from that of all the other samples, indicating a base change within the homopolymer. One major limitation of this approach Is that it can only be applied to MinION sequencing runs where multiple samples were analysed at the same time. In the future, it might be possible to create a map of expected AD at all positions within homopolymers, so that it can then be applied to individual samples.

As a final remark, while improving our method for sequencing *GBA* with ONT, I also contributed to the development of the software we were using. Suggestions made on the *Github* pages of both Nanopolish and Clair were accepted by the developers (Figure 49 and Figure 50) and will hopefully lead to advancement of the software. I collaborated with Dr Sedlazeck, developer of NGMLR, for the development of a method for detecting SVs in GBA (discussed in the following chapter) and was able to provide him with useful feedback on the functioning of NGMLR.

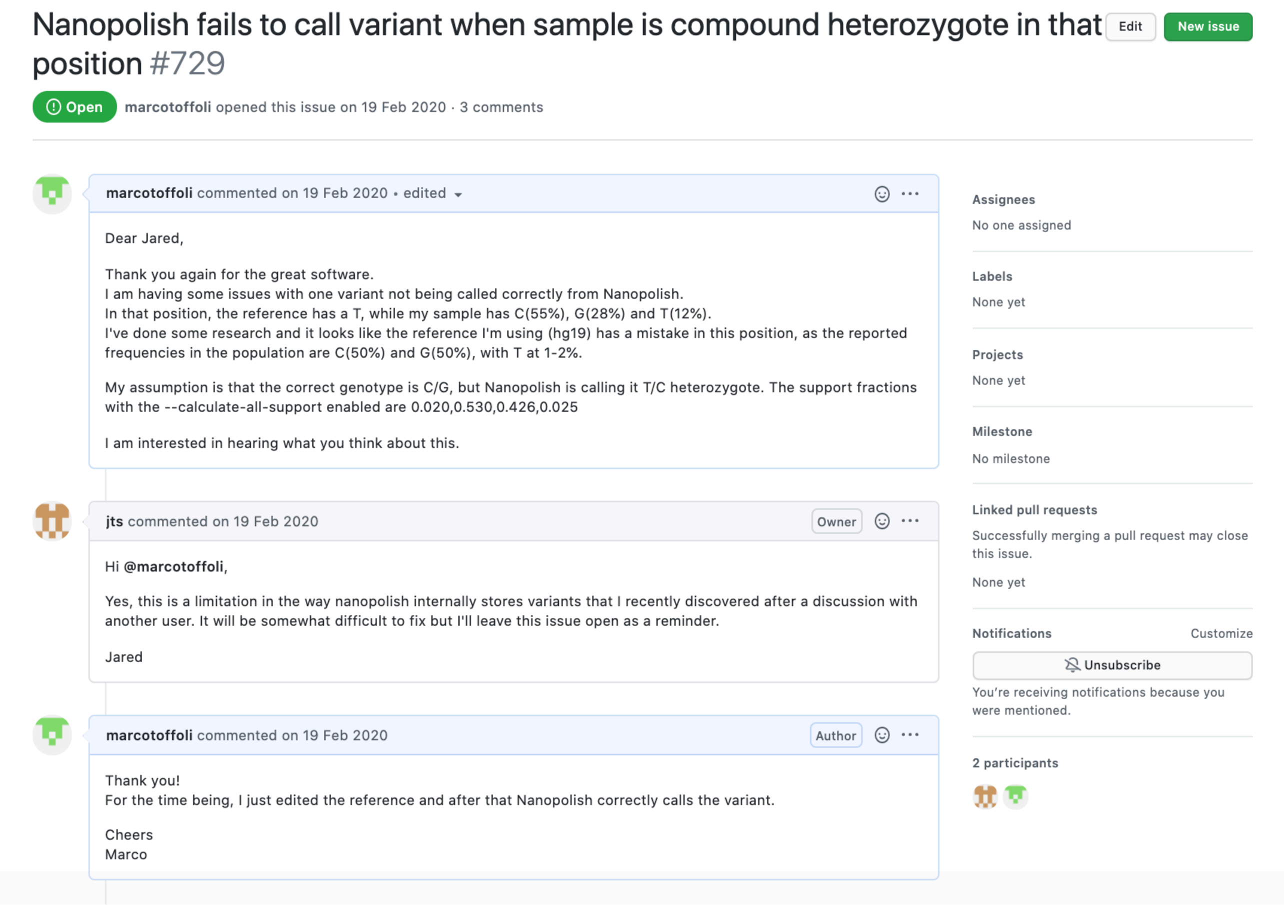
In the next Chapter, I discuss additional research on complex structural variants in the *GBA* gene, and a novel method I developed for their detection and characterisation.



1. Amplification PCR
2. Barcoding PCR
3. Sequencing with MinION (product: .fast5 files)
4. Basecalling with Guppy (product: .fastq files)
5. Demultiplexing with Guppy (product: .fastq files organised by barcode)
6. Alignment with NGMLR (product: .sam files)
7. Sorting and indexing with Samtools (product: .bam and .bai files)
8. Variants calling with Clair (product: .vcf files)
9. Phasing with Whasthap (product: phased .vcf files)
10. Analysis of 2 exonic homopolymers with fINDEL
11. Fusion and duplication PCR to detect reciprocal recombinants*

Figure 48: Flowchart of the optimised pipeline to fully sequence the GBA gene.

**More details on additional analysis of reciprocal recombinants in chapter 4.*



Nanopolish fails to call variant when sample is compound heterozygote in that position #729

marcotoffoli opened this issue on 19 Feb 2020 · 3 comments

marcotoffoli commented on 19 Feb 2020 • edited

Dear Jared,

Thank you again for the great software.
I am having some issues with one variant not being called correctly from Nanopolish.
In that position, the reference has a T, while my sample has C(55%), G(28%) and T(12%).
I've done some research and it looks like the reference I'm using (hg19) has a mistake in this position, as the reported frequencies in the population are C(50%) and G(50%), with T at 1-2%.

My assumption is that the correct genotype is C/G, but Nanopolish is calling it T/C heterozygote. The support fractions with the --calculate-all-support enabled are 0.020,0.530,0.426,0.025

I am interested in hearing what you think about this.

jts commented on 19 Feb 2020 Owner

Hi **@marcotoffoli**,

Yes, this is a limitation in the way nanopolish internally stores variants that I recently discovered after a discussion with another user. It will be somewhat difficult to fix but I'll leave this issue open as a reminder.

Jared

marcotoffoli commented on 19 Feb 2020 Author

Thank you!
For the time being, I just edited the reference and after that Nanopolish correctly calls the variant.

Cheers
Marco

Assignees
No one assigned

Labels
None yet

Projects
None yet

Milestone
No milestone

Linked pull requests
Successfully merging a pull request may close this issue.
None yet

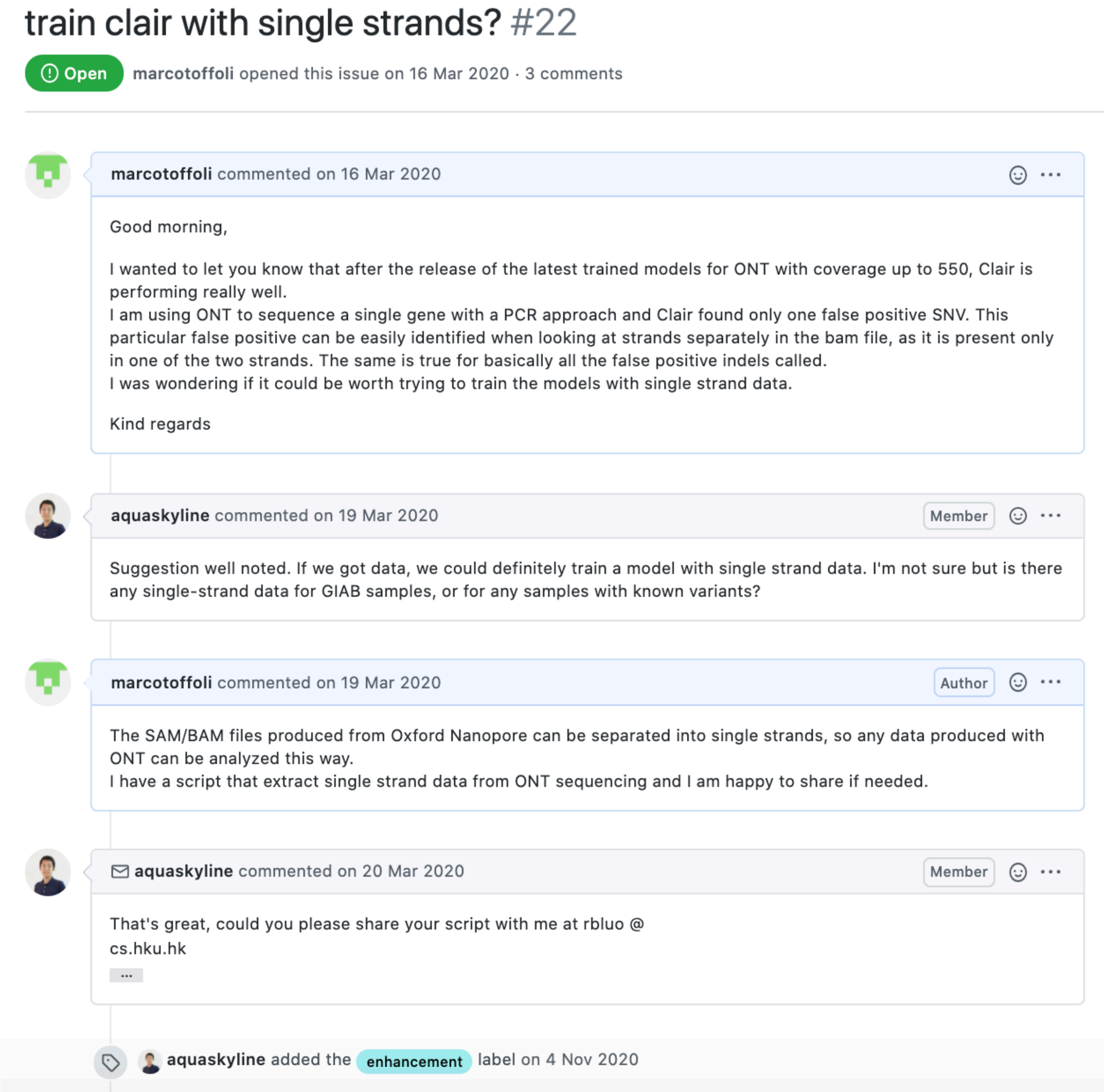
Notifications Customize

You're receiving notifications because you were mentioned.

2 participants

Figure 49: Suggestion I made to improve Nanopolish

-



train clair with single strands? #22

Open marcotoffoli opened this issue on 16 Mar 2020 · 3 comments

marcotoffoli commented on 16 Mar 2020

Good morning,

I wanted to let you know that after the release of the latest trained models for ONT with coverage up to 550, Clair is performing really well.
I am using ONT to sequence a single gene with a PCR approach and Clair found only one false positive SNV. This particular false positive can be easily identified when looking at strands separately in the bam file, as it is present only in one of the two strands. The same is true for basically all the false positive indels called.
I was wondering if it could be worth trying to train the models with single strand data.

Kind regards

aquaskyline commented on 19 Mar 2020 Member

Suggestion well noted. If we got data, we could definitely train a model with single strand data. I'm not sure but is there any single-strand data for GIAB samples, or for any samples with known variants?

marcotoffoli commented on 19 Mar 2020 Author

The SAM/BAM files produced from Oxford Nanopore can be separated into single strands, so any data produced with ONT can be analyzed this way.
I have a script that extract single strand data from ONT sequencing and I am happy to share if needed.

aquaskyline commented on 20 Mar 2020 Member

That's great, could you please share your script with me at rbluo @
cs.hku.hk
...

aquaskyline added the enhancement label on 4 Nov 2020

Figure 50: Suggestion I made to improve Clair

4. Characterisation of structural variants in the *GBA* gene

Content of this chapter will be reported in a separate publication, which is currently under submission and available on Medrxiv[165].

4.1.Overview and rationale

The *GBA* gene can be affected by non-reciprocal recombination (gene crossover) and reciprocal gene fusion and gene duplication recombinantion (Figure 3).

In chapter 3, I describe my contribution to the development and optimisation of a method for sequencing *GBA* with ONT. This method was able to detect and characterise non-reciprocal recombinant alleles with ONT[149].

However, from the analysis of the first samples we noticed that a complex SV was not detected. The SV was a reciprocal recombination resulting in a fusion between the gene and pseudogene, which caused the deletion of the binding site of the reverse primer used to amplify the *GBA* gene and the subsequent non-amplification of the recombinant allele (Figure 51). Moreover, reciprocal gene duplication alleles were also not amplified, as the primers we were using (in this chapter referred to as primer pair A) were not suitable(Figure 51).

In this chapter, I detail the development of a PCR approach for detecting reciprocal SVs in *GBA*, determine their breakpoints and predict their pathogenicity.

To gain orthogonal validation, I use dPCR and adaptive sampling with UNCALLED. UNCALLED (Utility for Nanopore Current Alignment to Large Expanses of DNA) is a software, based on the MinKNOW ReadUntil API[169], that analyses the sequence of the DNA molecule passing through each pore of the ONT device in real time. It can then trigger the premature ejection of the DNA molecule from the pore if it does not match a reference provided by the user, freeing up sequencing capacity. This process allows for purely *in silico* real time enrichment of a region of interest while sequencing whole genome DNA[170].

Sequencing of *GBA* from short read WGS is notoriously challenging[109]. The team of Dr Michael Eberle, at Illumina, is developing a novel caller for GBA, called Gauchian[165]. We were contacted to validate Gauchian with our ONT method and in the last part of this chapter, I describe the results of this cross-platform validation, which provides further confirmation of my findings.

4.2. Methods

4.2.1. Samples

DNA samples included in the analysis were collected through RAPSODI, as discussed in section 3.2. Additional DNA was extracted from post-mortem brain samples from the Queen Square Brain Bank (QSBB). For the cross-validation of Gauchian (see section 4.3.5), samples were sourced at the National Human Genome Research Institute (NHGRI - ordered from the Coriell Institute for Medical Research at www.coriell.org) and at the Parkinson's Progression Marker Initiative (PPMI)[171].

4.2.2. PCR approach

4.2.2.1. PCRs to detect reciprocal recombinants

To detect and amplify the reciprocal recombinants in *GBA*, I used a PCR approach. The primers sequences were obtained from a previous publication[154], and specifically designed to amplify the reciprocal recombinants. Two sets of primers were used, to amplify the fusion recombinants (GBA-nf/MTX1-r – primer pair B) and duplication recombinants (ΨGBA-nf/ ΨMTX1-r - primer pair C), respectively (Figure 51). The primers were edited to accommodate the ONT barcode adaptor sequence. Two additional PCRs, with primer pairs B and C, were carried out for all samples. The product of the PCR was then run on a 0.8% agarose gel. If an amplicon was detected by the presence of a band on the gel, this meant that sample carried a fusion recombinant (primer pair B) or a duplication recombinant (primer pair C). Primer sequences and PCR conditions can be found in Table 27.

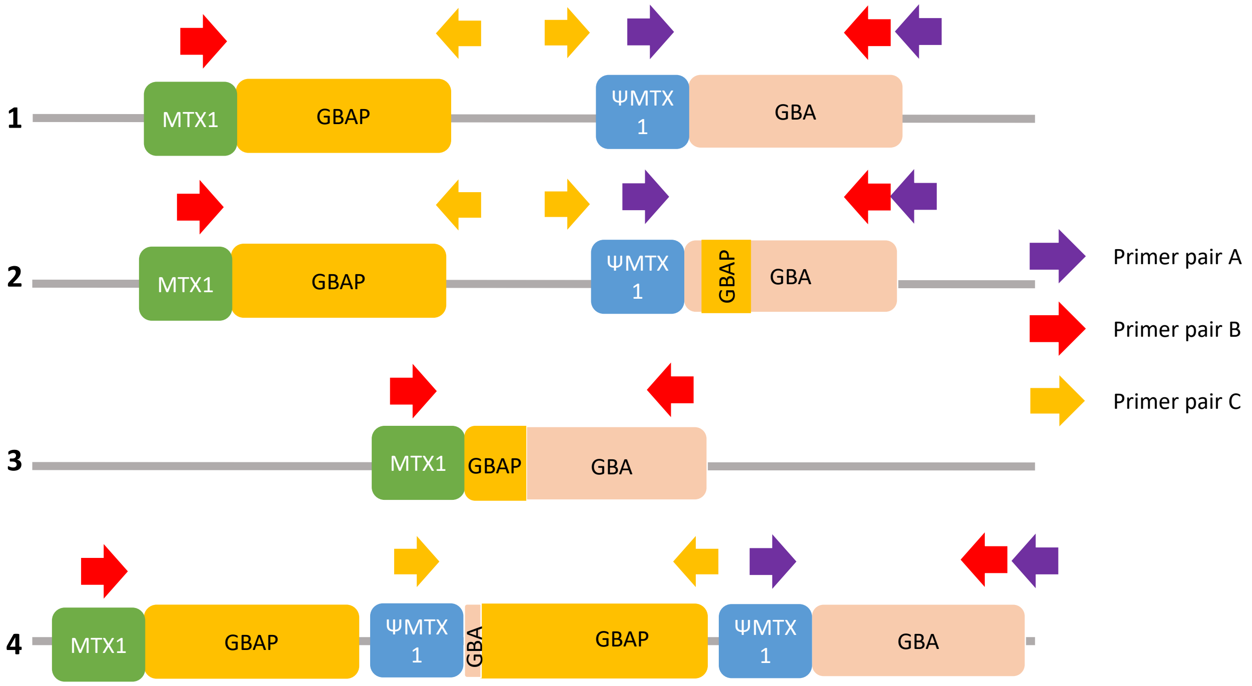


Figure 51: Primers used to detect the different recombinant alleles within the GBA gene.

1) Wild type allele. Only primer pair A produces an amplicon, as the other two pair are too distant (primer pair B) or inverted (primer pair C)

2) Non-reciprocal allele. Only primer pair A produces an amplicon, same as 1.

3) Reciprocal gene fusion allele. Only primer pair B produces an amplicon, as the binding site of the other primer pairs are lost (except for the forward primer of pair A)

4) Reciprocal gene duplication allele. Both primer pair A and C produce an amplicon.

Reagents	Volume		°C	Time	cycles
Template DNA (100ng)	x µL		95	3min	1
Nuclease free water	24.5µL - x		95	15sec	12
			62	15sec	
Barcode mix	0.5µL		65	10min	
			65	10min	1
LongAmp Taq 2x polymerase mix (NEB)	25µL		4	hold	
Primer pair B*	Forward	5'-TTTCTGTTGGTGCTGATATTGCATGTGTCCATTCTCCATGTCTTCA-3'			
	Reverse	5'-ACTTGCCTGTCGCTCTATCTTCAGCCTTCCTTCCTTCCCTGCAT-3'			
Primer pair C*	Forward	5'-TTTCTGTTGGTGCTGATATTGCGTGTCCGTTCTCCACATCCTTG-3'			
	Reverse	5'-ACTTGCCTGTCGCTCTATCTTCCCAACCTTTCTTCCTTCTTCTCAA-3'			

Table 27: PCR conditions and primers sequences to detect reciprocal recombinant alleles.

**primers are edited to carry the ONT adaptor sequence.*

4.2.2.2. Sequencing of recombinant alleles

If one of the two PCRs with primer pairs B and C produced an amplicon, these samples were barcoded and sequenced as described in chapter 3. However, the bioinformatics to analysis the sequencing data is different, as described in section 4.2.3.

4.2.3. In Silico characterisation of recombinants

4.2.3.1. Basecalling, Alignment to the human genome and SV calling

Basecalling of both the amplicon and UNCALLED samples was carried out with Guppy (version 4.2.2), while alignment to the human genome (version GRCh38.p13) with NGMLR (version 0.2.7)[157], LAST (Version 1243)[172], or Minimap2 (version 2.18)[167]. File manipulation was carried out with Samtools (version 1.10)[173] and aligned files were visually inspected with IGV (version 2.11.1)[174]. With LAST, the reference was indexed every 10th base. Three structural variant callers were tests: Sniffles (version 1.0.12a)[157], CuteSV (version 1.0.10)[175] and NanoSV[176].
A list of the commands used is provided in Appendix – Code.

4.2.3.1. Prediction of breakpoints with sentinel positions

The 3' ends of *GBA* and *GBAP1* are highly homologous, with the exception of a limited number of positions. I created a list of these positions, referred to as "sentinel" positions (Table 28 and Figure 52). Using LAST, samples were then aligned to a custom reference, masked for *GBAP1,* and inspected on IGV. If samples carried variants in any of the sentinel positions, that meant that the DNA in that position originated from the pseudogene, so the breakpoint within *GBA* had to be upstream of that position and downstream of the first sentinel position without evidence of SNVs on IGV (Figure 53). Once the method was established, it was then automated by using Clair to call variants on the fusion recombinants aligned to the custom reference after masking *GBAP1*.

Chr1:155233639
Chr1:155235203
Chr1:155235217
Chr1:155235252
Chr1:155235379
Chr1:155235412
Chr1:155235540
Chr1:155235727
Chr1:155235750
Chr1:155235878

Table 28: Sentinel positions (i.e. positions within the highly homology region where GBA and GBAP1 do not share the same sequence).

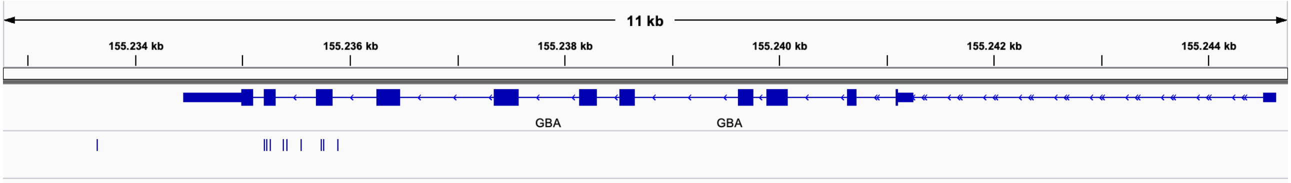


Figure 52: Visual representation of the sentinel positions at the 3' of GBA.

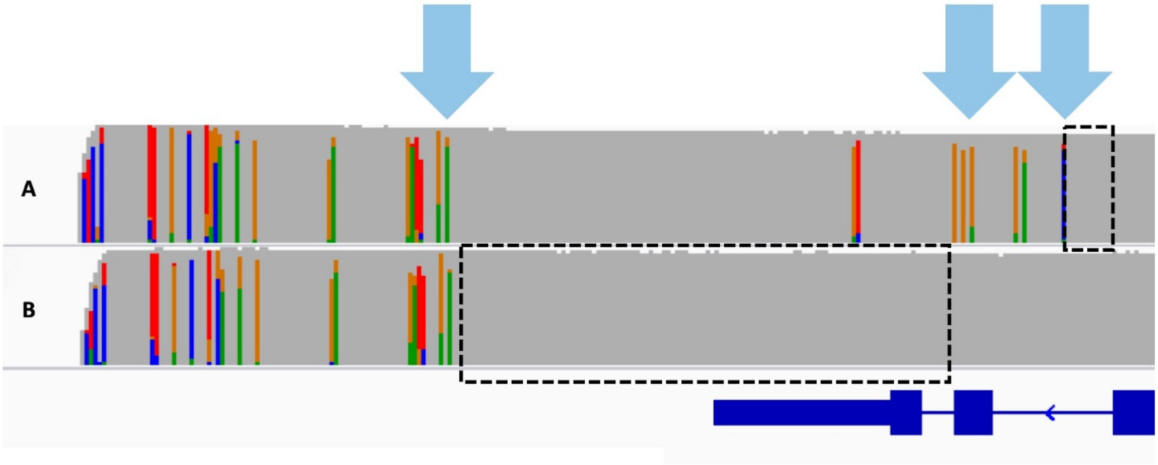


Figure 53: Visual determination of the breakpoint of fusion recombinants within the GBA gene. The light blue arrows highlight the sentinel positions, the top and middle rows show the coverage tracks of two samples carrying a fusion recombinant allele, with colours indicating positions where the alignment differs from the GBA reference. The bottom row shows the 3' end of GBA. The dotted boxes highlight the regions were the breakpoint is. A) Pathogenic fusion recombinant. The breakpoint lies in intron 9 of GBA (dotted box) and the recombination introduces SNVs in the

resulting allele. B) Non-pathogenic recombinant. The breakpoint can be anywhere between intron 10 and the region downstream of t e 3'UTR of GBA. However, since these regions of GBA and GBAP are 100% homologous, no SNVs are introduced in the resulting allele.

4.2.4. PCR free enrichment of genomic DNA for *GBA* locus – UNCALLED

4.2.4.1. Library preparation

Genomic DNA was prepared according to the protocol provided by Nanopore (protocol version GDE_9063_v109_revT_14Aug2019). The procedure is identical to the one described in chapter 3.2.5. I used 1500ng of input genomic DNA. The first flow-cell run with UNCALLED had a length of 10 hours and was used as a test, the following runs had a length of 48 hours. When the number of sequencing pores on a flow-cell dropped below 5% of the total before the end of the 48 hours, the flow-cell was washed with kit EXP-WSH003 and reloaded with the same library following the manufacturer's protocol (protocol version WFC_9088_v1_revF_18Sep2019).

4.2.4.2. UNCALLED

UNCALLED (version 2.2) was downloaded from https://github.com/skovaka/UNCALLED. The reference sequence, a region straddling *GBA*, *GBAP1* and a flanking region of 20Kb on both sides (hg38.chr1:155193567-155264811), was downloaded from the University of Santa Cruz (UCSC) Genomic Institute website (https://genome.ucsc.edu/). To improve performance, the reference needed to be masked for repeatitive short DNA sequences[170]. I tested different approaches to mask the reference, including the scripts *mask_internal.sh* and *mask_external.sh* (provided by the developers of UNCALLED at https://github.com/skovaka/UNCALLED/tree/master/masking) as well as a "hard-masked" version of the reference from UCSC (details in the Results section). The UNCALLED command Pafstats and reads from previous ONT runs were employed to compare different masking methods (see Results section 4.3.3.1). Chunk size was set at 3. Analysis of depth of coverage was carried out with Samtools depth. Reads were aligned to the human genome with LAST and NGMLR. For SV calling, I tested Sniffles, CuteSV and NanoSV.

A full list of commands for UNCALLED is provided in Appendix – Code.

4.2.5. Digital PCR to confirm presence and copy number of reciprocal recombinants

To confirm the presence and copy number of the reciprocal recombinants detected, a subset of samples carrying such variants were sent to Qiagen labs in Hilden (QIAGEN GMBH, APPLICATION LAB-MV2 NEU – INNOVATION STRASSE 2-40724 HILDEN, GERMANY) to carry out digital PCR (dPCR). Three targets were selected for analysis, two predicted to straddle the recombinant region (DCH101-0776005A and DCH101-0776012A) and one predicted to lie outside of the recombinant region and used as a reference (DCH101-1260927A). The primers for the dPCR were available in commercial kits provided by Qiagen. The positions of the primers used for analysis are reported in Table 29 and Figure 54.

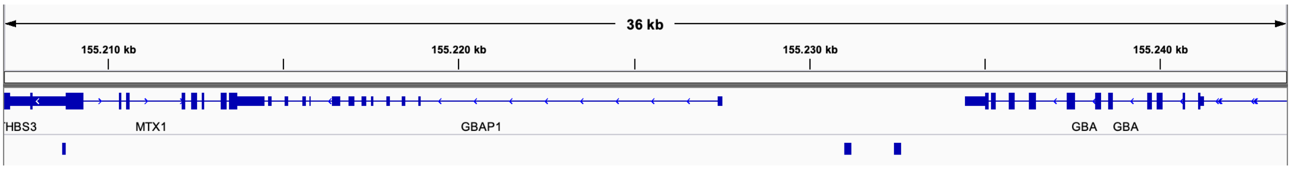


Figure 54: Regions covered by primers used for dPCR.

Kit	Region covered (hg38)	Genomic region	Predicted to be affected by recombination
DCH101-0776005A	155231010-55231209	*MTX1P1*	Yes
DCH101-0776012A	155232410-55232609	*MTX1P1*	Yes
DCH101-1260927A	155208699-55208804	*MTX1*	No

Table 29: primers used for confirmatory dPCR.

4.2.1. Cross-validation of results with novel Illumina caller for GBA

The team of dr Michael Eberle, at Illumina, is developing a novel caller for *GBA* using short read WGS data, called Gauchian[165]. A detailed explanation of the functioning of Gauchian is beyond the scope of this thesis. In brief, Gauchian overcomes the limitations of short read WGS sequencing of *GBA*, due to the high homology with the pseudogene, by calculating the copy number of each variant within the homology region. For example, p.L483P is natively present in *GBAP1* but not in *GBA*, so the expected copy number of p.L483P in a wild type individual is 2. A copy number of 3 indicates that this individual is heterozygous for p.L483P, as they must carry one of these in the *GBA* gene. Similarly, Gauchian can detect larger copy number gains and losses, corresponding to reciprocal and non-reciprocal recombinants.

I used the novel ONT pipeline described in this chapter to validate Gauchian, also obtaining further validation of my method. Results were also compared with WGS Illumina calls from BWA-GATK[177].

4.3. Results

4.3.1. Recombinants detected with PCR approach

In the RAPSODI cohort, 403 samples were tested with the fusion PCR and 414 with the duplication PCR. Additionally, 28 samples were obtained from the QSBB, including 6 samples with Lewy body dementia (LBD) and 22 with PD. Fourteen samples from non-affected controls were obtained from NHGRI and 22 samples from PPMI.

In total, 13 fusion alleles and 19 duplication alleles were detected. Of these, 9 fusion and 4 duplication recombinants were identified in the RAPSODI cohort, 2 fusion and 11 duplication among the NHGRI samples, 1 fusion and 1 duplication in the PPMI samples and 1 fusion and 3 duplications in the QSBB samples. Since the Coriell and

PPMI samples were selected *ad hoc* (see section 4.2.1), only the RAPSODI cohort and the QSBB samples was used to estimate prevalence of these variants. In the RAPSODI cohort, carriers' prevalence was 2.23% (AF 1.12%) for fusion recombinants and 0.96% for duplication recombinants. In the QSBB, prevalence was 3.57% for fusion recombinants and 10.71% for duplication recombinants.

4.3.1. Prediction of breakpoints of reciprocal recombinants

4.3.1.1. Testing different aligners and structural variant callers

Alignment of the product of PCR with primer pairs 2 and 3 is intrinsically problematic, as the amplicon contains merged sequence from *GBA* and *GBAP1* and does not have a correspondence in the reference human genome (Figure 51). I tested three different aligners (NGMLR, Minimap2 and LAST) and three different SV callers (Sniffles, CuteSV and NanoSV) to find a combination capable of calling the breakpoints of the fusion recombinants, in order to determine whether they are "pathogenic" or not.

Some variant callers are optimised for a specific aligner. In particular, Sniffles and CuteSV work better with data aligned with NGMLR, whereas NanoSV performs best with LAST.

Results of testing different aligners is displayed in Figure 55. With NGMLR, different parameters were tested, without dramatic changes in the resulting alignment. The most striking difference is that Minimap2 tends to align most of the reads to *GBA*, while LAST splits the reads between *GBA* and *GBAP1*. NGMLR is somewhat in between these two extremes.

When testing the different SV callers, the best combination was LAST-NanoSV, which called a deletion in all the fusion samples. The breakpoints predicted by LAST-NanoSV fell within the right region (as predicted with sentinel positions – see section 4.2.3.1) in 10 of the 13 cases (77%). Both Sniffles and CuteSV produced suboptimal results, calling multiple SVs in each sample. When considering only the call with the highest quality score, Sniffles called a variant with breakpoints falling in the right region in 6 of 13 samples (46%) and CuteSV in 3 of 13 samples (23%). In all cases, results were not ideal, and the best approach remained using the sentinel positions (as detailed in section 4.2.3.1).

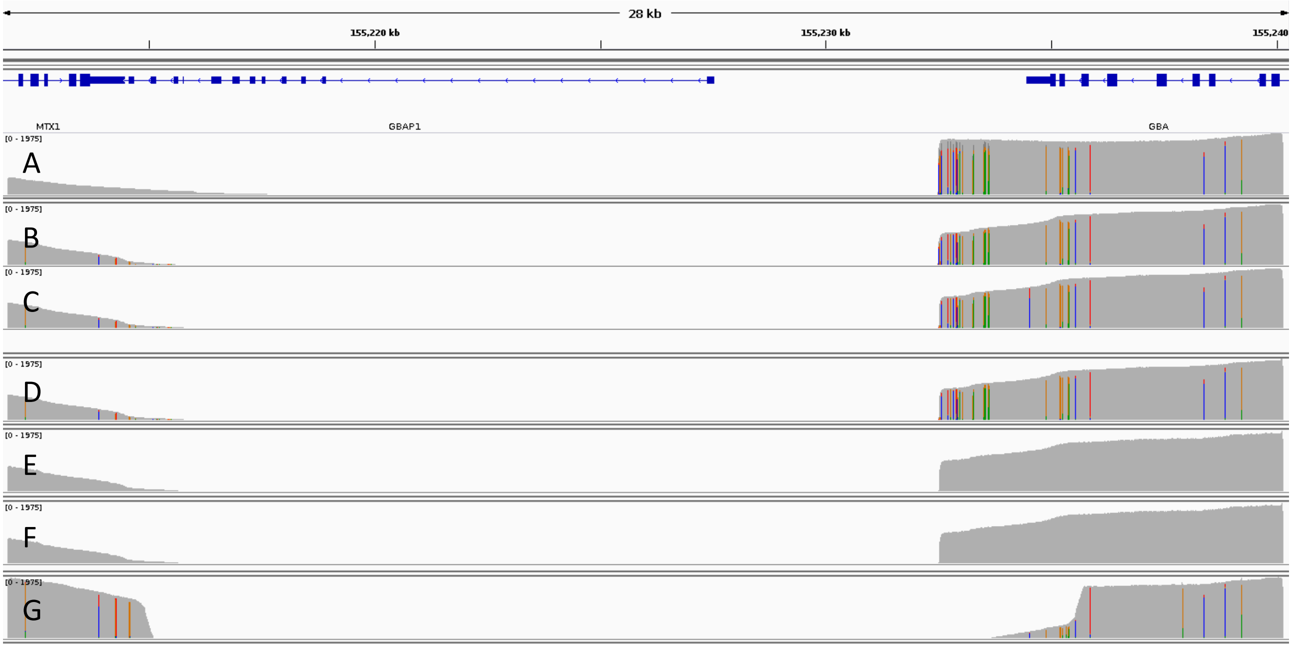


Figure 55: Reciprocal fusion recombinant, test with different aligners. Coordinates within Chromosome one and GBA and GBAP1 positions are reported at the top. The grey areas represent the proportion of reads aligned to GBA (to the right) and GBAP1 (to the left). As seen, LAST aligns more to GBAP1 compared to the other aligners. Images exported from IGV.

A) Minimap2.

B-F) NGMLR with different parameters. Default parameters (B), mismatch -7 (C), mismatch -7 -nolowcoverage (D), mismatch -14 -nolowcoverage (E), mismatch -20 -nolowcoverage (F).

G) LAST.

4.3.1.2. Sentinel positions

Using the sentinel positions as described in section 4.2.3, I was able to differentiate between fusion alleles altering the coding sequence of *GBA* (“pathogenic” recombinants) and those not altering the coding region (“non-pathogenic” recombinants). Of the 13 fusions detected, 7 were found to be non-pathogenic and 6 to be pathogenic. One of the non-pathogenic fusion alleles also carried p.L483P and an intronic variant in intron 9 (chr1:g.155235412G>A), suggesting a small non-reciprocal event alongside the larger reciprocal event coexisted in the same allele (Figure 56).

I applied the same approach to duplication recombinants, and the results showed that in all samples the duplicated region was all originating from *GBAP1* and did not include *GBA* sequence (Figure 57).

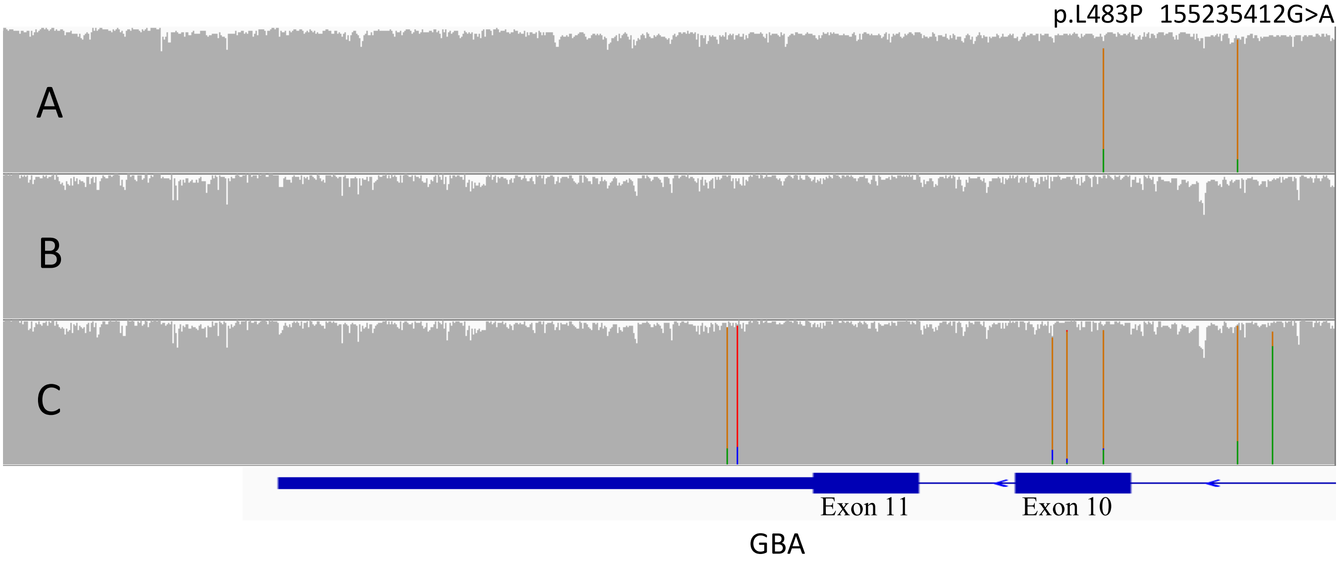


Figure 56: Three examples of reciprocal fusion recombinants. The coloured bars within the grey areas represent single nucleotide variants (p.L483P and chr1:155235412G>A are marked at the top). The corresponding position within the GBA gene is reported at the bottom.

A) non-pathogenic fusion allele which also carried the p.L483P and chr1:g.155235412G>A variants.

B) non-pathogenic fusion allele.

C) Pathogenic fusion allele, containing the 3 RecNciI SNVs and other intronic SNVs.

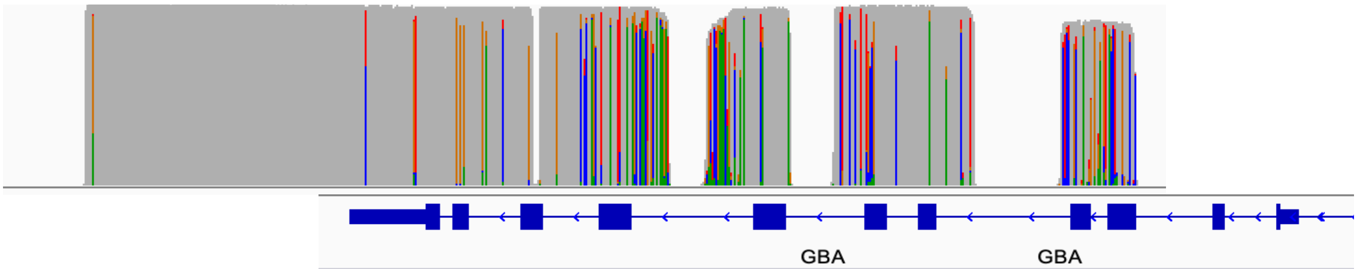


Figure 57: Reciprocal duplication recombinant, aligned to hg38 after masking GBAP1. The grey area represent the depth of coverage of the transcript across the GBA gene (showed at the bottom). Most of the duplicated material derives from GBAP1 (as shown by the coloured bars, representing SNVs, and by the large deletions, both part of GBAP1) and is merged with a region downstream of GBA (also see Figure 51 for reference). Image exported from IGV.

4.3.2. Phenotype and other *GBA* variants in carriers of non-pathogenic reciprocal recombinants

Among carriers of duplication recombinants, 3 also carried a p.L483P SNV, and 2 also carried the recTL+delta55 non-reciprocal recombinant. Of the non-pathogenic fusion carriers, 2 also carried the synonimous variant p.P491= and 2 carried p.L483P, of which one within a small non-reciprocal recombinant (Figure 56). Phenotype of non-pathogenic fusion and duplication recombinants in the RAPSODI cohort are reported in Table 30. In the RAPSODI cohort, the only one with a dedicated control group, neither duplication or non-pathogenic fusion recombinants were more frequent in PD or GD patients than in controls (fisher exact test p-values 0.41 and 0.19, respectively).

Sample	Structural variant	Other Variants	Pattern of recombinant (pathogenic fusions only)	Phenotype	Age at diagnosis (PD only)	Non motor symptoms (PD only)
NHGRI - 1	Duplication			Control		
NHGRI - 2	Duplication			Control		
NHGRI - 3	Duplication			Control		
NHGRI - 4	Duplication			Control		
NHGRI - 5	Duplication			Control		
NHGRI - 6	Duplication			Control		
NHGRI - 7	Duplication			Control		
NHGRI - 8	Duplication			Control		
NHGRI - 9	Duplication			Control		
NHGRI - 10	Duplication			Control		
PPMI - 1	Duplication	L483P		PD	60	RBD
QSBB - 1	Duplication			PD	63	Not available
QSBB - 2	Duplication			PD	67	Not available
QSBB - 3	Duplication			PD	64	Not available
RAPSODI - 1	Duplication			PD	58	Hyposmia, RBD, constipation
RAPSODI - 2	Duplication	L483P		PD	72	Depression
RAPSODI - 3	Duplication	L483P		GD relative		
RAPSODI - 4	Duplication			PD	59	Depression, Hyposmia, constipation
RAPSODI 5	Duplication			Control		
NHGRI - 11	Non-pathogenic fusion	p.pro491=		Control		
NHGRI - 12	Non-pathogenic fusion	p.pro491=		Control		
PPMI - 2	Non-pathogenic fusion			PD	63	Hyposmia, constipation
QSBB - 4	Non-pathogenic fusion	p.L483P		PD	61	Not available
RAPSODI - 6	Non-pathogenic fusion			GD relative		

RAPSODI - 7	Non-pathogenic fusion	p.L483P		GD		
RAPSODI - 8	Non-pathogenic fusion			PD	65	Hyposmia, constipation
NHGRI - 13	Pathogenic fusion		RecNcil	Control		
RAPSODI – 9	Pathogenic fusion		RecNcil	GD		
RAPSODI – 10	Pathogenic fusion		RecNcil	GD		
RAPSODI – 11	Pathogenic fusion		RecNcil	GD		
RAPSODI – 12	Pathogenic fusion		RecNcil	GD relative		
RAPSODI - 13	Pathogenic fusion		RecNcil	PD	56	Hyposmia, RBD

Table 30: Phenotype and other variants in carriers of non-pathogenic reciprocal recombinants.

PD: Parkinson disease, GD: Gaucher Disease

4.3.3. UNCALLED

In total, four samples were successfully sequenced with UNCALLED, as reported in Table 31. These samples were selected because a reciprocal recombinant (fusion in RAP3530 and hg03428 and duplication in hg03547 and hg03895) was detected with the PCR approach described in section 4.2.2. The sequencing yield and number of flow-cell wash and re-load cycles for each sample is summarised in Table 31.

Sample	Total duration of the run (hours)	Number of wash-load cycles	Total reads sequenced (millions)	Total bases sequenced (Gb)
RAP3530	11	1	1.74	4.48
hg03547	66.5	3	6.80	10.4
hg03895	66.3	3	4.23	6.72
hg03428	65.7	2	8.58	13.06

Table 31: Metrics of the four UNCALLED runs.

4.3.3.1. Masking the reference with UCLC browser improved performance

The *GBA* locus contains many simple repetitive regions as well as short interspersed nuclear elements (SINE) and long interspersed nuclear elements (LINE) as shown in

Figure 58. Since repetitive elements can decrease UNCALLED performance, masking the reference was required[170]. This process consists in replacing highly repetitive sequences with Ns, increase the speed of real time analysis and accuracy of UNCALLED. The developers of UNCALLED created two scripts called mask_internal.sh and mask_external.sh for masking the reference. The former iteratively masks high frequency k-mers within the reference and the latter masks sequences in the reference that are repeated in other regions in the human genome. I also tested a masked version of the reference downloaded from UCSC genome browser. The best iteration of mask_internal.sh improved signal processing speed, while mask_external.sh did not improve performance. However, the best performance was obtained with the reference masked by UCSC. Comparison of different masking methods is reported in Table 32.

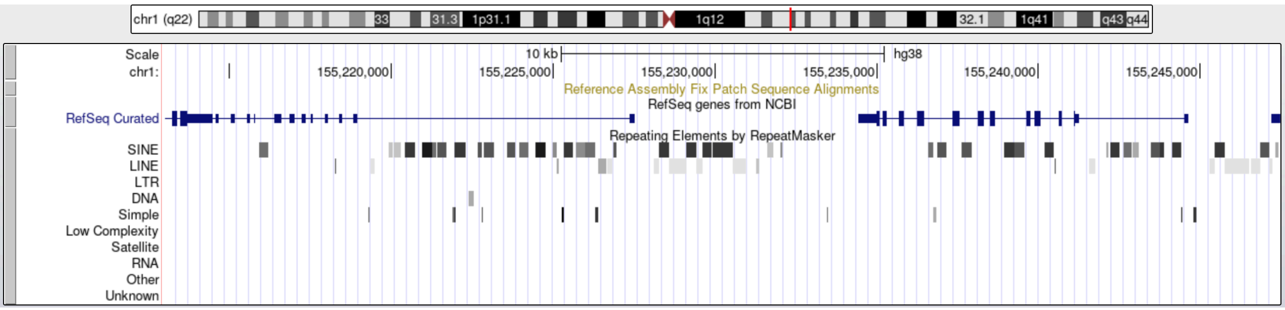


Figure 58: Repetitive regions in the GBA gene and pseudogene from UCSC genome browser (https://genome-euro.ucsc.edu/).

	Unmasked	Internal masking (100th iteration)	Internal masking (100th iteration) + external masking	UCSC masking
Mean base-pairs/second	5699.13	5983.06	5931.10	6568.84

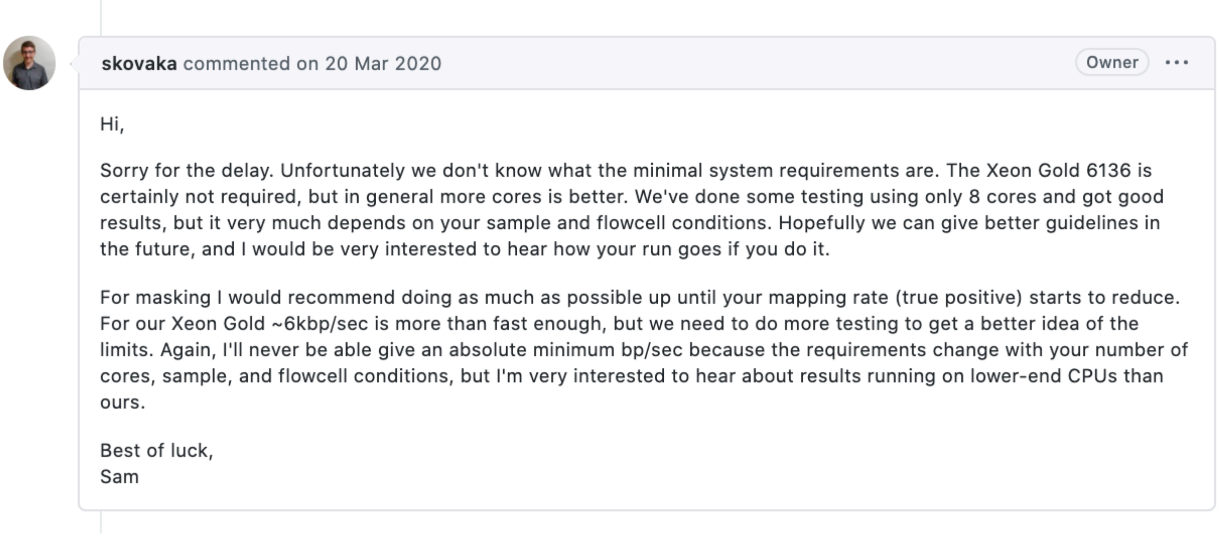
Table 32: Comparison of different methods for masking the reference for UNCALLED. Speed is represented as the Mean base-pairs that UNCALLED was able to analyse in a second on a simulation.

4.3.3.1. Evidence of enrichment with UNCALLED

UNCALLED has been designed for use on high tier computers with powerful Graphics processing units (GPUs)[170], which I did not have access to (Figure 59). For this reason, I sought evidence that UNCALLED was producing real-time enrichment of the *GBA* locus in the sequencing runs.

When comparing the mean depth of coverage over the whole genome and over the region of interest (hg38.chr1:155193567-155264811) enrichment was 2.1-5.7x. When considering only reads longer than 3Kb, enrichment increased to 5.8-47.9x (Table 33). This is consistent with the concept that reads aligning to off-target regions will be ejected prematurely and only a short part will be sequenced (Figure 61).

Additional proof of the correct functioning of UNCALLED was provided by the drop in N50 showed by MinKNOW during sequencing from 8.29Kb to 2.85Kb after starting UNCALLED (Figure 60). The N50 is a measurement of the length of the reads sequenced, so I attributed this change to the premature ejection of reads that were not on target by uncalled.


skovaka commented on 20 Mar 2020

Hi,

Sorry for the delay. Unfortunately we don't know what the minimal system requirements are. The Xeon Gold 6136 is certainly not required, but in general more cores is better. We've done some testing using only 8 cores and got good results, but it very much depends on your sample and flowcell conditions. Hopefully we can give better guidelines in the future, and I would be very interested to hear how your run goes if you do it.

For masking I would recommend doing as much as possible up until your mapping rate (true positive) starts to reduce. For our Xeon Gold ~6kbp/sec is more than fast enough, but we need to do more testing to get a better idea of the limits. Again, I'll never be able give an absolute minimum bp/sec because the requirements change with your number of cores, sample, and flowcell conditions, but I'm very interested to hear about results running on lower-end CPUs than ours.

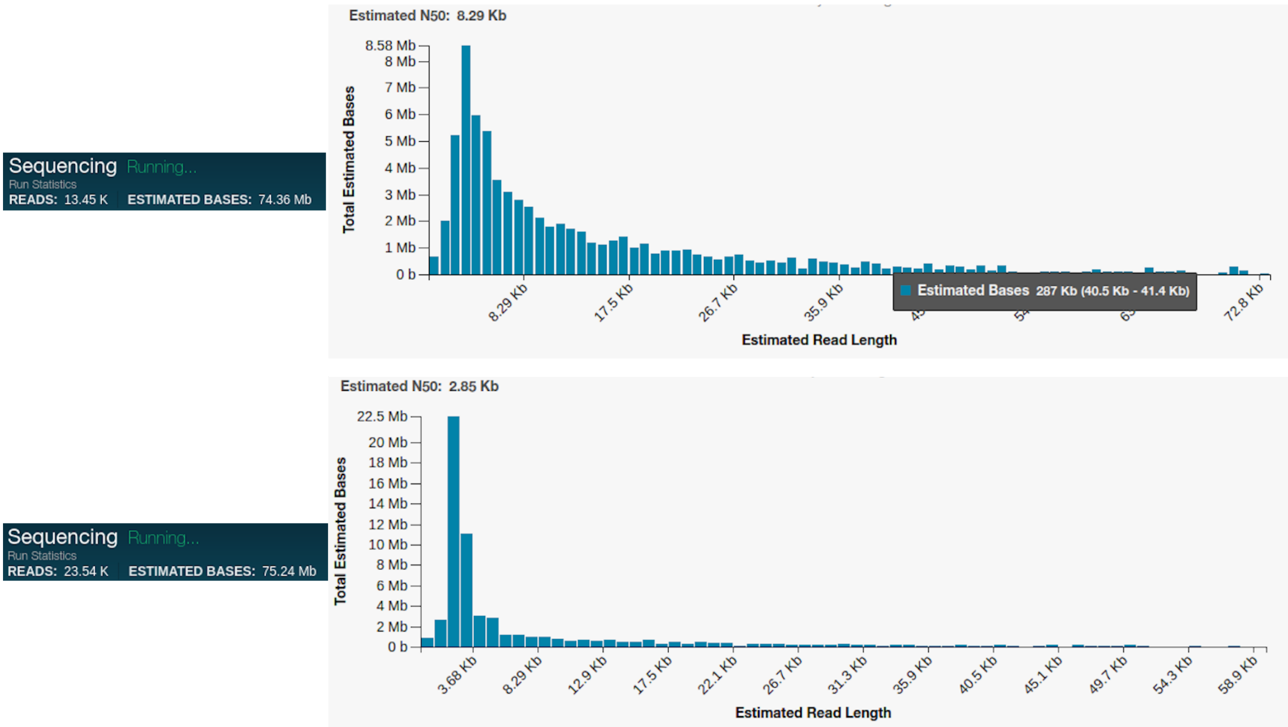
Best of luck,
Sam

Figure 59: The developer of UNCALLED talks about the system requirements for the software (from the Github page of UNCALLED).

Sample	Mean depth all reads – whole genome	Mean depth all reads - ROI	Estimated enrichment All reads	Mean depth reads > 3Kb – whole genome	Mean depth reads > 3Kb – ROI	Estimated enrichment reads > 3Kb
RAP3530	1.18715	3.45098	2.9x	0.44736	3.09001	6.9x
hg03547	2.62377	14.4584	5.5x	0.21519	10.3055	47.9x

hg03895	1.81837	10.3858	5.7x	0.28421	8.70733	30.6x
hg03428	3.45944	7.10502	2.1x	0.80779	4.68923	5.8x

Table 33: Enrichment with UNCALLED. ROI: Region Of Interest (hg38.chr1:155193567-155264811)



Sequencing Running...
Run Statistics
READS: 13.45 K ESTIMATED BASES: 74.36 Mb

Sequencing Running...
Run Statistics
READS: 23.54 K ESTIMATED BASES: 75.24 Mb

Figure 60: Evidence of enrichment with UNCALLED. N50 is 8.29 without UNCALLED and drops to 2.85 after UNCALLED is initiated.

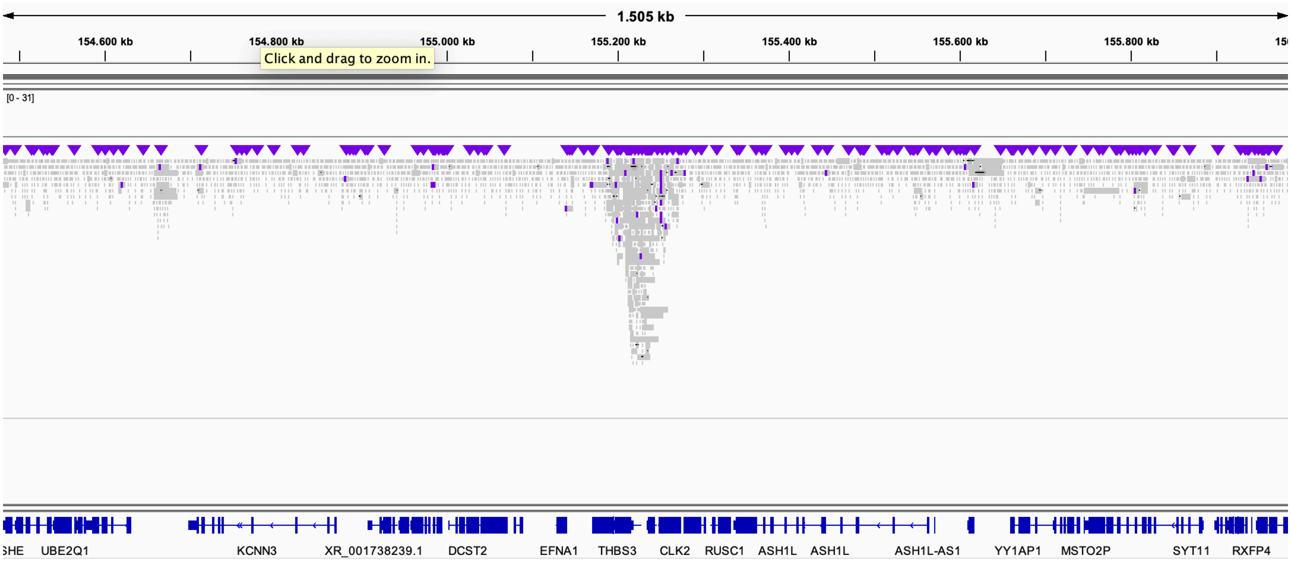


Figure 61: Enrichment with UNCALLED. The gray areas represent single reads, aligned to chr1 (coordinates reported at the top). The GBA region has higher coverage and longer reads compared to nearby regions. Image exported from IGV.

4.3.3.2. Structural variants calling

Upon visualising the LAST alignment on IGV, I was able to confirm the presence of all 4 structural variants. For the two fusion recombinants, the drop of coverage, suggesting the location of the breakpoint, was consistent with the prediction with sentinel positions (Figure 62). NanoSV on LAST alignment was able to correctly call the SV in the two fusion recombinants. It was also able to call multiple SVs in one of the two duplication recombinants, with extensively different breakpoints, and did not call any variants in the other. Sniffles and CuteSV on NGMLR alignment did not detect the correct SV in any of the 4 samples (Figure 63).

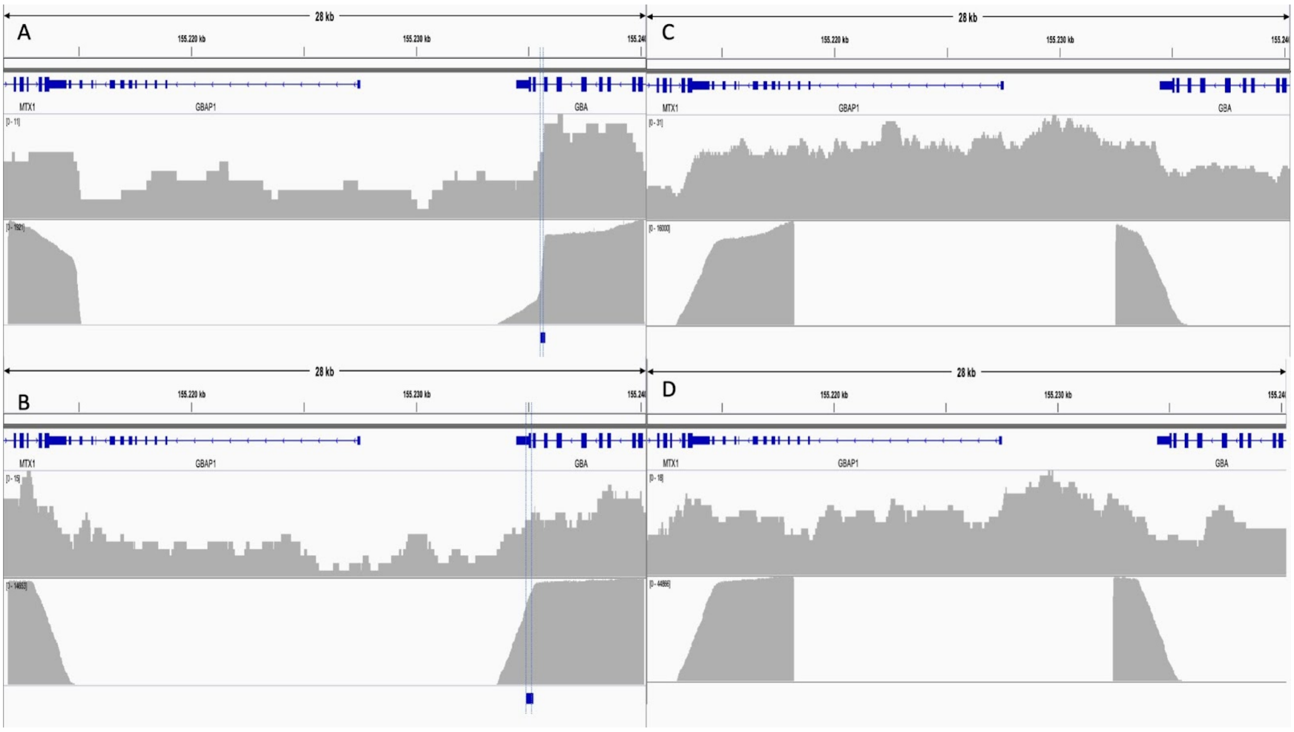


Figure 62: Reciprocal recombinants confirmed by adaptive sampling with UNCALLED.

The grey areas represent the depth of coverage at each position within Chromosome 1, coordinates and relative position of GBA and GBAP1 reported at the top of each panel.

Each panel shows the data produced with UNCALLED (top row of each panel) and the amplicons from PCRs with primer pairs 2 and 3 (bottom row of each panel). For the fusion recombinants (A and B), the position of the breakpoint predicted with sentinel position is provided (blue bars and dotted lines). A) "Pathogenic" fusion recombinant. B) Non-pathogenic fusion recombinant. C) Duplication recombinant (copy number 5). D) Duplication recombinant (copy number 3).

All alignment was carried out with LAST. Images exported from IGV.

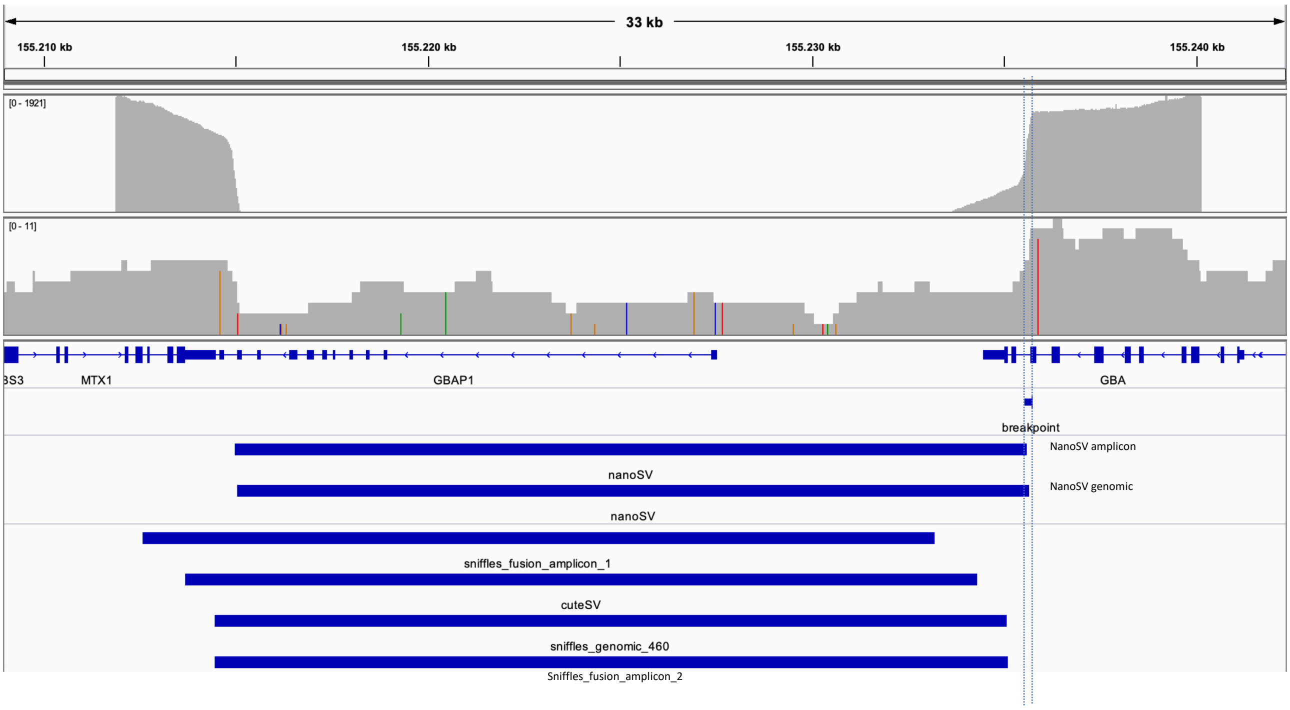


Figure 63: Comparison between different SV callers. The sample displayed is a reciprocal fusion recombinant. The top row shows the amplicon produced with primer pair 2. The mid row represents adaptive sampling with UNCALLED. The blue bars show the calls on amplicon and UNCALLED reads with three different SV callers: NanoSV, CuteSV and Sniffles. The dotted lines show the position of the breakpoint, detected with sentinel positions.

4.3.4. Confirmation of copy number in reciprocal recombinants with dPCR

Six samples were sent to Qiagen for dPCR: four carrying a reciprocal gene duplication recombinant, one with a reciprocal fusion recombinant and one without any recombinants detected. For all samples, the copy number detected across the regions predicted to be within the recombination event was consistent with the type of recombinant previously detected. Of interest, the 4 gene duplication recombinant carriers showed a copy number of 3, 5, 7 and 8, respectively. On the other hand, the sample carrying the fusion recombinant showed a copy number of 1 and the sample not carrying any recombinant showed a copy number of 2, as expected. Results of dPCR analysis are reported in Table 34.

		dPCR copy number*	
Sample	Recombinant detected with PCR and ONT approach	DCH101-0776005A	DCH101-0776012A
NA20756	Duplication	2.98	3.37
HG01912	Duplication	4.94	5.36
HG01889	Duplication	6.64	6.78
HG02284	Duplication	7.59	8.12
RAP4036	None	2.06	2.29
RAP3530	Fusion	0.92	1.04

Table 34: Results of dPCR.

**results are normalised by copy number detected with kit DCH101-1260927A, which straddles a region outside of the recombination event.*

4.3.5. Cross-platform validation of results with Novel Illumina caller for *GBA*

To validate Gauchian, and obtain further validation of the ONT pipeline described in this chapter, 37 samples were sequenced with Gauchian, based on pre-existing Illumina WGS data obtained from the AMP-PD Knowledge portal. Additionally, 5 brain samples from QSBB were whole genome sequenced with Illumina and then analysed with Gauchian.

The samples included in the cross-validation included a selection of fusion and duplication recombinants, SNVs, wild type alleles and samples where Gauchian and BWA-GATK were not concordant (Table 35). For all 42, Gauchian and the pipeline described in this chapter produced fully concordant results.

Variant	Number of samples	Comments
Non-pathogenic fusion recombinants	4	Including 1 also carrying p.L483P
Pathogenic fusion recombinants	1	
Duplication recombinants	14	Including 2 also carrying a gene conversion and 1 with p.L483P
SNVs	9	Including 2 where BWA-GATK missed p.L483P
Gene conversions (RecTL + delta55)	2	
wild type alleles	12	Including 2 where BWA-GATK wrongly called p.A495P

Table 35: Details of samples used for cross-validation of Gauchian.

4.4. Discussion

The *GBA* locus harbours frequent and complex recombination events, due to the presence of *GBAP1*[178]. Resolving these recombinants is challenging with both short and long read sequencing, and in most cohorts they are missed completely[109]. The ONT method I helped develop and validate, using primer pair A for PCR as described in chapter 3, is able to detect non-reciprocal recombinants, but misses reciprocal events. In this chapter, I described an inexpensive and relatively easy way to detect reciprocal recombinants, only requiring two additional PCRs with primer pairs B and C. The ability to detect reciprocal recombinants is essential for pathogenic fusion recombinants, that cause the severe, neuronopathic forms of GD and are strong risk factors for sporadic PD[107], and for this reason the method described in this chapter greatly improves the ONT pipeline for sequencing *GBA*. Moreover, this method is capable of detecting reciprocal recombinants not affecting the coding region of *GBA*, including “non-pathogenic” fusion alleles and all duplication alleles. These recombinants have an uncertain clinical and biological significance. While they leave the *GBA* gene intact, they cause substantial deletion or multiplication of *GBAP1* and nearby sequence, with a number of potential effects. GBAP1 contains multiple microRNA (miRNA) binding sites. These small molecules can alter expression of various genes, and an interesting hypothesis is that recombination events that multiply or delete these sites might alter expression of *GBA*[179].

Moreover, I was able to detect “non-pathogenic” reciprocal recombinants also harbouring SNVs and even small non-reciprocal recombinants. This is not unexpected, and was previously described[115], but nonetheless stresses the complexity of structural variants in the *GBA* locus. Given the relatively high frequency of this association, one interesting question is whether “non-pathogenic” reciprocal recombinants alter the risk of disease in *GBA* variant carriers.

This is an exciting new perspective, that has the potential to help us understand the incomplete penetrance and variable phenotype of PD cases associated with *GBA* variants.

In our cohort, the sample size was too low to draw any conclusion on potential pathogenicity of these variants based on their prevalence in the different groups. A direction for future research would be to apply our ONT method to detect reciprocal recombinants in larger cohorts and clarify their role in disease.

Our method was applied to validate Gauchian, a novel caller for *GBA* based on Illumina short read WGS data[165]. The validation process showed fully concordant results between ONT and Gauchian in 42 samples carrying different *GBA* variants, including reciprocal and non-reciprocal recombinants. There are many repositories of WGS data from people with Parkinson, like the AMP-PD consortium, featuring a cohort of over 3000 PD cases and controls. These resources are invaluable to understand the genetic background of PD, but to date the calls of *GBA* variants were unreliable due to misalignment to *GBAP1*[109]. Once publicly available, Gauchian will enable reliable analysis of the prevalence of *GBA* variants in these repositories, including the "non-pathogenic" reciprocal recombinants.

Importantly, both Gauchian and ONT were able to correctly genotype variants that were previously miscalled by BWA-GATK, one of the most widely used pipelines for Illumina data[177]. This suggests that both methods are an improvement of the existing technology.

Adaptive sampling is a way of obtaining enrichment of regions of interest with ONT[170]. One of the main advantages of adaptive sampling is that it allows the visualisation of native DNA molecules, without the need of amplification with PCR or CAS9, which can cause artefacts or miss certain variants (e.g. reciprocal recombinants with primer pair A). UNCALLED, the adaptive sampling software I used, requires high computing power to select DNA molecules in real time, so its functioning on smaller computers was not granted. Moreover, while now UNCALLED is integrated in newer ONT devices, it is still not fully supported for the smaller ONT device, the MinION. It is hard to estimate the exact extent of the enrichment I was able to achieve, as the samples analysed all had copy number gains or losses within the region of interest, which cause coverage drops or raises even without adaptive sampling. Nonetheless, I was able to provide evidence that UNCALLED can be used reliably with a MinION on an affordable computer, with only 16 GB of RAM and without a GPU.

With UNCALLED, I provided further validation of the ONT method, allowing “real life” visualisation of the reciprocal recombinants as drops and peaks of coverage. With further refinement of the method, it may be possible to visualise the exact breakpoints of the recombinants and estimate copy number gain of duplications.

The main limitation of this PCR approach for detecting reciprocal recombinants in *GBA* is the somewhat unreliable nature of the PCR: when the PCR reaction with primer pairs 2 and 3 does not produce an amplicon, this can be due to the absence of reciprocal recombinants, or to failure of the PCR itself. This is a risk especially with DNA extracted from saliva, which often is of suboptimal quality. Thorough quality control of the DNA prior to PCR, consistent use of multiple positive and negative controls and carrying out repeats of each PCR reaction are all important steps to avoid missing reciprocal recombinants due to PCR failure. It is reassuring to note that the prevalence of reciprocal recombinants we detected in the RAPSODI cohort is consistent with that observed with Gauchian in the much bigger AMP-PD cohort (data not reported in this thesis)[165].

Another intrinsic limitation of this method is the inability to phase specific reciprocal variants. While for fusion recombinants the resulting amplicon contains the residual *GBA* sequence in that allele and thus allows phasing with other variants in *GBA*, in duplication recombinants only the duplicated region is amplified, so phasing with *GBA* is not possible (Figure 51).

All in all, the method presented here, in combination with the main pipeline presented in chapter 3, allows detection of all classes of variants in the *GBA* gene. When targeted GBA sequencing is sufficient, ONT long read sequencing is a valid and reliable option. To date, except for Gauchian, no other sequencing methods are able to fully resolve *GBA*[109].

5. Analysis of deep intronic variants in *GBA* and their effect on the phenotype of Parkinson disease

Content of this chapter has been adapted from a recent publication I authored[125], with permission of the publisher.

5.1. Overview and Rationale

One of the advantages of long read sequencing of *GBA* is that it provides information about intronic regions of the gene. The penetrance variability of *GBA* variants remains largely unexplained and in is possible that other genetic factors, including DIVs, can affect the risk of PD.

DIVs (i.e. intronic variants that are located at distance from exons) can cause disease through different mechanisms, including the creation of donor or acceptor splicing sites and the formation of novel exons or loss of existing exons[180,181], and affecting miRNA binding sites[120]. Indeed, many disease have been linked to DIVs[182–184], inlcuding *GBA*[119].

Two main intronic haplotypes have been described in *GBA* for the first time in 1990[122]. These haplotypes are characterised by at least 3 intronic variants in *GBA* (rs9628662, rs762488 and rs2009578)[122,123], with the reference haplotype (featuring the reference nucleotide at the 3 positions) and the alternate allele (featuring the alternate nucleotide at the 3 positions) having a frequency of roughly 70% and 30%, respectively[122,123].

It is unknown whether these haplotypes affect the risk of PD.

Recently, a study correlated these two haplotypes with the age at diagnosis and onset of PD showing that the reference/major haplotype was associated with a earlier age at onset and diagnosis of PD[124]. In this section of my PhD, I tried to replicate these findings in the RAPSODI cohort. To seek further validation, I expanded the analysis to the Accelerating Medicines Partnership for Parkinson’s disease initiative (AMP-PD) cohort.

5.2.Methods

5.2.1. Participants

I included participants to the RAPSODI study with a diagnosis of PD that did not carry any coding variants in *GBA* (details on recruitment in chapter 2).

I identified additional participants from the AMP-PD database (https://amp-pd.org). AMP-PD participants were included if they had a diagnosis of “Idiopathic PD” or “Parkinson’s Disease” and if they did not carry any coding *GBA* variant. AMP-PD data were downloaded on the 20/07/2020 (version 2019_v1release_1015).

5.2.2. Definition of the haplotypes

For RAPSODI participants, full *GBA* sequencing data were obtained as described in chapter 3.

AMP-PD included short read WGS data, from which I extracted *GBA* sequences.

I defined haplotype A by the presence of the alternate genotype, and haplotype B by the reference genotype at the 3 intronic variants rs9628662, rs762488, and rs2009578.

Participants that carried at least 1 allele which did not fall into this classification, or for which quality of the alignment at any of the 3 positions was not good enough for confident calling, were excluded from the analysis.

R version 4.0.2 and the R package vcfR version 1.12.0 were used to analyse the databases and assign each allele to one of two haplotypes.

5.2.3. Statistical analysis

I used R version 4.0.2 for statistical analysis. I investigated two different models of effect of the haplotypes: an additive model and a dominant (haplotype B) model. For the additive model, the number of alleles carrying haplotype B was the dependent variable and linear regression was used to assess an effect on age at diagnosis of symptoms.

For the dominant effect of haplotype B model, participants were divided into two groups, one with homozygotes for haplotype A and one with heterozygotes and homozygotes for haplotype B and ANOVA was used for the analysis. The same analysis was repeated on the full cohorts and only on patients with an age at diagnosis above 50.

5.3. Results

5.3.1. Sample size and haplotypes

In total, I analysed data from 1417 PD patients (100 from RAPSODI and 1317 from AMP-PD). Of them, 141 were homozygous for haplotype A, 573 were heterozygous and 693 were homozygous for haplotype B. For 10 patients, it was impossible to determine the haplotype, because the sequencing quality was poor or incomplete, or because their genotype at the 3 positions defining the haplotypes were not all wild type or all mutated. The allelic frequencies were 0.302 for haplotype A and 0.691 for haplotype B in the joint cohorts. When considering each cohort separately, allelic frequencies were 0.265 / 0.685 in RAPSODI and 0.305 / 0.692 in AMP-PD. Numbers of participants with each haplotype are reported in Table 36.

A great number of rarer intronic variants were identified, determining over 50 unique haplotypes (pictured in Figure 64).

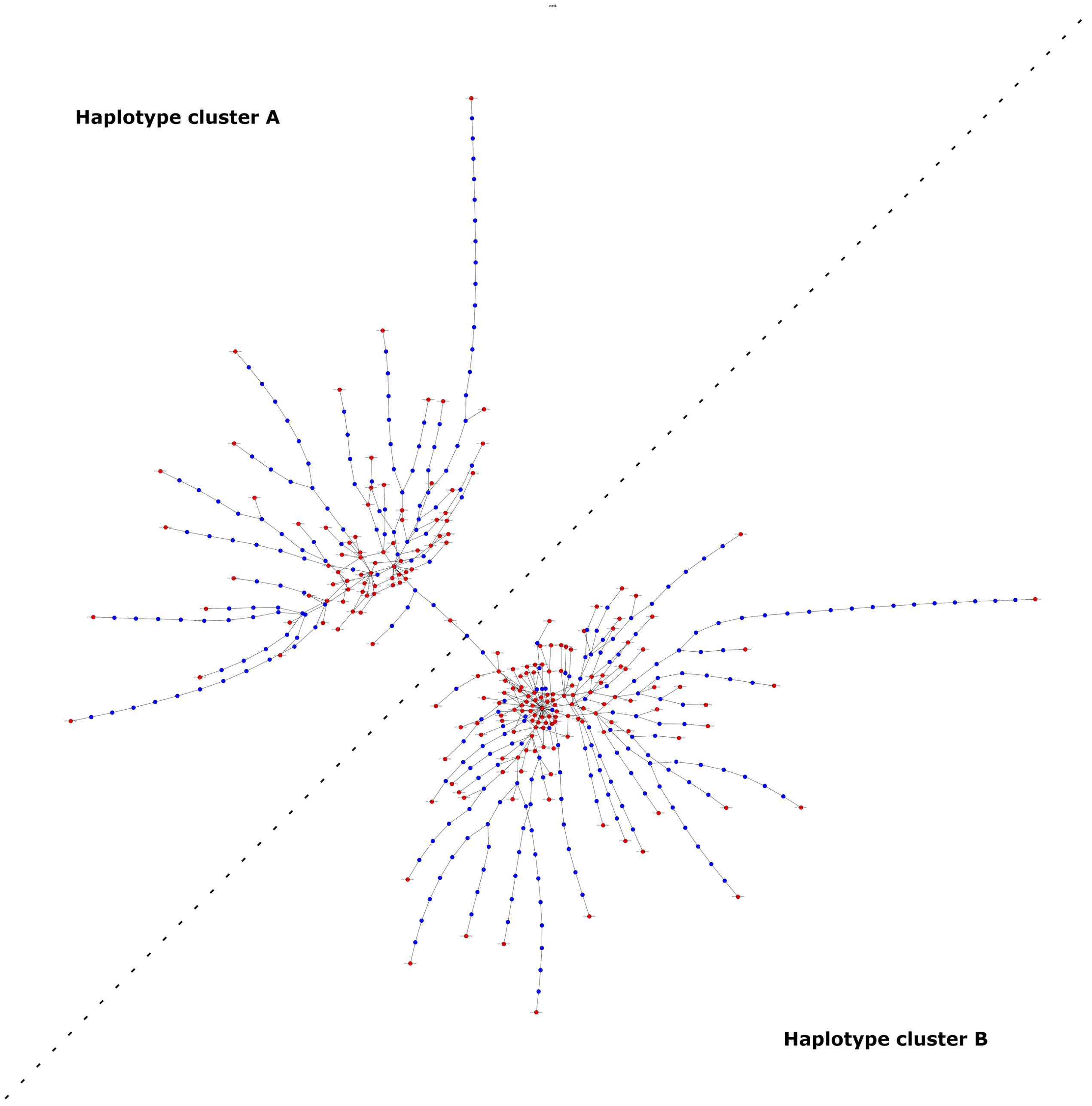


Figure 64: All individual haplotypes detected. Red dots represent observed unique haplotypes and blue dots represent hypothetical haplotypes (reproduced from Toffoli et al[125], under Creative Commons CC BY license).

5.3.2. Effect of the haplotypes on age at diagnosis of PD

Table 36 and Figure 65 show the age at diagnosis in the different groups.

I did not observe any differences with either the additive model or the dominant model.

I noticed that some participants were diagnosed with PD at a very young age. These participants have a higher chance of carrying variants in other PD causing genes (e.g.

PINK/Parkin). For this reason, I repeated the analysis after excluding participants with an age at diagnosis earlier than 50. Again, no significant differences were observed.

ALL AGES									
	RAPSODI Cohort			AMP-PD Cohort			Merged Cohorts (RAPSODI + AMP-PD)		
	Number of Participants	Mean Age at Diagnosis	Median Age at Diagnosis	Number of Participants	Mean Age at Diagnosis	Median Age at Diagnosis	Number of Participants	Mean Age at Diagnosis	Median Age at Diagnosis
Hom haplotype A	10	61.5	59.5	131	60.1	61	141	60.2	61
Heterozygous	33	59.9	60	540	59.5	61	573	59.5	61
Hom haplotype B	52	61.3	61	641	60.1	61	693	60.2	61
Other	5			5			10		
Total	100			1317			1417		
p-value additive model	0.7292			0.5819			0.528		
p-value dominant haplotype B	0.7892383			0.7981861			0.7552578		
	AGE AT DIAGNOSIS > 50								
	RAPSODI Cohort			AMP-PD Cohort			Merged Cohorts (RAPSODI + AMP-PD)		
	Number of Participants	Mean Age at Diagnosis	Median Age at Diagnosis	Number of Participants	Mean Age at Diagnosis	Median Age at Diagnosis	Number of Participants	Mean Age at Diagnosis	Median Age at Diagnosis
Hom haplotype A	10	61.5	59.5	104	64	64	114	63.8	64
Heterozygous	29	61.8	61	445	63	63	474	62.9	63
Hom haplotype B	48	62.7	61	526	63.6	63	574	63.5	63
Other	4			5			9		
Total	91			1080			1171		
p-value additive model	0.5504			0.7416			0.6431		
p-value dominant haplotype B	0.7378356			0.3714943			0.4654556		

Table 36: Number of participants carrying each haplotype and age at diagnosis of PD.

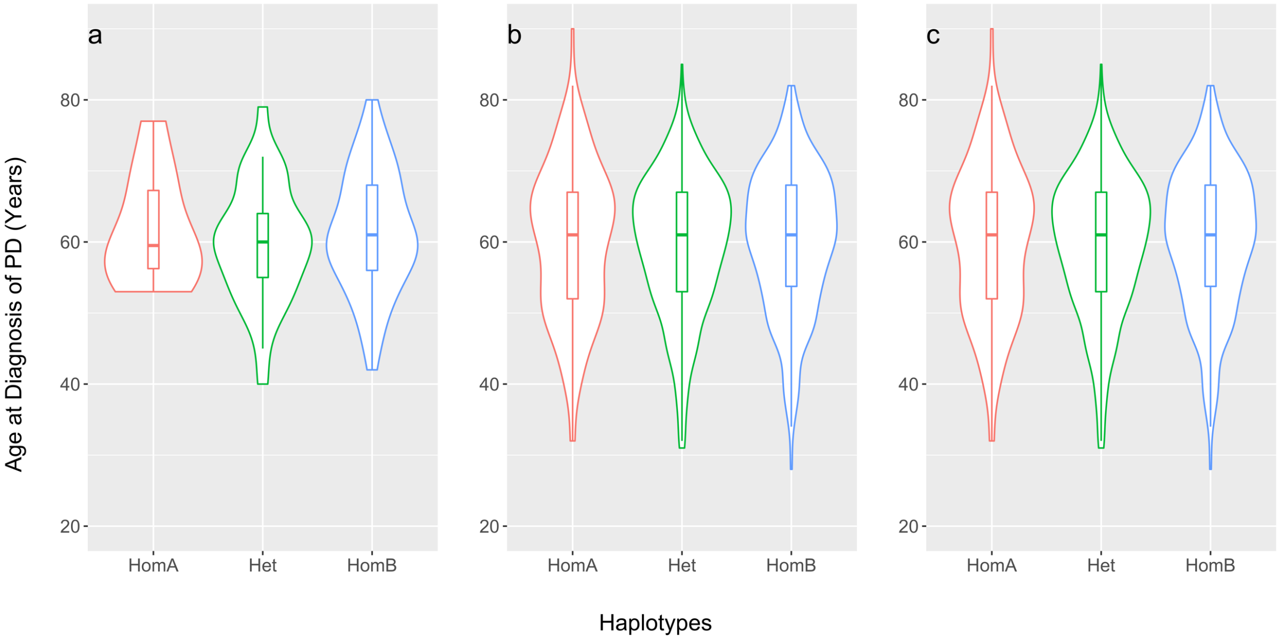


Figure 65: Age at diagnosis of PD and haplotypes. A) RAPSODI cohort. B) AMP-PD cohort. C) Joint cohorts.

5.4. Discussion

The idea that intronic variants and haplotypes in *GBA* can affect the risk and phenotype of PD is interesting and, if true, could in part explain the different penetrance of *GBA* mutations and why only a minority of *GBA* carriers develop PD in their lifetime.

The report by Schierding et al. that two common intronic haplotypes in *GBA* affected age at onset and diagnosis in their PD cohort was promising and suggested further investigation[124].

In an attempt to validate this finding, I run the same analysis in the RAPSODI cohort and sought additional confirmation by looking at the AMP-PD database, containing aggregated data of independent cohorts of PD patients from different parts of the world (https://amp-pd.org).

I tested both a dominant model and an additive model, and did not find any effect on age at diagnosis of PD.

Moreover, to exclude the possibility that a small number of early onset PD patients could skew the analysis, I re-run the same models after excluding individuals with an age at onset below 50. Again, I did not detect any significant effect of the haplotypes and I was unable to confirm the hypothesis formulated by Schierding et al[124].

One potential limitation of this study is that short read WGS data are unreliable in some regions within *GBA*[109], as extensively discussed in chapter 4, and the AMP-PD calls might be imprecise. However, the 3 haplotypes I considered in this analysis are all outside of the regions where *GBA* and *GBAP1* are highly homologous, suggesting that misalignment is less likely. Moreover, the allele frequencies observed in both AMP-PD and RAPSODI were very similar, and consistent with those reported in non-Finnish Europeans in gnomAD data[185], where the minor allele frequencies for the 3 variants are 0.295, 0.294, and 0.287.

One way to explain the discrepancies between the results reported in this chapter and those produced by Schierding et all[124] is the composition of the cohorts studied. Indeed, there were significant ethnic differences between the patients they studied, all from New Zealand and Australia, and those included in my analysis, mostly white Europeans and North-Americans. Moreover, both RAPSODI and AMP-PD included a significant amount of individuals of Ashkenazi Jewish ancestry. It is possible that the effect observed by the New Zealand group is specific to the ethnic composition of their cohort, or it might have been masked in our cohort due to the high number of Ashkenazi Jews, for example. Further studies are granted to clarify these conflicting findings.

Of note, we observed a high number of rare intronic variants and unique haplotypes (Figure 64). DIVs can have a functional effect in different ways. They can generate new splicing donor and acceptor sites, causing alternative splicing, and they can disrupt transcription regulatory motifs.

Further research is granted to determine whether these rarer intronic variants affect the risk and phenotype of PD and other diseases.

This analysis showed how the application of the *GBA* long read sequencing I developed and refined can provide further insight into the pathogenesis of PD.

6. Final remarks

My whole PhD focussed on two simple questions: why are the penetrance and phenotype of *GBA*-associated PD so variable? Is it possible to predict who, among *GBA*-variant carriers, will develop PD?

To answer these questions, I contributed to the development of RAPSODI, a portal for the remote assessment of participants, focussing on non-motor symptoms of PD. The idea behind RAPSODI is that, in order to understand what factors affect the risk of PD, and what signs herald the development of the disease, a large number of participants need to be followed-up for a long period of time. As in-person, face to face assessment proved unsustainable[128], RAPSODI allowed us to maintain a large cohort of participants constantly engaged, monitored and with limited resources. During the setup of RAPSODI, I encountered many difficulties, including the impossibility to meet and recruit new participants for over a year, due to the limitations brought by the COVID-19 pandemic.

However, the most significant problem I had to face was sequencing the *GBA* gene. The original plan was to outsource Sanger sequencing of the saliva samples collected through RAPSODI to a separate UCL laboratory. This solution was not ideal, partly due to the relatively high cost of Sanger sequencing of all exons of *GBA*, but also because we were encountering a high number of false negative calls (e.g. GD patients with only one or no *GBA* variants detected). For these reasons, I decided to develop a novel method for sequencing *GBA* with ONT long read technology. What started as a mere mean to an end, uncovered a whole new world of opportunities and interesting lines of research.

The novel method I developed is able to detect intronic variants and complex recombinants, even when they do not affect the coding sequence of *GBA*. I was able to investigate some of these potential genetic modifiers of the risk of PD myself, and more research will be carried out in the future.

All in all, I feel that my research advanced the knowledge on the role of *GBA* in the pathogenesis of PD, and hopefully will lead to more exciting findings in the future.
I was able to contribute to the development of different software, and in particular Gauchian, the novel *GBA* caller based on Illumina short read WGS, which has the potential to be a major contribution to research involving *GBA*. The method I developed for analysing *GBA* with ONT is publicly available and has already been applied by other research groups[186].
Finally, RAPSODI will continue recruiting and assessing participants for many years, and hopefully will provide further insight into the pathogenesis of PD.

Bibliography

1. Parkinson, J. (1817). An Essay on the Shaking Palsy Member.
2. Dorsey, E.R., Constantinescu, R., Thompson, J.P., Biglan, K.M., Holloway, R.G., Kieburtz, K., Marshall, F.J., Ravina, B.M., Schifitto, G., Siderowf, A., et al. (2007). Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. Neurology *68*, 384–386.
3. Dickson, D.W. (2012). Parkinson's Disease and Parkinsonism: Neuropathology. Cold Spring Harbor Perspectives in Medicine *2*, a009258–a009258.
4. Braak, H., Del Tredici, K., Rüb, U., de Vos, R.A.I., Jansen Steur, E.N.H., and Braak, E. Staging of brain pathology related to sporadic Parkinson's disease. Neurobiology of Aging *24*, 197–211.
5. Tolosa, E., Wenning, G., and Poewe, W. (2006). The diagnosis of Parkinson's disease. The Lancet Neurology *5*, 75–86.
6. Pfeiffer, R.F. (2016). Non-motor symptoms in Parkinson's disease. Parkinsonism & Related Disorders *22*, S119–S122.
7. Mahlknecht, P., Seppi, K., and Poewe, W. (2015). The Concept of Prodromal Parkinson's Disease. Journal of Parkinson's Disease *5*, 681–697.
8. Armstrong, M.J., and Okun, M.S. (2020). Diagnosis and Treatment of Parkinson Disease: A Review. JAMA *323*, 548–560.
9. Toffoli, M., Vieira, S.R.L., and Schapira, A.H.V. (2020). Genetic causes of PD: A pathway to disease modification. Neuropharmacology *170*, 108022.
10. Aarsland, D., Creese, B., Politis, M., Chaudhuri, K.R., Ffytche, D.H., Weintraub, D., and Ballard, C. (2017). Cognitive decline in Parkinson disease. Nature Reviews Neurology *13*, 217–231.
11. Emre, M., Aarsland, D., Brown, R., Burn, D.J., Duyckaerts, C., Mizuno, Y., Broe, G.A., Cummings, J., Dickson, D.W., Gauthier, S., et al. (2007). Clinical diagnostic criteria for dementia associated with Parkinson's disease. Movement Disorders *22*, 1689–1707.
12. Litvan, I., Goldman, J.G., Tröster, A.I., Schmand, B.A., Weintraub, D., Petersen, R.C., Mollenhauer, B., Adler, C.H., Marder, K., Williams-Gray, C.H., et al. (2012). Diagnostic criteria for mild cognitive impairment in Parkinson's disease: Movement Disorder Society Task Force guidelines. Movement Disorders *27*, 349–356.
13. Svenningsson, P., Westman, E., Ballard, C., and Aarsland, D. (2012). Cognitive impairment in patients with Parkinson's disease: Diagnosis, biomarkers, and treatment. The Lancet Neurology *11*, 697–707.
14. Williams-Gray, C.H., Mason, S.L., Evans, J.R., Foltynie, T., Brayne, C., Robbins, T.W., and Barker, R.A. (2013). The CamPaIGN study of Parkinson's disease: 10-year outlook in an incident population-based cohort. Journal of Neurology, Neurosurgery and Psychiatry *84*, 1258–1264.
15. Pigott, K., Rick, J., Xie, S.X., Hurtig, H., Chen-Plotkin, A., Duda, J.E., Morley, J.F., Chahine, L.M., Dahodwala, N., Akhtar, R.S., et al. (2015). Longitudinal study of normal cognition in Parkinson disease. Neurology *85*, 1276–1282.
16. Kempster, P.A., O'Sullivan, S.S., Holton, J.L., Revesz, T., and Lees, A.J. (2010). Relationships between age and late progression of Parkinson's disease: a clinico-pathological study. Brain *133*, 1755–1762.
17. Watson, G.S., and Leverenz, J.B. (2010). Profile of cognitive impairment in parkinson's disease. In Brain Pathology, (NIH Public Access), pp. 640–645.
18. Muslimović, D., Post, B., Speelman, J.D., and Schmand, B. (2005). Cognitive profile of patients with newly diagnosed Parkinson disease. Neurology *65*, 1239–1245.
19. Diamond, A. (2013). Executive functions. Annual Review of Psychology *64*, 135–168.
20. Phillips, L.H., Wynn, V.E., McPherson, S., and Gilhooly, K.J. (2001). Mental planning and the Tower of London task. The Quarterly Journal of Experimental Psychology Section A *54*, 579–597.
21. Tombaugh, T. (2004). Trail Making Test A and B: Normative data stratified by age and education. Archives of Clinical Neuropsychology *19*, 203–214.

22. Williams-Gray, C.H., Evans, J.R., Goris, A., Foltynie, T., Ban, M., Robbins, T.W., Brayne, C., Kolachana, B.S., Weinberger, D.R., Sawcer, S.J., et al. (2009). The distinct cognitive syndromes of Parkinson's disease: 5 year follow-up of the CamPaIGN cohort. Brain *132*, 2958–2969.
23. Aarsland, D., and Kurz, M.W. (2010). The epidemiology of dementia associated with parkinson's disease. In Brain Pathology, (Brain Pathol), pp. 633–639.
24. Nyhus, E., and Barceló, F. (2009). The Wisconsin Card Sorting Test and the cognitive assessment of prefrontal executive functions: A critical update. Brain and Cognition *71*, 437–451.
25. Uc, E.Y., McDermott, M.P., Marder, K.S., Anderson, S.W., Litvan, I., Como, P.G., Auinger, P., Chou, K.L., and Growdon, J.C. (2009). Incidence of and risk factors for cognitive impairment in an early parkinson disease clinical trial cohort. Neurology *73*, 1469–1477.
26. Bohnen, N.I., Kaufer, D.I., Hendrickson, R., Ivanco, L.S., Lopresti, B.J., Constantine, G.M., Mathis, C.A., Davis, J.G., Moore, R.Y., and DeKosky, S.T. (2006). Cognitive correlates of cortical cholinergic denervation in Parkinson's disease and parkinsonian dementia. Journal of Neurology *253*, 242–247.
27. Cowan, N. (2008). Chapter 20 What are the differences between long-term, short-term, and working memory? Progress in Brain Research *169*, 323–338.
28. Whittington, C., Podd, J., and Stewart-Williams, S. (2006). Memory deficits in Parkinson's disease. Journal of Clinical and Experimental Neuropsychology *28*, 738–754.
29. Postuma, R.B., Aarsland, D., Barone, P., Burn, D.J., Hawkes, C.H., Oertel, W., and Ziemssen, T. (2012). Identifying prodromal Parkinson's disease: pre-motor disorders in Parkinson's disease. Movement Disorders : Official Journal of the Movement Disorder Society *27*, 617–626.
30. Aarsland, D., Brønnick, K., Larsen, J.P., Tysnes, O.B., and Alves, G. (2009). Cognitive impairment in incident, untreated parkinson disease: The norwegian parkwest study. Neurology *72*, 1121–1126.
31. Muslimović, D., Post, B., Speelman, J.D., and Schmand, B. (2007). Motor procedural learning in Parkinson's disease. Brain *130*, 2887–2897.
32. Weil, R.S., Schrag, A.E., Warren, J.D., Crutch, S.J., Lees, A.J., and Morris, H.R. (2016). Visual dysfunction in Parkinson's disease. Brain *139*, 2827–2843.
33. Uc, E.Y., Rizzo, M., Anderson, S.W., Sparks, J.D., Rodnitzky, R.L., and Dawson, J.D. (2006). Impaired visual search in drivers with Parkinson's disease. Annals of Neurology *60*, 407–413.
34. Williams-Gray, C.H., Foltynie, T., Lewis, S.J.G., and Barker, R.A. (2006). Cognitive deficits and psychosis in Parkinson's disease: A review of pathophysiology and therapeutic options. CNS Drugs *20*, 477–505.
35. Tröster, A.I., Fields, J.A., Paolo, A.M., and Koller, W.C. (2006). Absence of the apolipoprotein E ε4 allele is associated with working memory impairment in Parkinson's disease. Journal of the Neurological Sciences *248*, 62–67.
36. Kehagia, A.A., Barker, R.A., and Robbins, T.W. (2012). Cognitive impairment in Parkinson's disease: The dual syndrome hypothesis. Neurodegenerative Diseases *11*, 79–92.
37. Kumar, K., Djarmati-Westenberger, A., and Grünewald, A. (2011). Genetics of Parkinson's Disease. Seminars in Neurology *31*, 433–440.
38. Hernandez, D.G., Reed, X., and Singleton, A.B. (2016). Genetics in Parkinson disease: Mendelian versus non-Mendelian inheritance. Journal of Neurochemistry *139 Suppl*, 59–74.
39. Blauwendraat, C., Nalls, M.A., and Singleton, A.B. (2020). The genetic architecture of Parkinson's disease. The Lancet Neurology *19*, 170–178.
40. Chang, D., Nalls, M.A., Hallgrímsdóttir, I.B., Hunkapiller, J., Brug, M. van der, Cai, F., Kerchner, G.A., Ayalon, G., Bingol, B., Sheng, M., et al. (2017). A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. Nature Genetics *49*, 1511–1516.
41. Heinzel, S., Berg, D., Gasser, T., Chen, H., Yao, C., and Postuma, R.B. (2019). Update of the MDS research criteria for prodromal Parkinson's disease. Movement Disorders mds.27802.
42. Toffoli, M., Vieira, S.R.L., and Schapira, A.H.V. (2020). Genetic causes of PD: A pathway to disease modification. Neuropharmacology *170*, 108022.

43. Kumar, K., Djarmati-Westenberger, A., and Grünewald, A. (2011). Genetics of Parkinson's Disease. Seminars in Neurology *31*, 433–440.
44. Satake, W., Nakabayashi, Y., Mizuta, I., Hirota, Y., Ito, C., Kubo, M., Kawaguchi, T., Tsunoda, T., Watanabe, M., Takeda, A., et al. (2009). Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. Nature Genetics *41*, 1303–1307.
45. Stirnemann, J., Belmatoug, N., Camou, F., Serratrice, C., Froissart, R., Caillaud, C., Levade, T., Astudillo, L., Serratrice, J., Brassier, A., et al. (2017). A Review of Gaucher Disease Pathophysiology, Clinical Presentation and Treatments. International Journal of Molecular Sciences *18*, 441.
46. Meikle, P.J., Hopwood, J.J., Clague, A.E., and Carey, W.F. (2016). Prevalence of Lysosomal Storage Disorders. *281*, 249–254.
47. Poorthuis, B.J., Wevers, R.A., Kleijer, W.J., Groener, J.E., de Jong, J.G., van Weely, S., Niezen-Koning, K.E., and van Diggelen, O.P. The frequency of lysosomal storage diseases in The Netherlands. Human Genetics *105*, 151–156.
48. Dionisi-Vici, C., Rizzo, C., Burlina, A.B., Caruso, U., Sabetta, G., Uziel, G., and Abeni, D. (2002). Inborn errors of metabolism in the Italian pediatric population: A national retrospective survey. Journal of Pediatrics *140*, 321–327.
49. Beutler, E., and Grabowski, G.A. (2001). Gaucher disease. In The Metabolic and Molecular Basis of Inherited Diseases., McGraw-Hill, ed. (New York), pp. 3635–3668.
50. Riboldi, G.M., and Di Fonzo, A.B. (2019). GBA, Gaucher Disease, and Parkinson's Disease: From Genetic to Clinic to New Therapeutic Approaches. Cells *8*, 364.
51. Bembi, B., Zambito Marsala, S., Sidransky, E., Ciana, G., Carrozzi, M., Zorzon, M., Martini, C., Gioulis, M., Pittis, M.G., and Capus, L. (2003). Gaucher's disease with Parkinson's disease: clinical and pathological aspects. Neurology *61*, 99–101.
52. Neudorfer, O., Giladi, N., Elstein, D., Abrahamov, a, Turezkite, T., Aghai, E., Reches, a, Bembi, B., and Zimran, a (1996). Occurrence of Parkinson's syndrome in type I Gaucher disease. QJM : Monthly Journal of the Association of Physicians *89*, 691–694.
53. Tayebi, N., Walker, J., Stubblefield, B., Orvisky, E., LaMarca, M.E., Wong, K., Rosenbaum, H., Schiffmann, R., Bembi, B., and Sidransky, E. (2003). Gaucher disease with parkinsonian manifestations: does glucocerebrosidase deficiency contribute to a vulnerability to parkinsonism? Molecular Genetics and Metabolism *79*, 104–109.
54. Choi, J.M., Kim, W.C., Lyoo, C.H., Kang, S.Y., Lee, P.H., Baik, J.S., Koh, S.-B., Ma, H.-I., Sohn, Y.H., Lee, M.S., et al. (2012). Association of mutations in the glucocerebrosidase gene with Parkinson disease in a Korean population. Neuroscience Letters *514*, 12–15.
55. Kumar, K.R., Ramirez, A., Göbel, A., Kresojević, N., Svetel, M., Lohmann, K., M Sue, C., Rolfs, A., Mazzulli, J.R., Alcalay, R.N., et al. (2013). Glucocerebrosidase mutations in a Serbian Parkinson's disease population. European Journal of Neurology *20*, 402–405.
56. Lwin, A., Orvisky, E., Goker-Alpan, O., LaMarca, M.E., and Sidransky, E. (2004). Glucocerebrosidase mutations in subjects with parkinsonism. Molecular Genetics and Metabolism *81*, 70–73.
57. Mao, X.-Y., Burgunder, J.-M., Zhang, Z.-J., An, X.-K., Zhang, J.-H., Yang, Y., Li, T., Wang, Y.-C., Chang, X.-L., and Peng, R. (2010). Association between GBA L444P mutation and sporadic Parkinson's disease from Mainland China. Neuroscience Letters *469*, 256–259.
58. Moraitou, M., Hadjigeorgiou, G., Monopolis, I., Dardiotis, E., Bozi, M., Vassilatis, D., Vilageliu, L., Grinberg, D., Xiromerisiou, G., Stefanis, L., et al. (2011). β-Glucocerebrosidase gene mutations in two cohorts of Greek patients with sporadic Parkinson's disease. Molecular Genetics and Metabolism *104*, 149–152.
59. Ziegler, S.G., Eblan, M.J., Gutti, U., Hruska, K.S., Stubblefield, B.K., Goker-Alpan, O., LaMarca, M.E., Sidransky, E., Shira G. Ziegler, Michael J. Eblan, B.A., Usha Gutti, M.S., Kathleen S. Hruska, Ph.D., B., K. Stubblefield, B.S., Ozlem Goker-Alpan, M.D., Mary E LaMarca, B.A., and E.S., et al. (2007). Glucocerebrosidase mutations in Chinese subjects from Taiwan with

sporadic Parkinson disease. Molecular Genetics and Metabolism *91*, 195–200.

60. Sidransky, E., Nalls, M.A.A., Aasly, J.O.O., Aharon-Peretz, J., Annesi, G., Barbosa, E.R.R., Bar-Shira, A., Berg, D., Bras, J., Brice, A., et al. (2009). Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. The New England Journal of Medicine *361*, 1651–1661.

61. Chen, J., Li, W., Zhang, T., Wang, Y., Jiang, X., and Xu, Z. (2014). Glucocerebrosidase gene mutations associated with Parkinson's disease: a meta-analysis in a Chinese population. PloS One *9*, e115747.

62. Do, C.B., Tung, J.Y., Dorfman, E., Kiefer, A.K., Drabant, E.M., Francke, U., Mountain, J.L., Goldman, S.M., Tanner, C.M., Langston, J.W., et al. (2011). Web-based genome-wide association study identifies two novel loci and a substantial genetic component for parkinson's disease. PLoS Genetics *7*,.

63. Nalls, M. a, Pankratz, N., Lill, C.M., Do, C.B., Hernandez, D.G., Saad, M., DeStefano, A.L., Kara, E., Bras, J., Sharma, M., et al. (2014). Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. Nature Genetics *46*, 989–993.

64. Simon-Sanchez, J., Schulte, C., Bras, J.J.M., Simón-Sánchez, J., Schulte, C., Bras, J.J.M., Sharma, M., Gibbs, J.R., Berg, D., Paisan-Ruiz, C., et al. (2009). Genome-wide association study reveals genetic risk underlying Parkinson's disease. … Genetics *41*, 1308–1312.

65. O'Regan, G., DeSouza, R.-M., Balestrino, R., and Schapira, A.H. (2017). Glucocerebrosidase Mutations in Parkinson Disease. Journal of Parkinson's Disease *7*, 411–422.

66. Kim, S., Yun, S.P., Lee, S., Umanah, G.E., Bandaru, V.V.R., Yin, X., Rhee, P., Karuppagounder, S.S., Kwon, S.-H., Lee, H., et al. (2018). GBA1 deficiency negatively affects physiological α-synuclein tetramers and related multimers. Proceedings of the National Academy of Sciences *115*, 798–803.

67. Rocha, E.M., Smith, G.A., Park, E., Cao, H., Brown, E., Hallett, P., and Isacson, O. (2015). Progressive decline of glucocerebrosidase in aging and Parkinson's disease. Annals of Clinical and Translational Neurology *2*, 433–438.

68. Suzuki, M., Fujikake, N., Takeuchi, T., Kohyama-Koganeya, A., Nakajima, K., Hirabayashi, Y., Wada, K., and Nagai, Y. (2015). Glucocerebrosidase deficiency accelerates the accumulation of proteinase K-resistant α-synuclein and aggravates neurodegeneration in a *Drosophila* model of Parkinson's disease. Human Molecular Genetics *24*, 6675–6686.

69. Gegg, M.E., Sweet, L., Wang, B.H., Shihabuddin, L.S., Sardi, S.P., and Schapira, A.H.V. (2015). No evidence for substrate accumulation in Parkinson brains with GBA mutations. Movement Disorders *30*, 1085–1089.

70. Huebecker, M., Moloney, E.B., Van Der Spoel, A.C., Priestman, D.A., Isacson, O., Hallett, P.J., and Platt, F.M. (2019). Reduced sphingolipid hydrolase activities, substrate accumulation and ganglioside decline in Parkinson's disease. Molecular Neurodegeneration *14*,.

71. Schöndorf, D.C., Aureli, M., McAllister, F.E., Hindley, C.J., Mayer, F., Schmid, B., Sardi, S.P., Valsecchi, M., Hoffmann, S., Schwarz, L.K., et al. (2014). IPSC-derived neurons from GBA1-associated Parkinson's disease patients show autophagic defects and impaired calcium homeostasis. Nature Communications *5*, 4028.

72. Manning-Boğ, A.B., Schüle, B., and Langston, J.W. (2009). Alpha-synuclein-glucocerebrosidase interactions in pharmacological Gaucher models: a biological link between Gaucher disease and parkinsonism. Neurotoxicology *30*, 1127–1132.

73. Bendikov-Bar, I., and Horowitz, M. (2012). Gaucher disease paradigm: from ERAD to comorbidity. Human Mutation *33*, 1398–1407.

74. Fernandes, H.J.R., Hartfield, E.M., Christian, H.C., Emmanoulidou, E., Zheng, Y., Booth, H., Bogetofte, H., Lang, C., Ryan, B.J., Sardi, S.P., et al. (2016). ER Stress and Autophagic Perturbations Lead to Elevated Extracellular α-Synuclein in GBA-N370S Parkinson's iPSC-Derived Dopamine Neurons. Stem Cell Reports *6*, 342–356.

75. Gan-Or, Z., Amshalom, I., Kilarski, L.L., Bar-Shira, A., Gana-Weisz, M., Mirelman, A., Marder, K., Bressman, S., Giladi, N., and Orr-Urtreger, A. (2015). Differential effects of severe vs

mild GBA mutations on Parkinson disease. Neurology *84*, 880–887.
76. Sidransky, E., and Lopez, G. (2012). The link between the GBA gene and parkinsonism. The Lancet Neurology *11*, 986–998.
77. Duran, R., Mencacci, N.E., Angeli, A. V., Shoai, M., Deas, E., Houlden, H., Mehta, A., Hughes, D., Cox, T.M., Deegan, P., et al. (2013). The glucocerobrosidase E326K variant predisposes to Parkinson's disease, but does not cause Gaucher's disease. Movement Disorders *28*, 232–236.
78. Gegg, M.E., Burke, D., Heales, S.J.R.R., Cooper, J.M., Hardy, J., Wood, N.W., and Schapira, A.H.V. V (2012). Glucocerebrosidase deficiency in substantia nigra of parkinson disease brains. Annals of Neurology *72*, 455–463.
79. Mazzulli, J.R., Xu, Y.H., Sun, Y., Knight, A.L., McLean, P.J., Caldwell, G.A., Sidransky, E., Grabowski, G.A., and Krainc, D. (2011). Gaucher disease glucocerebrosidase and α-synuclein form a bidirectional pathogenic loop in synucleinopathies. Cell *146*, 37–52.
80. Toffoli, M., Smith, L., and Schapira, A.H. V. (2020). The biochemical basis of interactions between Glucocerebrosidase and alpha-synuclein in GBA 1 mutation carriers. Journal of Neurochemistry jnc.14968.
81. Gan-Or, Z., Giladi, N., Rozovski, U., Shifrin, C., Rosner, S., Gurevich, T., Bar-Shira, A., and Orr-Urtreger, A. (2008). Genotype-phenotype correlations between GBA mutations and Parkinson disease risk and onset. Neurology *70*, 2277–2283.
82. Cilia, R., Tunesi, S., Marotta, G., Cereda, E., Siri, C., Tesei, S., Zecchinelli, A.L., Canesi, M., Mariani, C.B., Meucci, N., et al. (2016). Survival and dementia in GBA -associated Parkinson's disease: The mutation matters. Annals of Neurology *80*, 662–673.
83. Thaler, A., Gurevich, T., Bar Shira, A., Gana Weisz, M., Ash, E., Shiner, T., Orr-Urtreger, A., Giladi, N., and Mirelman, A. (2017). A "dose" effect of mutations in the GBA gene on Parkinson's disease phenotype. Parkinsonism & Related Disorders *36*, 47–51.
84. Brockmann, K., Srulijes, K., Pflederer, S., Hauser, A., Schulte, C., Maetzler, W., Gasser, T., and Berg, D. (2015). GBA -associated Parkinson's disease: Reduced survival and more rapid progression in a prospective longitudinal study. Movement Disorders *30*, 407–411.
85. Jesús, S., Huertas, I., Bernal-Bernal, I., Bonilla-Toribio, M., Cáceres-Redondo, M.T., Vargas-González, L., Gómez-Llamas, M., Carrillo, F., Calderón, E., Carballo, M., et al. (2016). GBA Variants Influence Motor and Non-Motor Features of Parkinson's Disease. PLOS ONE *11*, e0167749.
86. Chahine, L.M., Qiang, J., Ashbridge, E., Minger, J., Yearout, D., Horn, S., Colcher, A., Hurtig, H.I., Lee, V.M.Y., Van Deerlin, V.M., et al. (2013). Clinical and Biochemical Differences in Patients Having Parkinson Disease With vs Without GBA Mutations. JAMA Neurology *70*, 852.
87. Li, Y., Sekine, T., Funayama, M., Li, L., Yoshino, H., Nishioka, K., Tomiyama, H., and Hattori, N. (2014). Clinicogenetic study of GBA mutations in patients with familial Parkinson's disease. Neurobiology of Aging *35*, 935.e3-935.e8.
88. Brockmann, K., Srulijes, K., Hauser, A.K., Schulte, C., Csoti, I., Gasser, T., and Berg, D. (2011). GBA-associated PD presents with nonmotor characteristics. Neurology *77*, 276–280.
89. Liu, G., Boot, B., Locascio, J.J., Jansen, I.E., Winder-Rhodes, S., Eberly, S., Elbaz, A., Brice, A., Ravina, B., van Hilten, J.J., et al. (2016). Specifically neuropathic Gaucher's mutations accelerate cognitive decline in Parkinson's. Annals of Neurology *80*, 674–685.
90. Mata, I.F., Leverenz, J.B., Weintraub, D., Trojanowski, J.Q., Chen-Plotkin, A., Van Deerlin, V.M., Ritz, B., Rausch, R., Factor, S.A., Wood-Siverio, C., et al. (2016). GBA Variants are associated with a distinct pattern of cognitive deficits in Parkinson's disease. Movement Disorders *31*, 95–102.
91. Alcalay, R.N., Caccappolo, E., Mejia-Santana, H., Tang, M.X., Rosado, L., Orbe Reilly, M., Ruiz, D., Ross, B., Verbitsky, M., Kisselev, S., et al. (2012). Cognitive performance of GBA mutation carriers with early-onset PD: The CORE-PD study. Neurology *78*, 1434–1440.
92. Oeda, T., Umemura, A., Mori, Y., Tomita, S., Kohsaka, M., Park, K., Inoue, K., Fujimura, H., Hasegawa, H., Sugiyama, H., et al. (2015). Impact of glucocerebrosidase mutations on motor and

nonmotor complications in Parkinson's disease. Neurobiology of Aging *36*, 3306–3313.
93. Revel-Vilk, S., Szer, J., Mehta, A., and Zimran, A. (2018). How we manage Gaucher Disease in the era of choices. British Journal of Haematology *182*, 467–480.
94. Sardi, S.P., Viel, C., Clarke, J., Treleaven, C.M., Richards, A.M., Park, H., Olszewski, M.A., Dodge, J.C., Marshall, J., Makino, E., et al. (2017). Glucosylceramide synthase inhibition alleviates aberrations in synucleinopathy models. Proceedings of the National Academy of Sciences of the United States of America *114*, 2699–2704.
95. Sanofi: Press Releases, Friday, February 5, 2021.
96. Riboldi, G.M., and Di Fonzo, A.B. (2019). GBA, Gaucher Disease, and Parkinson's Disease: From Genetic to Clinic to New Therapeutic Approaches. Cells *8*, 364.
97. Horowitz, M., Elstein, D., Zimran, A., and Goker-Alpan, O. (2016). New Directions in Gaucher Disease. Human Mutation 1–16.
98. Maegawa, G.H.B., Tropak, M.B., Buttner, J.D., Rigat, B.A., Fuller, M., Pandit, D., Tang, L., Kornhaber, G.J., Hamuro, Y., Clarke, J.T.R., et al. (2009). Identification and characterization of ambroxol as an enzyme enhancement agent for Gaucher disease. The Journal of Biological Chemistry *284*, 23502–23516.
99. Migdalska-Richards, A., Daly, L., Bezard, E., and Schapira, A.H. V. (2016). Ambroxol effects in glucocerebrosidase and α-synuclein transgenic mice. Annals of Neurology *80*, 766–775.
100. Mullin, S., Smith, L., Lee, K., D'Souza, G., Woodgate, P., Elflein, J., Hällqvist, J., Toffoli, M., Streeter, A., Hosking, J., et al. (2020). Ambroxol for the Treatment of Patients With Parkinson Disease With and Without Glucocerebrosidase Gene Mutations. JAMA Neurology *77*, 427.
101. Cheng, H.-C., Ulane, C.M., and Burke, R.E. (2010). Clinical progression in Parkinson disease and the neurobiology of axons. Annals of Neurology *67*, 715–725.
102. Davis, M.Y., Johnson, C.O., Leverenz, J.B., Weintraub, D., Trojanowski, J.Q., Chen-Plotkin, A., Van Deerlin, V.M., Quinn, J.F., Chung, K.A., Peterson-Hiller, A.L., et al. (2016). Association of GBA Mutations and the E326K Polymorphism With Motor and Cognitive Progression in Parkinson Disease. JAMA Neurol *73*, 1217–1224.
103. Ruskey, J.A., Greenbaum, L., Roncière, L., Alam, A., Spiegelman, D., Liong, C., Levy, O.A., Waters, C., Fahn, S., Marder, K.S., et al. (2019). Increased yield of full GBA sequencing in Ashkenazi Jews with Parkinson's disease. Eur J Med Genet *62*, 65–69.
104. Armstrong, L.C., Komiya, T., Bergman, B.E., Mihara, K., and Bornstein, P. (1997). Metaxin Is a Component of a Preprotein Import Complex in the Outer Membrane of the Mammalian Mitochondrion. Journal of Biological Chemistry *272*, 6510–6518.
105. Long, G.L., Winfield, S., Adolph, K.W., Ginns, E.I., and Bornstein, P. (1996). Structure and Organization of the Human Metaxin Gene (MTX) and Pseudogene. Genomics *33*, 177–184.
106. Stone, D.L., Tayebi, N., Orvisky, E., Stubblefield, B., Madike, V., and Sidransky, E. (2000). Glucocerebrosidase gene mutations in patients with type 2 Gaucher disease. Human Mutation *15*, 181–188.
107. Hruska, K.S., LaMarca, M.E., Scott, C.R., and Sidransky, E. (2008). Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA). Human Mutation *29*, 567–583.
108. Straniero, L., Rimoldi, V., Melistaccio, G., Di Fonzo, A., Pezzoli, G., Duga, S., and Asselta, R. (2020). A rapid and low-cost test for screening the most common Parkinson's disease-related GBA variants. Parkinsonism & Related Disorders *80*, 138–141.
109. Zampieri, S., Cattarossi, S., Bembi, B., and Dardis, A. (2017). GBA Analysis in Next-Generation Era: Pitfalls, Challenges, and Possible Solutions. Journal of Molecular Diagnostics *19*, 733–741.
110. Woo, E.G., Tayebi, N., and Sidransky, E. (2021). Next-Generation Sequencing Analysis of GBA1: The Challenge of Detecting Complex Recombinant Alleles. Front Genet *12*, 684067.
111. Bodian, D.L., Klein, E., Iyer, R.K., Wong, W.S.W., Kothiyal, P., Stauffer, D., Huddleston, K.C., Gaither, A.D., Remsburg, I., Khromykh, A., et al. (2016). Utility of whole-genome

sequencing for detection of newborn screening disorders in a population cohort of 1,696 neonates. Genet Med *18*, 221–230.

112. Coop, G., and Przeworski, M. (2007). An evolutionary view of human recombination. Nat Rev Genet *8*, 23–34.

113. Carvalho, C.M.B., and Lupski, J.R. (2016). Mechanisms underlying structural variant formation in genomic disorders. Nature Reviews Genetics *17*, 224–238.

114. Latham, T., Grabowski, G.A., Theophilus, B.D., and Smith, F.I. (1990). Complex alleles of the acid beta-glucosidase gene in Gaucher disease. American Journal of Human Genetics *47*, 79–86.

115. Tayebi, N., Stubblefield, B.K., Park, J.K., Orvisky, E., Walker, J.M., LaMarca, M.E., and Sidransky, E. (2003). Reciprocal and Nonreciprocal Recombination at the Glucocerebrosidase Gene Region: Implications for Complexity in Gaucher Disease. The American Journal of Human Genetics *72*, 519–534.

116. Dobkin, C., Pergolizzi, R.G., Bahre, P., and Bank, A. (1983). Abnormal splice in a mutant human β-globin gene not at the site of a mutation. Proceedings of the National Academy of Sciences of the United States of America *80*, 1184–1188.

117. Bergsma, A.J., In't Groen, S.L.M., Verheijen, F.W., Van Der Ploeg, A.T., and Pijnappel, W.W.M.P. (2016). From Cryptic Toward Canonical Pre-mRNA Splicing in Pompe Disease: A Pipeline for the Development of Antisense Oligonucleotides. Molecular Therapy - Nucleic Acids *5*, e361.

118. Antonellis, A., Dennis, M.Y., Burzynski, G., Huynh, J., Maduro, V., Hodonsky, C.J., Khajavi, M., Szigeti, K., Mukkamala, S., Bessling, S.L., et al. (2010). A rare myelin protein zero (MPZ) variant alters enhancer activity in Vitro and in Vivo. PLoS ONE *5*,.

119. Malekkou, A., Sevastou, I., Mavrikiou, G., Georgiou, T., Vilageliu, L., Moraitou, M., Michelakakis, H., Prokopiou, C., and Drousiotou, A. (2020). A novel mutation deep within intron 7 of the GBA gene causes Gaucher disease. Molecular Genetics and Genomic Medicine *8*, 1090.

120. Vaz-Drago, R., Custódio, N., and Carmo-Fonseca, M. (2017). Deep intronic mutations and human disease. Human Genetics *136*, 1093–1111.

121. Sidransky, E., Nalls, M.A.A., Aasly, J.O.O., Aharon-Peretz, J., Annesi, G., Barbosa, E.R.R., Bar-Shira, A., Berg, D., Bras, J., Brice, A., et al. (2009). Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. The New England Journal of Medicine *361*, 1651–1661.

122. Zimran, A., Gelbart, T., and Beutler, E. (1990). Linkage of the PvuII polymorphism with the common Jewish mutation for Gaucher disease. American Journal of Human Genetics *46*, 902–905.

123. Beutler, E., West, C., and Gelbart, T. (1992). Polymorphisms in the human glucocerebrosidase gene. Genomics *12*, 795–800.

124. Schierding, W., Farrow, S., Fadason, T., Graham, O., Pitcher, T., Qubisi, S., Davidson, A.J., Perry, J.K., Anderson, T., Kennedy, M., et al. (2020). Common variants co-regulate expression of GBA and modifier genes to delay Parkinson's disease onset. Movement Disorders mds.28144.

125. Toffoli, M., Higgins, A., Lee, C., Koletsi, S., Chen, X., Eberle, M., Sedlazeck, F.J., Mullin, S., Proukakis, C., and Schapira, A.H.V. (2021). Intronic Haplotypes in the GBA Gene Do Not Predict Age at Diagnosis of Parkinson's Disease. Movement Disorders *36*, 1456–1460.

126. Beavan, M., McNeill, A., Proukakis, C., Hughes, D.A., Mehta, A., and Schapira, A.H. V. (2015). Evolution of Prodromal Clinical Markers of Parkinson Disease in a GBA Mutation–Positive Cohort. JAMA Neurology *72*, 201.

127. McNeill, A., Duran, R., Proukakis, C., Bras, J., Hughes, D., Mehta, A., Hardy, J., Wood, N.W., and Schapira, A.H.V. V (2012). Hyposmia and cognitive impairment in Gaucher disease patients and carriers. Movement Disorders *27*, 526–532.

128. Avenali, M., Toffoli, M., Mullin, S., McNeil, A., Hughes, D.A., Mehta, A., Blandini, F., and Schapira, A.H. V (2019). Evolution of prodromal parkinsonian features in a cohort of GBA mutation-positive individuals: a 6-year longitudinal study. Journal of Neurology, Neurosurgery & Psychiatry *90*, 1091–1097.

129. Noyce, A.J., R'Bibo, L., Peress, L., Bestwick, J.P., Adams-Carr, K.L., Mencacci, N.E.,

Hawkes, C.H., Masters, J.M., Wood, N., Hardy, J., et al. (2017). PREDICT-PD: An online approach to prospectively identify risk indicators of Parkinson's disease. Movement Disorders *32*, 219–226.

130. Postuma, R.B., Arnulf, I., Hogl, B., Iranzo, A., Miyamoto, T., Dauvilliers, Y., Oertel, W., Ju, Y.-E., Puligheddu, M., Jennum, P., et al. (2012). A Single-Question Screen for REM Sleep Behavior Disorder: A Multicenter Validation Study. Movement Disorders : Official Journal of the Movement Disorder Society *27*, 913.

131. Stiasny-Kolster, K., Mayer, G., Schäfer, S., Möller, J.C., Heinzel-Gutenbrunner, M., and Oertel, W.H. (2007). The REM sleep behavior disorder screening questionnaire-A new diagnostic instrument. Movement Disorders *22*, 2386–2393.

132. Martinez-Martin, P., Rodriguez-Blazquez, C., Alvarez-Sanchez, M., Arakaki, T., Bergareche-Yarza, A., Chade, A., Garretto, N., Gershanik, O., Kurtis, M.M., Martinez-Castrillo, J.C., et al. (2013). Expanded and independent validation of the Movement Disorder Society–Unified Parkinson's Disease Rating Scale (MDS-UPDRS). Journal of Neurology *260*, 228–236.

133. Zigmond, A.S., and Snaith, R.P. (1983). The Hospital Anxiety and Depression Scale. Acta Psychiatrica Scandinavica *67*, 361–370.

134. Hasan, H., Burrows, M., Athauda, D.S., Hellman, B., James, B., Warner, T., Foltynie, T., Giovannoni, G., Lees, A.J., and Noyce, A.J. (2019). The BRadykinesia Akinesia INcoordination (BRAIN) Tap Test: Capturing the Sequence Effect. Movement Disorders Clinical Practice *6*, 462–469.

135. Noyce, A.J., Nagy, A., Acharya, S., Hadavi, S., Bestwick, J.P., Fearnley, J., Lees, A.J., and Giovannoni, G. (2014). Bradykinesia-Akinesia Incoordination Test: Validating an Online Keyboard Test of Upper Limb Function. PLoS ONE *9*, e96260.

136. Wesnes, K.A., Brooker, H., Ballard, C., McCambridge, L., Stenton, R., and Corbett, A. (2017). Utility, reliability, sensitivity and validity of an online test system designed to monitor changes in cognitive function in clinical trials. International Journal of Geriatric Psychiatry *32*, e83–e92.

137. Doty, R.L. (2007). Office procedures for quantitative assessment of olfactory function. American Journal of Rhinology *21*, 460–473.

138. Hruska, K.S., LaMarca, M.E., Scott, C.R., and Sidransky, E. (2008). Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA). Human Mutation *29*, 567–583.

139. Nomura, T., Inoue, Y., Kagimura, T., Uemura, Y., and Nakashima, K. (2011). Utility of the REM sleep behavior disorder screening questionnaire (RBDSQ) in Parkinson's disease patients. Sleep Medicine *12*, 711–713.

140. McNeill, A., Duran, R., Proukakis, C., Bras, J., Hughes, D., Mehta, A., Hardy, J., Wood, N.W., and Schapira, A.H.V.V. (2012). Hyposmia and cognitive impairment in Gaucher disease patients and carriers. Movement Disorders *27*, 526–532.

141. Beavan, M., McNeill, A., Proukakis, C., Hughes, D.A., Mehta, A., and Schapira, A.H.V. (2015). Evolution of Prodromal Clinical Markers of Parkinson Disease in a GBA Mutation–Positive Cohort. JAMA Neurology *72*, 201.

142. Avenali, M., Toffoli, M., Mullin, S., McNeil, A., Hughes, D.A., Mehta, A., Blandini, F., and Schapira, A.H.V. (2019). Evolution of prodromal parkinsonian features in a cohort of GBA mutation-positive individuals: a 6-year longitudinal study. Journal of Neurology, Neurosurgery & Psychiatry *90*, 1091–1097.

143. Seidel, M., Brooker, H., Lauenborg, K., Wesnes, K., and Sjögren, M. (2021). Cognitive Function in Adults with Enduring Anorexia Nervosa. Nutrients *13*, 859.

144. Wesnes, K.A., Brooker, H., Ballard, C., McCambridge, L., Stenton, R., and Corbett, A. (2017). Utility, reliability, sensitivity and validity of an online test system designed to monitor changes in cognitive function in clinical trials. International Journal of Geriatric Psychiatry *32*, e83–e92.

145. Muslimović, D., Post, B., Speelman, J.D., and Schmand, B. (2005). Cognitive profile of patients with newly diagnosed Parkinson disease. Neurology *65*, 1239–1245.

146. Malek, N., Weil, R.S., Bresner, C., Lawton, M.A., Grosset, K.A., Tan, M., Bajaj, N., Barker, R.A., Burn, D.J., Foltynie, T., et al. (2018). Features of GBA-associated Parkinson's disease at presentation in the UK Tracking Parkinson's study. J Neurol Neurosurg Psychiatry *89*, 702–709.
147. Alcalay, R.N., Caccappolo, E., Mejia-Santana, H., Tang, M.X., Rosado, L., Orbe Reilly, M., Ruiz, D., Ross, B., Verbitsky, M., Kisselev, S., et al. (2012). Cognitive performance of GBA mutation carriers with early-onset PD: The CORE-PD study. Neurology *78*, 1434–1440.
148. Mata, I.F., Leverenz, J.B., Weintraub, D., Trojanowski, J.Q., Chen-Plotkin, A., Van Deerlin, V.M., Ritz, B., Rausch, R., Factor, S.A., Wood-Siverio, C., et al. (2016). GBA Variants are associated with a distinct pattern of cognitive deficits in Parkinson's disease. Movement Disorders *31*, 95–102.
149. Leija-Salazar, M., Sedlazeck, F.J.F.J., Toffoli, M., Mullin, S., Mokretar, K., Athanasopoulou, M., Donald, A., Sharma, R., Hughes, D., Schapira, A.H.V.A.H.V. V, et al. (2019). Evaluation of the detection of GBA missense mutations and other variants using the Oxford Nanopore MinION. Molecular Genetics & Genomic Medicine *7*, e564.
150. Do, J., McKinney, C., Sharma, P., and Sidransky, E. (2019). Glucocerebrosidase and its relevance to Parkinson disease. Molecular Neurodegeneration *14*, 36.
151. Ip, C.L.C., Loose, M., Tyson, J.R., de Cesare, M., Brown, B.L., Jain, M., Leggett, R.M., Eccles, D.A., Zalunin, V., Urban, J.M., et al. (2015). MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Research *4*, 1075.
152. Cretu Stancu, M., van Roosmalen, M.J., Renkens, I., Nieboer, M.M., Middelkamp, S., de Ligt, J., Pregno, G., Giachino, D., Mandrile, G., Espejo Valle-Inclan, J., et al. (2017). Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat Commun *8*, 1326.
153. Jain, M., Olsen, H.E., Paten, B., and Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biology *17*, 239.
154. Jeong, S.-Y., Kim, S.-J., Yang, J.-A., Hong, J.-H., Lee, S.-J., and Kim, H.J. (2011). Identification of a novel recombinant mutation in Korean patients with Gaucher disease using a long-range PCR approach. Journal of Human Genetics *56*, 469–471.
155. QIAquick PCR Purification Kit and QIAquick PCR & Gel Cleanup Kit Quick-Start Protocol - (EN) - QIAGEN.
156. AMPure XP Performance, PCR Purification - Beckman Coulter.
157. Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., and Schatz, M.C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. Nature Methods *15*, 461–468.
158. Luo, R., Wong, C.-L., Wong, Y.-S., Tang, C.-I., Liu, C.-M., Leung, C.-M., and Lam, T.-W. (2020). Exploring the limit of using a deep neural network on pileup data for germline variant calling. Nature Machine Intelligence *2*, 220–227.
159. Loman, N.J., Quick, J., and Simpson, J.T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods *12*, 733–735.
160. Martin, M., Patterson, M., Garg, S., O Fischer, S., Pisanti, N., Klau, G., Schöenhuth, A., and Marschall, T. (2016). WhatsHap: fast and accurate read-based phasing. BioRxiv 085050.
161. Stortchevoi, A., Kamelamela, N., and Levine, S.S. (2020). SPRI Beads-based Size Selection in the Range of 2-10kb. J Biomol Tech *31*, 7–10.
162. Leija-Salazar, M., Sedlazeck, F.J.F.J., Toffoli, M., Mullin, S., Mokretar, K., Athanasopoulou, M., Donald, A., Sharma, R., Hughes, D., Schapira, A.H.V.A.H.V.V., et al. (2019). Evaluation of the detection of GBA missense mutations and other variants using the Oxford Nanopore MinION. Molecular Genetics & Genomic Medicine *7*, e564.
163. Hidden Markov model.
164. Auwera, G.A.V. der, Carneiro, M.O., Hartl, C., Poplin, R., Angel, G. del, Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols

in Bioinformatics *43*, 11.10.1-11.10.33.
165. Toffoli, M., Chen, X., Sedlazeck, F.J., Lee, C.-Y., Mullin, S., Higgins, A., Koletsi, S., Garcia-Segura, M.E., Sammler, E., Scholz, S.W., et al. (2021). Comprehensive analysis of GBA using a novel algorithm for Illumina whole-genome sequence data or targeted Nanopore sequencing.
166. Gilpatrick, T., Lee, I., Graham, J.E., Raimondeau, E., Bowen, R., Heron, A., Downs, B., Sukumar, S., Sedlazeck, F.J., and Timp, W. (2020). Targeted nanopore sequencing with Cas9-guided adapter ligation. Nat Biotechnol *38*, 433–438.
167. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.
168. Huang, Y.-T., Liu, P.-Y., and Shih, P.-W. (2021). Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing. Genome Biology *22*, 95.
169. Loose, M., Malla, S., and Stout, M. (2016). Real-time selective sequencing using nanopore technology. Nature Methods *13*, 751–754.
170. Kovaka, S., Fan, Y., Ni, B., Timp, W., and Schatz, M.C. (2020). Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. Nature Biotechnology 1–11.
171. Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., Coffey, C., Kieburtz, K., Flagg, E., Chowdhury, S., et al. (2011). The Parkinson Progression Marker Initiative (PPMI). Progress in Neurobiology *95*, 629–635.
172. Martin Frith / last.
173. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
174. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative Genomics Viewer. Nat Biotechnol *29*, 24–26.
175. Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., Liu, Y., Liu, B., and Wang, Y. (2020). Long-read-based human genomic structural variation detection with cuteSV. Genome Biology *21*, 189.
176. Mapping and phasing of structural variation in patient genomes using nanopore sequencing | Nature Communications.
177. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv:1303.3997 [q-Bio].
178. Tayebi, N., Stubblefield, B.K., Park, J.K., Orvisky, E., Walker, J.M., LaMarca, M.E., and Sidransky, E. (2003). Reciprocal and Nonreciprocal Recombination at the Glucocerebrosidase Gene Region: Implications for Complexity in Gaucher Disease. The American Journal of Human Genetics *72*, 519–534.
179. Thomson, D.W., and Dinger, M.E. (2016). Endogenous microRNA sponges: evidence and controversy. Nat Rev Genet *17*, 272–283.
180. Romano, M., Buratti, E., and Baralle, D. (2013). Role of pseudoexons and pseudointrons in human cancer. Int J Cell Biol *2013*, 810572.
181. Dhir, A., and Buratti, E. (2010). Alternative splicing: role of pseudoexons in human disease and potential therapeutic strategies. FEBS J *277*, 841–855.
182. Qian, X., Wang, J., Wang, M., Igelman, A.D., Jones, K.D., Li, Y., Wang, K., Goetz, K.E., Birch, D.G., Yang, P., et al. (2021). Identification of Deep-Intronic Splice Mutations in a Large Cohort of Patients With Inherited Retinal Diseases. Frontiers in Genetics *12*, 276.
183. Mendes de Almeida, R., Tavares, J., Martins, S., Carvalho, T., Enguita, F.J., Brito, D., Carmo-Fonseca, M., and Lopes, L.R. (2017). Whole gene sequencing identifies deep-intronic variants with potential functional impact in patients with hypertrophic cardiomyopathy. PLoS One *12*, e0182946.
184. Kumar, R.D., Burrage, L.C., Bartos, J., Ali, S., Schmitt, E., Nagamani, S.C.S., and LeMons, C. (2021). A deep intronic variant is a common cause of OTC deficiency in individuals with previously negative genetic testing. Mol Genet Metab Rep *26*, 100706.

185. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.
186. Schierding, W., Farrow, S., Fadason, T., Graham, O., Pitcher, T., Qubisi, S., Davidson, A.J., Perry, J.K., Anderson, T., Kennedy, M., et al. (2020). Common variants co-regulate expression of GBA and modifier genes to delay Parkinson's disease onset. Movement Disorders mds.28144.

Appendix – Code

Script to run Guppy on Myriad

```
#!/bin/bash -l

# Batch script to run a GPU job on Myriad

#$ -S /bin/bash

#$ -l gpu=2

#$ -l h_rt=24:00:00

#$ -l mem=5G

#$ -l tmpfs=200G

#$ -pe smp 12

#$ -N guppyGPU

#$ -wd /home/skgttof/Scratch

cd $TMPDIR

module unload compilers mpi
module load compilers/gnu/4.9.2
module load cuda/9.0.176-patch4/gnu-4.9.2

/home/skgttof/Scratch/ont-guppy/bin/guppy_basecaller --input_path /home/skgttof/Scratch/fast5 --save_path /home/skgttof/Scratch/fastq -r --flowcell FLO-MIN106 --kit SQK-LSK109 --ipc_callers 16 --num_callers 12 --gpu_runners_per_device 2 --device "cuda:all:100%" && /home/skgttof/Scratch/ont-guppy/bin/guppy_barcoder -i /home/skgttof/Scratch/fastq -s /home/skgttof/Scratch/fastq_barcoded --barcode_kits EXP-PBC096

tar zcvf $HOME/Scratch/files_from_job_$JOB_ID.tar.gz $TMPDIR
```

The wrapping script used to sequence the GBA gene with ONT

```
#!/bin/bash
set -o pipefail
set -e

###############readme
#this script will take demultiplexed fastq files (from Guppy) and do alignment with NGMLR and variants calling with
Clair

#usage: yeah -q /path/to/folder_with_fastq  -w /path/to/folder_for_output -r /path/to/reference_for_alignment -b
/path/to/bed_file_with_full_GBA_coverage -Q threshold_for_QG_filtering_of_called_variants -t target_region(format
chrx:nnnnnnnnn-nnnnnnnnn)[default chr1:155234452-155241249] -C /path/to/clair.py -m
/path/to/clair/modules/folder

#before starting you need all dependencies installed and added to PATH (NGMLR, bgzip, tabiz, annovar, whatshap,
vcftools, bcftools, Clair).

#I isntalled Clair without conda environments so that it works by just calling python /path/to/clair.py. If you installed it
differently you will have to update the script to activate and deactivate environments etc

###########setting options
target=chr1:155234452-155241249

while getopts 'q:w:r:b:Q:t:C:m:' OPTION
do
  case "$OPTION" in
  q)
    FQ="$OPTARG"
    ;;
  w)
    WD="$OPTARG"
    ;;
  r)
    ref="$OPTARG"
    ;;
  b)
    bed="$OPTARG"
    ;;
  t)
    target="$OPTARG"
    ;;
  Q)
    qual="$OPTARG"
    ;;
  C)
    CLAIR="$OPTARG"
    ;;
  m)
    MODEL="$OPTARG/model"
    ;;
  ?)
  echo "option missing" >&2
```

```
   exit 1
   ;;
   esac
done
shift "$(($OPTIND -1))"

echo preparing files

mkdir $WD/mergedfastq $WD/aligned $WD/sorted_indexed $WD/clair $WD/downsampled_bams $WD/finaloutput
$WD/coverage $WD/downsampled_sorted_indexed $WD/QG_filter $WD/phased $WD/phased_unfiltered

for D in $FQ/*
do

   find $D -name *.fastq -exec cat {} \; > $WD/mergedfastq/${D#"$(dirname "$D")"/}.fastq

done

echo aligning

for f in $WD/mergedfastq/*.fastq
do

   ngmlr \
   -t 8 \
   -r $ref \
   -q $f \
   -o $WD/aligned/$(basename $f .fastq).sam \
   -x ont

done

echo sorting and indexing

for f in $WD/aligned/*.sam
do

   samtools sort $f \
       -o $WD/sorted_indexed/$(basename $f .sam).bam

done

for f in $WD/sorted_indexed/*.bam
do

   samtools index $f

done

echo adjusting coverage

for f in $WD/sorted_indexed/*.bam
```

```
do

      var="$(coverageBed -mean -a $bed -b $f | awk '{print $4}')"
   int=${var%.*}

      if [ $int -gt 550 ]

   then

   var1="$(echo "1/($var/550)" | bc -l)"
      samtools view -s $var1 -b $f -o $WD/downsampled_bams/${f#"$(dirname "$f")"/}

   else

   cp $f $WD/downsampled_bams/${f#"$(dirname "$f")"/}

   fi

done

echo sorting and indexing 2

for f in $WD/downsampled_bams/*.bam
do

   samtools sort $f \
       -o $WD/downsampled_sorted_indexed/${f#"$(dirname "$f")"/}

done

for f in $WD/downsampled_sorted_indexed/*.bam
do

   samtools index $f

done

echo calling variants

target1=${target%:*}
Sacile=${target#*:}
target2=${Sacile%-*}
target3=${target#*-}

CONTIG_NAME=chr1

for f in $WD/downsampled_sorted_indexed/*.bam
do

   VARIANT_CALLING_OUTPUT_PATH=$WD/clair/$(basename $f .bam).vcf
   BAM_FILE_PATH=$f
   SAMPLE_NAME=$(basename $f .bam)
```

```
    python $CLAIR callVarBam --chkpnt_fn "$MODEL" --ref_fn "$ref" --bam_fn "$BAM_FILE_PATH" --ctgName "$target1" --sampleName "$SAMPLE_NAME" --call_fn "$VARIANT_CALLING_OUTPUT_PATH" --minCoverage 20 --threshold 0.2 --delay 0 --ctgStart "$target2" --ctgEnd "$target3" --dcov 1000

done

echo filtering variants

for f in $WD/clair/*.vcf
do

     bcftools view -i "GQ > $qual" $f -o $WD/QG_filter/$(basename $f .vcf).vcf

done

echo phasing

for f in $WD/QG_filter/*.vcf
do

    sample=$(basename $f .vcf)

    whatshap phase $f $WD/sorted_indexed/$sample.bam \
        --reference $ref \
        -o $WD/phased/$sample.vcf \
        --ignore-read-groups

done

echo phasing non-filtered vcfs

for f in $WD/clair/*.vcf
do

    sample=$(basename $f .vcf)

    whatshap phase $f $WD/sorted_indexed/$sample.bam \
        --reference $ref \
        -o $WD/phased_unfiltered/$sample.vcf \
        --ignore-read-groups

done

echo finalising

for f in  $WD/phased/*.vcf
do

    bgzip $f

done
```

```
for f in $WD/phased/*gz
do

   tabix -p vcf $f

done

vcf-merge $WD/phased/*.gz > $WD/finaloutput/final.vcf

for f in $WD/downsampled_sorted_indexed/*.bam
do

   coverageBed \
   -mean \
   -header \
   -a $bed \
   -b $f \
   > $WD/coverage/$(basename $f .bam).vcf

done

more $WD/coverage/* | cat > $WD/finaloutput/coverage.vcf

for f in  $WD/phased_unfiltered/*.vcf
do

   bgzip $f

done

for f in $WD/phased_unfiltered/*gz
do

   tabix -p vcf $f

done

echo oh yeah!
```

Albacore

```
read_fast5_basecaller.py
   --flowcell FLO-MIN106
   --kit SQK-LSK108
   --barcoding
   --output_format fastq
   --input /path/to/fast5
   --save_path /path/to/fastq
   --worker_threads 8 -r
```

Minimap2

```
minimap2
   --MD
   -ax map-ont /path/to/reference.fa
   path/to/.fastq > path/to/output.bam
```

LAST

```
last-train -P8 -Q0
     /path/to/reference
     /path/to/fastq
     > /path/to/.par

lastal
   -Q0
   -P 8
   -p /path/to/.par
   /path/to/reference
   /path/to/
   | last-split > /path/to/output.maf

maf-convert sam -d /path/to/output.maf > /path/to/output.sam
```

Nanopolish

```
nanopolish index
   -d path/to/fast5
   path/to/fastq

nanopolish variants
   -g /path/to/reference.fa
   -r /path/to/.fastq
   -b /path/to/sorted.bam
   --ploidy 2
   -o /path/to/output.vcf
   --fix-homopolymers
```

Script to detect drops in homopolymers

```
#!/bin/bash

#creates CSV file with depth of coverage per base in the specified range.
#USAGE: depthCSV.sh </path/to/folder/with/sortedandindexed/bamfiles> </path/to/output/directory> <path to bed file with region(s) of interest>
#requires python3 installed and the script fINDEL.py in the same folder as the main script
#also requires samtools added to PATH

set -o pipefail

bamdir=$1
outputdir=$2
region=$3

mkdir $outputdir/csv

for f in $bamdir/*.bam
do
   samtools depth $f -b $region | awk 'BEGIN {OFS = ","} {print $1, $2, $3}' >> ${outputdir}/csv/$(basename $f .bam).csv
done

echo 'Sample,Position,Depth,mean Depth flanking 100 positions,fraction of flancking mean depth' >> $outputdir/results.csv

for f in $outputdir/csv/*.csv
do
   python3 ./findel.py $f >> $outputdir/results.csv
done
#!/bin/bash
```

fINDEL

```
#path needs to be defined in bash shell

import os
import csv
import sys

path = sys.argv[1]

ROI1 = list(range(155239990, 155239996))
ROI2 = list(range(155239657, 155239662))
ROI = ROI1 + ROI2

depth = {}

with open(path, mode='r') as data:
    for index, value in enumerate(csv.reader(data)):
            depth[int(value[1])] = int(value[2])
for i in ROI:
    calcsum = 0
    for n in range(i-50, i+50, 1):
        calcsum += int(depth[n])
    calcsum -= int(depth[i])
    mean = calcsum / 100
    print(f'{os.path.basename(path)},{i},{depth[i]},{mean},{depth[i] / mean}')
```

Sniffle

```
sniffles \
    -m /path/to/.bam \
    -v /path/to/.vcf \
    -n -1 \
    -s 2 \
    --skip_parameter_estimation \
    -t 8
```

NanoSV

```
python3.6 /home/minionpc/nanosv/nanosv/NanoSV.py \
        $f \
    -t 8 \
    -o ${f}_LAST_nanoSV.vcf \
    -c /home/minionpc/Desktop/rds/UNCALLED/nanoSVconfig.ini
```

CuteSV

```
cuteSV <bamfile> \
      <reference> \
      <output.vcf> \
      <workingdirectory> \
   --max_cluster_bias_INS 100 \
   --diff_ratio_merging_INS 0.3 \
   --max_cluster_bias_DEL 100 \
   --diff_ratio_merging_DEL 0.3 \
   --threads 8
```

UNCALLED

```
#To prepare reference with uncalled
bwa index -p GBAextended /path/to/reference.fa
uncalled index -i /path/to/reference.fa -x GBAextended

#to mask reference with UNCALLED script
mask_internal.sh /path/to/reference.fa 10 100 /path/to/output.fa

#to do a dry run to test performance

uncalled pafstats
   -r /path/to/file.paf
   /path/to/output.paf

#To run UNCALLED with MinKNOW
uncalled realtime
   --port 8000
   -t 8
   --enrich
   -c 3
   -x GBAextended > /path/to/output.paf
```