

**Next generation sequencing approaches for
novel gene discovery
in South African Parkinson's disease families**

by

Nikita Simone Pillay

(Student Number: 4078453)

*Dissertation submitted in fulfilment of the requirement for the degree of
Doctor of Philosophy in Bioinformatics at the South African National Bioinformatics Institute,
Faculty of Natural Sciences, University of the Western Cape*



Supervisor: Professor Alan Christoffels
South African National Bioinformatics Institute
Faculty of Natural Sciences
University of the Western Cape

Co-Supervisor: Professor Soraya Bardien
Division of Molecular Biology and Human Genetics
Faculty of Medicine and Health Sciences
Stellenbosch University

December 2022

I. Declaration

I, **Nikita Simone Pillay**, declare that *Next generation sequencing approaches for novel gene discovery in South African Parkinson's disease families* is based solely on my own work (except where acknowledgements indicate otherwise), that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Date: 9th of December 2022

Signed: _____



II. Acknowledgements

I would like to express my sincerest gratitude to the following individuals, groups and organisations for their assistance throughout the duration of my PhD.

To Prof. Alan Christoffels, for affording me the opportunity to conduct my PhD at SANBI. Your patience and guidance during this project has been greatly appreciated.

To Prof. Soraya Bardien, for being a source of constant optimism and keen supervision. Your generous provision of time and support has played an instrumental role in ensuring my research study came to fruition. I am profoundly thankful to have been a part of your research group.

To neurologist, Prof. Jonathan Carr and research nurse, Debbie Acker, for your roles in recruiting the family (ZA 15) that served as the backbone of my study.

To our collaborators, Prof. Owen Ross, Prof. Mary-Claire King, Dr. Caitlin Viljoen, Dr. Ananyo Choudhury and Dr. Suzanne Lesage, for their invaluable assistance regarding the sequencing of the study participants and the subsequent screening analyses.

To the staff at the Central Analytical Facilities (CAF) unit at Stellenbosch University for performing the Sanger sequencing in this study.

To the Parkinson's Disease Research Group, for being a constant source of academic and social support throughout this study.

To my dearest friends and family, for always believing in me.

This research project was supervised by Prof. Alan Christoffels (South African National Bioinformatics Institute at the University of the Western Cape) and co-supervised by Prof. Soraya Bardien (Division of Molecular Biology and Human Genetics at Stellenbosch University). This study was funded by the South African Research Chairs Initiative (SARCHI) of the Department of Science and Technology and National Research Foundation of South Africa

III. Abstract

Introduction: In the last decade, next-generation sequencing (NGS) approaches have revolutionised the study of human genomics, particularly aiding the understanding of genetic diseases. Parkinson's disease (PD) is a complex neurodegenerative disorder with a heterogeneous genetic disposition. This disorder is clinically characterised by the progressive loss of dopaminergic neurons in the *substantia nigra pars compacta* (SNpc). Subsequently, this results in a severe decrease of available dopamine that manifests as a myriad of both motor and non-motor symptoms. Several genes, including α -synuclein (*SNCA*), parkin (*PRKN*), leucine-rich repeat kinase 2 (*LRRK2*), PTEN induced putative kinase 1 (*PINK1*), and protein deglycase (*DJ-1*), are confirmed as disease-causing in autosomal recessive (AR), autosomal dominant (AD), early-onset (EO), and late-onset (LO) forms of the disorder. Thus far, monogenic causes of the disease have been found to affect a small proportion of all individuals with PD. However, the discovery of these PD-associated genes has led to an increased understanding of the biological systems underlying the disease, which can improve the diagnosis, prognosis and clinical management of affected individuals.

To date, the limited number of PD studies performed in sub-Saharan Africa (SSA) have mostly consisted of single PD gene screening analyses to determine whether these genes are implicated in PD in individuals of African ancestry. These studies have determined that the majority of these individuals do not possess known causes of the disease and it is postulated that they may harbour novel disease-associated variants or genes. Whole exome sequencing (WES) studies incorporating PD-affected families that display Mendelian inheritance patterns are useful for the determination of novel pathogenic variants. However, the underlying disease-causing mechanisms in which these novel variants operate are rarely examined post-NGS analysis. Consequently, the present study aimed to use WES and *in silico* analysis approaches in a PD-affected family of South African Xhosa ancestry to identify a novel variant or gene that may be linked to the onset of disease, as well as the possible functional effect of that variant.

Methods and Results: WES was performed on two PD-affected siblings and two unaffected siblings of a South Africa family designated as ZA 15. A WES workflow consisting of BWA-MEM, GATk HaplotypeCaller and Ensembl-VEP was used to analyse the WES data. Variant call files (VCFs) were screened for the presence of variants in PD-associated genes to eliminate known causes of disease. Filtering of the VCFs using stringent criteria (heterozygous, exonic, non-synonymous variants shared by only the affected siblings with a Phred score > 30 , present in population databases with a minor allele frequency (MAF) < 0.01 and a CADD score > 20), produced a list of 68 variants of interest. Subsequent gene and protein expression analysis determined that 24 of the variants were expressed in

neuro-specific tissue. Co-segregation analysis revealed that only 20 of the variants co-segregated with the disease in the family. Screening of these variants through ancestry-specific and PD-specific private cohorts resulted in 3 remaining candidates of interest, namely; AHNAK nucleoprotein 2 (*AHNAK2*) p.D1540H, mesencephalic astrocyte-derived neurotrophic factor (*MANF*) p.A13V and zinc finger DHHC-type containing 11 (*ZDHHC11*) p.R276P. Subsequently, a single variant (p.A13V in the *MANF* gene) was prioritised for further study as the gene is known to be expressed in the SNpc (the main neuronal region implicated in PD), where it exerts a protective effect on dopaminergic neurons.

Bioinformatic *in silico* analysis was done to determine p.A13V's possible impact on MANF's protein structure/function. Conservation analysis using the ConSurf server revealed the variant had variable conservation and occurs in the interior of the protein structure (i.e. buried). Secondary structure analysis using Project HOPE, PredictProtein, SignalP 3.0 and Phobius indicated, importantly, that the variant is present in the hydrophobic core of the *signal peptide* of the protein. Furthermore, the variant was predicted to be destabilising at the sequence level with a change in Gibbs free energy ($\Delta\Delta G$) of -0.2 and -0.21 obtained from MuPro and I-Mutant 3.0, respectively. Robetta was found to produce the best theoretical models of all the servers used (Robetta, I-TASSER and DeepPotential), according to the scores generated by TM-Align. These structural protein models passed all the basic quality checks by Verify3D, Q-MEAN, Procheck and ERRAT and were deemed appropriate for further structural analysis. DUET, DynaMut and MaestroWeb predicted a destabilising effect of the variant on the wildtype structure, while MaestroWeb also indicated an increase in rigidity of the signal peptide, close to the cleavage site. Root mean square deviations (RMSD), root mean square fluctuations (RMSF) and principal component analysis (PCA) using GROMACS indicated a deviation in structural conformation and flexibility between the wildtype and mutant models.

Discussion and Conclusions: The molecular destabilisation caused to the MANF protein structure upon introduction of the p.A13V variant, particularly at the signal peptide cleavage site and towards the C-terminal of the protein, could potentially impact the protein's translocation and expression. Previous studies have linked mutations in the hydrophobic core of the signal peptide to mRNA degradation via the Regulation of Aberrant Protein Production (RAPP) pathway, which could lead to decreased expression levels of the protein. However, if the variant interferes with the cleavage of the signal peptide (which would prevent recognition by signal peptide receptors), protein translocation would be affected resulting in an accumulation of the protein in the endoplasmic reticulum (ER) which could aggravate ER stress. Furthermore, if the signal peptide is cleaved off but the variant prevents degradation of these molecules, they could aggregate and cause cytotoxicity. The possible interference of the protein's neuroprotective properties (in regards to its role as an ER stress regulator

and its potential link to mitochondrial function) could cause a PD-pathology and therefore, these findings necessitate further laboratory-based functional analysis.

Limitations of this study include the limited sample size (sequencing of only two affected and two unaffected siblings) and the sole use of WES for mutation screening which may miss exonic duplications/insertions and other more complex rearrangements. Although knockout studies have been previously performed on *MANF*, it is necessary to determine the possible effect of the p.A13V variant on the protein's expressivity and trafficking. Thus, the recommendations for future study include analysis of translocation, expression levels, signal peptide aggregation, mitochondrial function association and the possible induced phenotype in an animal model with a high rate of homology between the two *MANF* genes, such as *D. melanogaster*.

Our study served as a benchmark for the analysis of PD-affected families of diverse ancestry. The use of WES and *in silico* analysis in an African ancestry family affected with PD proved to be useful in identifying a potentially new PD susceptibility factor. However, it also highlighted the necessity for the inclusion of diverse African population data (particularly in large population databases) for improved NGS analysis. In conclusion, determining the complex genetic architecture underlying PD, particularly in under-represented populations, is critical to provide insight into novel PD molecular mechanisms, detection of PD biomarkers, and elucidation of novel drug targets. Ultimately, this knowledge will change the course of future clinical diagnoses and therapeutic modalities for this currently, incurable disorder.

Keywords: African Ancestry; Bioinformatics; Familial PD; *In Silico* Analysis; *MANF*; Next-Generation Sequencing (NGS); Novel Variants; Parkinson's Disease (PD); Signal Peptide; Whole Exome Sequencing (WES)

IV. Abbreviations

$\Delta\Delta G$	Delta Delta G (Change In Gibb's Free Energy)
Å	Angstrom
°C	Degrees Celsius
µl	Microlitre
µM	Micromolar
3'	Three-Prime
5'	Five-Prime
AA	Amino Acid
AAO	Age At Onset
ACMG	American College Of Medical Genetics
AD	Autosomal Dominant
AFR	African/African-American
AGO2	Argonaute-2
AMR	Admixed American
AR	Autosomal Recessive
ATF6	Activating Transcription Factor 6 Alpha
ATP	Adenosine Triphosphate
ATP13A2	Atpase Cation Transporting 13A2
BAM	Binary Alignment Map
BLAST	Basic Local Alignment Search Tool
BLAT	Blast-Like Alignment Tool
bp	Base Pair
BWA	Burrows-Wheeler Aligner
BWA-MEM	Bwa's Maximal Exact Match
CADD	Combined Annotation Dependent Depletion
CAF	Central Analytical Facility
cDNA	Complementary Deoxyribonucleic Acid
CHCHD2	Coiled-Coil-Helix-Coiled-Coil-Helix Domain Containing 2
ClinVar	Public Archive Of Interpretations Of Clinically Relevant Variants
CNV	Copy Number Variation
COVID-19	Coronavirus Disease 2019
DA	Dopamine (3,4-Dihydroxyphenethylamine)
dH ₂ O	Deionised Water
DJ-1	Protein Deglycase
<i>DmMANF</i>	<i>Manf</i> Ortholog In Drosophila Melanogaster
DNA	Deoxyribonucleic Acid
DNAJC6	Dnaj Heat Shock Protein Family (Hsp40) Member C6
DNAJC13	Dnaj Heat Shock Protein Family (Hsp40) Member C13
dNTP	Deoxynucleoside Triphosphate
DOB	Date Of Birth
EAS	East Asian
Ensembl-VEP	Ensembl's Variant Effect Predictor
EO-PD	Early Onset Parkinson's Disease
ER	Endoplasmic Reticulum
EUR	European
ExAC	Exome Aggregation Consortium
fathmm	Functional Analysis Through Hidden Markov Models
FBXO7	F-Box Protein 7
FMPD	French And Mediterranean Parkinson's Disease Genetic Study Group
GATk	Genome Analysis Toolkit
GBA	Glucosylceramidase Beta
GC	Guanine/Cytosine
g	Grams

gDNA	Genomic Deoxyribonucleic Acid
gERP++	Genomic Evolutionary Rate Profiling
gnomAD	Genome Aggregation Database
GO	Gene Ontology
GP2	Global Parkinson's Genetics Program
GRCh37/38 (hg19/38)	Genome Reference Consortium Human Build 37/38
GRP78/BiP	78-Kda Glucose-Regulated Protein/Binding Immunoglobulin Protein
gTEX	Genotype-Tissue Expression
gVCF	Genomic Variant Call File
GWAS	Genome-Wide Association Studies
H	Hydrogen
H ₂ O	Water
H ₃ Africa	Human Heredity And Health In Africa
HPA	Human Phenotype Atlas
HRM	High-Resolution Melt
ID	Identification Number
InDels	Insertions/Deletions
IRE1	Inositol-Requiring Enzyme 1
kcal/mol	Kilocalorie Per Mole
KEGG	Kyoto Encyclopaedia Of Genes And Genomes
LB	Lewy Bodies
LO-PD	Late-Onset Parkinson's Disease
LRRK2	Leucine-Rich Repeat Kinase 2
MA	Mutation Assessor
MAF	Minor Allele Frequency
<i>MANF</i>	Mesencephalic Astrocyte Derived Neurotrophic Factor
MAPT	Microtubule Associated Protein Tau
M-CAP	Mendelian Clinically Applicable Pathogenicity
mCSM	Cutoff Scanning Matrix
MD	Molecular Dynamics
MgCl ₂	Magnesium Chloride
MGI	Mouse Genome Informatics
min	Minute
mL	Millilitre
MLPA	Multiplex Ligation-Dependent Probe Amplification
mM	Millimolar
MPP+	1-Methyl-4-Phenylpyridinium
MPTP	1-Methyl-4-Phenyl-1,2,3,6-Tetrahydropyran
MRI	Magnetic Resonance Imaging
mRNA	Messenger Ribonucleic Acid
MSA	Multiple Sequence Alignment
n	Sample Size
N/A	Not Applicable
NCBI	The National Center For Biotechnology Information
ng	Nanogram
ng/μl	Nanogram Per Microlitre
NFE	Non-Finnish European
NGS	Next-Generation Sequencing
NHGRI	National Human Genome Research Institute
nm	Nanometre
NN	Neural Network
NPT	Amount Of Substance (N), Pressure (P) And Temperature (T) (Isothermal–Isobaric Ensemble)
NVT	Amount Of Substance (N), Volume (V) And Temperature (T) (Canonical Ensemble)
OMIM	Online Mendelian Inheritance In Man
OPLS-AA	Optimised Potentials For Liquid Simulations - All Atom
PARK (n)	Pd-Associated Loci
PC	Principal Components
PCA	Principal Component Analysis

PCR	Polymerase Chain Reaction
PD	Parkinson's Disease
PDB	Protein Data Bank
PERK	Protein Kinase Rna- Like Endoplasmic Reticulum Kinase
PINK1	Pten Induced Putative Kinase 1
PLA2G6	85 Kda Calcium-Independent Phospholipase A2
pmol	Picomole
PolyPhen-2	Polymorphism Phenotyping v2
ps	Picosecond
PRKN	Parkin Rbr E3 Ubiquitin Protein Ligase
PROVEAN	Protein Variation Effect Analyzer
QMEAN	Qualitative Model Energy Analysis
QPP	Queensland Parkinson's Project
QS	Quality Score
RAPP	Regulation Of Aberrant Protein Production
REM	Rapid Eye Movement
RI	Reliability Index
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
RNA	Ribonucleic Acid
SA	South Africa
SAHGP	South African Human Genome Project
SAM	Sequence Alignment/Map
SANBI	South African National Bioinformatics Institute
SARS-Cov-2	Severe Acute Respiratory Syndrome Coronavirus 2
SAS	South Asian
SD	Standard Deviation
SH-SH5Y	Human Neuroblastoma Cell Line
SIFT	Sorts Intolerant From Tolerant
SNCA	Alpha-Synuclein
SNP	Single Nucleotide Polymorphisms
SNpc	Substantia Nigra Pars Compacta
SNV	Single Nucleotide Variation
SPME	Smooth Particle-Mesh Ewald
SRP	Signal Recognition Particle
SSA	Sub-Saharan Africa
SVM	Support Vector Machine
SYNJ1	Synaptojanin 1
Ta	Annealing Temperature
Tm	Melting Temperature
UCHL1	Ubiquitin C-Terminal Hydrolase L1
UPR	Unfolded Protein Response
UPS	Ubiquitin-Proteasome System
USA	United States Of America
V	Volts
VCF	Variant Call File
VPS35	Vps35 Retromer Complex Component
VPS13C	Vacuolar Protein Sorting 13 Homolog C
VUS	Variant Of Unknown Significance
w/v	Weight By Volume
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
ZA 15	South African-Xhosa Family Affected With Parkinson's Disease

V. List of Figures

		Page
<i>Chapter 2</i>		
<u>Figure 2.1</u>	Overview of a next-generation sequencing workflow	5
<u>Figure 2.2</u>	The number of publications related to NGS (indexed on PubMed) correlated to the cost of a single sequenced genome (2001-2021)	6
<u>Figure 2.3</u>	A comparison between whole genome sequencing, whole exome sequencing and targeted gene panels	7
<u>Figure 2.4</u>	Basic steps for NGS analysis	9
<u>Figure 2.5</u>	Diagrammatic representation of the scope of bioinformatic tools available for each step of WES data analysis	11
<u>Figure 2.6</u>	Dopamine levels in a normal vs. a PD-affected neuron	15
<u>Figure 2.7</u>	Motor and non-motor symptoms present in PD-affected individuals	15
<u>Figure 2.8</u>	The number of publications related to NGS and PD (indexed on PubMed) correlated to the cost of a single sequenced genome (2001-2021)	20
<i>Chapter 3</i>		
<u>Figure 3.1</u>	Outline of the methodological approach for Aim 1 of the PhD project	26
<u>Figure 3.2</u>	Pedigree of Xhosa family ZA 15	28
<u>Figure 3.3</u>	WES data analysis workflow used for the analysis of family ZA 15	29
<u>Figure 3.4</u>	Workflow for variant filtering to eliminate known causes of PD and identify novel candidate variants	32
<u>Figure 3.5</u>	Summary of the main findings of the present study based on WES analysis of ZA 15	58

Chapter 4

<u>Figure 4.1</u>	A brief overview of the methodology for in silico analysis of the p.A13V variant in MANF	63
<u>Figure 4.2</u>	Secondary structure prediction of the MANF protein using the PredictProtein server	70
<u>Figure 4.3</u>	Simplified structure of a signal peptide	70
<u>Figure 4.4</u>	Signal peptide secondary structure prediction for the MANF protein using the Phobius server	71
<u>Figure 4.5</u>	Signal peptide secondary structure prediction for the MANF protein using the SignalP 3.0 server	72
<u>Figure 4.6</u>	AA quality plots generated for the theoretical protein structures by Verify3D	75
<u>Figure 4.7</u>	Q-MEAN PDB structure comparison plots generated for the theoretical protein structures	75
<u>Figure 4.8</u>	Q-MEAN local similarity plots generated for the theoretical protein structures	76
<u>Figure 4.9</u>	Ramachandran plots generated using ProCheck	77
<u>Figure 4.10</u>	Error values calculated for each AA using the ERRAT server	78
<u>Figure 4.11</u>	Output depicting the $\Delta\Delta G$ and molecular flexibility of the p.A13V variant on the wildtype structure using Dynamut	79
<u>Figure 4.12</u>	Comparing the RMSD between the wildtype and variant structure	81
<u>Figure 4.13</u>	Comparing the root mean square fluctuation between the wildtype and variant structure	82
<u>Figure 4.14</u>	PCA of the wildtype and variant models of MANF	83
<u>Figure 4.15</u>	Summary of the results obtained through in-silico analysis of p.A13V in MANF	85

Chapter 5

<u>Figure 5.1</u>	Model for the translational cycle of a secretory protein (MANF)	93
--------------------------	--	-----------

VI. List of Tables

		Page
Chapter 2		
<u>Table 2.1</u>	Genes found to be associated with PD and non-PD Parkinsonism	18-19
Chapter 3		
<u>Table 3.1</u>	Pre-screening results for known PD genes in the proband (ID 43.59/74.53) using various mutation screening techniques	37
<u>Table 3.2</u>	Population database MAFs and pathogenicity prediction scores for <i>LRRK2</i> p.E889D	38
<u>Table 3.3</u>	Summary statistics produced for each sample that was whole exome sequenced (BAM and gVCF files)	39
<u>Table 3.4</u>	Non-synonymous, exonic variants found in the known/putative PD genes that are shared by the affected individuals (74.53 and 74.54) only	39
<u>Table 3.5</u>	In silico pathogenicity prediction scores across the prioritised 24 variants with a CADD score > 20 and expressed in neuro-specific tissue	42
<u>Table 3.6</u>	MAF population frequencies, in public databases, for the 24 prioritised variants with a CADD score > 20 and expressed in neuro-specific tissue	43
<u>Table 3.7</u>	Sanger sequencing of the 24 prioritised variants in the proband (ID 43.59/74.53)	46
<u>Table 3.8</u>	Co-segregation of the variants in the ZA 15 family members	49
<u>Table 3.9</u>	Variant screening in various 'non-PD' private population cohorts	52
<u>Table 3.10</u>	Variant screening in various PD-specific private population cohorts	54
<u>Table 3.11</u>	Comparison of the gene expression profiles and functional processes for each gene of interest	56

Chapter 4

<u>Table 4.1</u>	Predicted effect of the p.A13V variant on the MANF protein using MuPro and I-Mutant 3.0	69
<u>Table 4.2</u>	Structural alignment and validation of the generated models of wildtype and mutant MANF protein	74
<u>Table 4.3</u>	$\Delta\Delta G$ scores for the p.A13V variant on the wildtype protein structure using DUET/SDM/mCSM	79



Table of Contents

Declaration	I
Acknowledgements	II
Abstract	III
List of Abbreviations	IV
List of Figures	V
List of Tables	VI
CHAPTER 1	1
1.1 Brief summary of the literature	1
1.2 Rationale of the PhD research project	2
1.3 Aims and objectives of the PhD research project	2
Aim 1: To identify novel pathogenic variants and/or genes for PD in a South African Xhosa family affected with familial PD using WES and bioinformatic analyses	2
Aim 1 objectives	2
Aim 2: To perform <i>in silico</i> analysis on the prioritised variant to determine if the functional effect on the protein could be pathogenic	3
Aim 2 objectives	3
1.4 Dissertation overview	3
Chapter 1: Rationale of the Dissertation	3
Chapter 2: Literature Review	3
Chapter 3: Whole exome sequencing analysis of a South African Xhosa family affected with Parkinson's disease	3
Chapter 4: <i>In-Silico</i> mutation analysis of p.A13V in mesencephalic astrocyte-derived neurotrophic factor (MANF)	4
Chapter 5: Discussion and Conclusions	4
CHAPTER 2	5
2.1 The advent of next-generation sequencing (NGS)	5
2.1.1 What is NGS?	5
2.1.2 Comparing the types of NGS in complex disease research	7
2.2 Bioinformatics analysis of NGS data in complex disease research	9
2.2.1 NGS analysis steps	9
2.2.2 NGS analysis using bioinformatic tools	10
2.2.3 Functional analysis of prioritised variants	12
2.3 Disease under investigation: Parkinson's disease (PD)	14
2.3.1 Clinical features of PD	14
2.3.2 Diagnosis and treatment	16

2.4 PD genetics	16
2.4.1 Identified causes of PD	16
2.4.2 The discovery of the established PD genes	18
2.4.3 Strategies for the discovery of novel PD genes and susceptibility factors	19
2.5 The state of PD research in sub-Saharan Africa (SSA)	20
2.5.1 Strategies for novel PD gene discovery in SA	21
Significance of study	22
CHAPTER 3	23
Abstract	23
3.1 Introduction	25
3.2 Methods and materials	25
3.2.1 Study participants (South African Xhosa family - ZA 15).....	27
3.2.1.1 Ethical considerations for study	27
3.2.1.2 Selection criteria for study participants	27
3.2.1.3 Pedigree of South African Xhosa family (ZA 15) affected with Parkinson’s disease	28
3.2.2 WES and data analysis pipeline	28
3.2.2.1 WES	28
3.2.2.2 WES data analysis workflow	29
3.2.3 Variant filtering	32
3.2.3.1 Filter VCF for heterozygous SNPs that are shared between the affected siblings.....	32
3.2.3.2 Filter VCF for known and putative PD genes	33
3.2.3.3 Filter VCF for novel potential disease variants	33
3.2.4 Gene expression and pathway analysis	33
3.2.4.1 Tissue expression	33
3.2.4.2 Pathway analysis	33
3.2.4.3 Gene-disease association	34
3.2.4.4 Gene/protein interactions	34
3.2.5 Variant screening using wet-laboratory techniques	34
3.2.5.1 Sanger sequencing	34
3.2.5.2 Screening of variants in private cohorts	36
3.3 Results	36
3.3.1 Study participants (South African Xhosa family - ZA 15).....	36
3.3.1.1 Descriptive overview of the family	36
3.3.1.2 Pre-screening of the proband resulted in no variant/s of significance in known or putative PD genes	37
3.3.2 WES analysis was performed on 4 individuals in ZA 15	38

3.3.2.1 Summary statistics of BAM and VCF files depicted a high rate of sequence alignment and read quality	38
3.3.3 Variant filtering and prioritisation yielded 68 variants of interest	40
3.3.3.1 Gene expression and pathway analysis revealed that 24 variants were expressed in the brain	40
3.3.3.2 In-silico pathogenicity prediction scores found 13 variants to be pathogenic across >5 pathogenicity prediction tools	41
3.3.4 Co-segregation analysis in family ZA 15	45
3.3.4.1 Sanger sequencing revealed that 20 variants co-segregated in family ZA 15	45
3.3.5 Allele frequencies of variants in private PD and non-PD cohorts	50
3.3.5.1 Private cohort screening further reduced the number of candidates to 3 variants	50
3.3.6 Prioritising a single variant in family ZA 15 for further <i>in-silico</i> protein analysis resulted in the nomination of the p.A13V variant in MANF	55
3.4 Conclusion	57
CHAPTER 4	59
Abstract	59
4.1 Introduction	61
4.2 Methods and materials	62
4.2.1 Dataset identifiers	64
4.2.2 Secondary structure analysis	64
4.2.2.1 Phylogenetic analysis of the protein sequence	64
4.2.2.2 Functional and stability effects prediction analysis	64
4.2.2.3 Protein domain analysis	65
4.2.3 Tertiary structure analysis	65
4.2.3.1 Ab initio structural modelling of the wildtype and variant proteins	65
4.2.3.2 Structural validation of protein models	65
4.2.3.3 Determining the effect of the variant on the wildtype structure	66
4.2.3.4 Comparing the theoretical wildtype and mutant structures	66
4.3 Results	68
4.3.1 Secondary structure analysis	68
4.3.1.1 Phylogenetic conservation analysis reveals that p.A13V is a buried residue with variable conservation	68
4.3.1.2 Functional and stability effects prediction analysis indicate that the variant is situated in the signal peptide and is destabilising	68
4.3.1.3 Protein domain analysis confirmed the presence of the p.A13V variant in the hydrophobic core of the signal peptide	70
4.3.2 Tertiary Structure Analysis	73
Part 1 (Ab Initio Protein Modelling and Validation of the MANF Protein)	73
4.3.2.1 Ab initio structural modelling of the wildtype and variant protein introduced the signal peptide domain onto the protein structures	73

4.3.2.2 Structural validation of protein models indicated that Robetta produced the highest quality theoretical structures	73
Part 2 (Mutation Analysis of the Theoretical Protein Structures)	78
4.3.2.3 Determining the effect of the mutation on the wildtype structure indicated that the variant is destabilising	78
4.3.2.4 Comparing the wildtype and mutant structures using Pymol and MD simulations indicated a difference in polarity, potential flexibility and conformation	80
4.4 Conclusion	84
CHAPTER 5	86
5.1 Understudied populations in PD genetic research	86
5.2 Main findings	88
5.3 Methodological approach in the present study	89
5.4 p.A13V in MANF as a candidate for PD	90
5.5 Limitations of the Study	95
5.6 Future work	96
5.7 Concluding remarks	97
References	98
Appendices	122
Appendix A: Ethics approval from Stellenbosch University	122
Appendix B: Ethics approval from the University of the Western Cape	123
Appendix C: Data Transfer Agreement between Stellenbosch University and the University of the Western Cape	124
Appendix D: Perspective article published in <i>Frontiers in Genetics: Neurogenomics</i> Section 131	131
Appendix E: List of known and putative PD genes obtained from literature	145
Appendix F: WES analysis steps, tools and commands	146
Appendix G: Pathway and gene expression analysis of the prioritised 24 variants with a CADD score > 20 and expressed in neuro-specific tissue	148
Appendix H: Primer sequences designed for Sanger sequencing	155
Appendix I: MANF analysis figures	157
Appendix J: ACMG classification of p.A13V in MANF	159

1.1 Brief summary of the literature

The study of human genetic variation through DNA sequencing has evolved significantly in recent decades and has allowed researchers to investigate the genetic mutations underlying many complex disorders (Muzzey *et al.*, 2015). Next-generation sequencing (NGS) technology has subsequently allowed for high-throughput, parallel sequencing that provides a quicker, cost-effective method for large-scale sequencing projects (Muzzey *et al.*, 2015). Prior to the advent of NGS, positional cloning and linkage analysis in large, multiplex pedigrees were used for co-segregation analyses to identify monogenic disease genes (Pang *et al.*, 2017). Genome-wide association studies (GWAS) followed and focused on single nucleotide polymorphisms (SNPs) or variants associated with a disease that was present in a population (Petersen *et al.*, 2017). However, these methods only identified a small number of disease genes, thus, the genetic cause of many diseases remained largely unknown (Petersen *et al.*, 2017). In the last decade, NGS, particularly in the form of whole exome sequencing (WES), has been particularly useful to detect novel mutations in families with disorders that depict Mendelian inheritance (Fernandez-Marmiesse *et al.*, 2017).

Parkinson's disease (PD) is a complex, though relatively prevalent, neurodegenerative motor disorder with a highly heterogeneous aetiology, although Mendelian forms of the disorder do exist (Alcalay *et al.*, 2020). According to many reports, only 5-10% of all PD cases can be attributed to established PD genes, indicating that many undiscovered genetic anomalies influence the onset of PD (Alcalay *et al.*, 2020). PD studies in sub-Saharan Africa have shown that most PD-affected individuals rarely have the common mutations or causative genes known to be implicated in PD (Williams *et al.*, 2018). Thus, it is hypothesised that the use of NGS approaches may potentially identify novel genetic causes underlying PD in African ancestry populations. This knowledge may lead to important biological insights that will ultimately improve the diagnosis and treatment of this disorder.

Although NGS approaches have been successful in the study of multi-genic diseases, the overwhelming majority of genomic research has been limited to European and Asian populations (Schoonen *et al.*, 2019). Notably, this monopolisation of scientific information can skew the inferences made about genetic disorders. African ancestry populations harbour an abundance of genetic diversity, and genomic research targeted at these populations may provide insight into the 'missing heritability' of many rare or complex disorders (Bentley *et al.*, 2020).

1.2 Rationale of the PhD research project

This study aims to identify and use, an appropriate bioinformatic approach for NGS analysis of South African individuals affected with familial PD, thereby allowing the analysis to be effective and reproducible. This study also aims to fill a knowledge gap regarding the genetic architecture underlying the disease in South African PD families. Results from this study could provide insight into novel genetic factors that instigate PD onset and progression in individuals from under-researched ethnic backgrounds, but also in global PD populations. Further functional study into novel candidate genes could lead to a formative basis for newer or targeted therapeutic modalities through understanding and manipulating the mutational effects on biological targets associated with the disease. Ultimately, it is a goal that the use of NGS could lead to precision or stratified medicine where the treatment and prevention of a particular disease is optimised by considering individual variability at the genetic level.

1.3 Aims and objectives of the PhD research project

Aim 1: To identify novel pathogenic variants and/or genes for PD in a South African Xhosa family affected with familial PD using WES and bioinformatic analyses

Aim 1 objectives

1. Analyse previous candidate gene screening data of the proband to eliminate known causes of PD
2. Perform WES on the prioritised members of the family
3. Create a custom WES analysis workflow to analyse the generated WES data
4. Formulate a stringent variant filtering approach to identify novel, potentially disease-causing variants
5. Screen the prioritised variants through ancestry-matched and PD-specific cohorts to determine allele frequencies
6. Perform Sanger sequencing to confirm NGS results and co-segregation of variants within the family
7. Identify the top variant/s for further functional/*in silico* analysis

Aim 2: To perform *in silico* analysis on the prioritised variant to determine if the functional effect on the protein could be pathogenic

Aim 2 objectives

1. Perform conservation analysis on the protein of interest
2. Perform secondary structure and stability prediction analysis to determine the effect of the variant on protein stability
3. Create theoretical 3-dimensional models of both the wildtype and mutant protein using webservice-based tools
4. Validate the generated wildtype and mutant protein models
5. Determine the effect of the variant on the wildtype structure
6. Perform molecular dynamics (MD) simulations on the wildtype and mutant models to determine the impact of the variant on the generated protein structures

1.4 Dissertation overview

This dissertation is divided into 5 chapters and is written in British English. All bioinformatic analysis and wet-laboratory experiments pertaining to this study was done by the PhD candidate.

Chapter 1: Rationale of the dissertation

Chapter 2: Literature review

The literature review encompasses the advent of NGS technology and its impact on disease genomics as well as the current state of PD genetics. We explore how the use of WES in families affected with PD has enabled researchers with an efficient method of determining novel variants or genes underlying the pathobiology of the disease. Furthermore, the best-practice bioinformatic tools and WES analysis methods are explored, as well as the implications of determining the potential causes of PD in an understudied population such as the one examined in our study.

Chapter 3: Whole exome sequencing analysis of a South African Xhosa family affected with Parkinson's disease

Chapter three describes a South African family of Xhosa ancestry (designated family ZA 15) who are affected with PD, where selected family members underwent WES. The WES analysis, co-segregation analysis and subsequent population screening of selected variants, are described in

detail. The results of the analysis culminated in the prioritisation of three variants of interest, from which one was selected for further study.

Chapter 4: *In-silico* mutation analysis of p.A13V in mesencephalic astrocyte-derived neurotrophic factor (MANF)

Chapter four describes the prioritisation of a single variant for *in silico* mutation analysis based on its potential link to PD. This chapter focuses on conservation analysis of the amino acids, secondary structure analysis of the implicated protein, stability effect prediction of the variant using the protein sequence, protein domain analysis, theoretical modelling of the wildtype and mutant protein, as well as variant effect analysis on the theoretical models. These findings serve as a precursor to ‘wet-lab’ functional studies to determine whether the variant is the cause of disease within family ZA 15.

Chapter 5: Discussion and conclusions

The final chapter explores the implications of the findings and how the variant could potentially lead to a PD phenotype. Furthermore, the limitations of the current study are laid out and recommendations for future work based on our findings and the direction of PD genetic research are elucidated.



2.1 The advent of next-generation sequencing (NGS)

The ability to decipher the genomic assembly of organisms, particularly humans, has accelerated the progress of biomedical research exploring the genetic basis of disease. Simply, DNA sequencing is the task of using biochemical methods and sequencing machinery to determine the actual order of nucleotides within DNA (Giani *et al.*, 2020). Various sequencing methods have expanded research efforts to include population-wide genomics, mapping, diagnostics of genetic disease, and ultimately, the strive towards personalised medicine (Kulski, 2016). Over time, sequencing technologies have progressively developed to include next-generation, high-throughput approaches that perform highly efficient and increasingly accurate methods of sequencing.

2.1.1 What is NGS?

NGS is the deep sequencing technology that allows for the parallel sequencing of millions of short DNA fragments or reads, and additionally promises a lower cost, faster output, and higher throughput of sequencing as compared to traditional ‘first generation’ sequencing methods (Kulski, 2016; Gutierrez-Rodrigues and Calado, 2018). In simplified terms, the process of NGS consists of four fundamental steps, namely, (1) nucleic acid extraction, (2) library preparation, (3) sequencing and, ultimately, (4) computational analysis of the sequencing output data (**Figure 2.1**) (Fernández-Marmiesse *et al.*, 2018; Giani *et al.*, 2019).

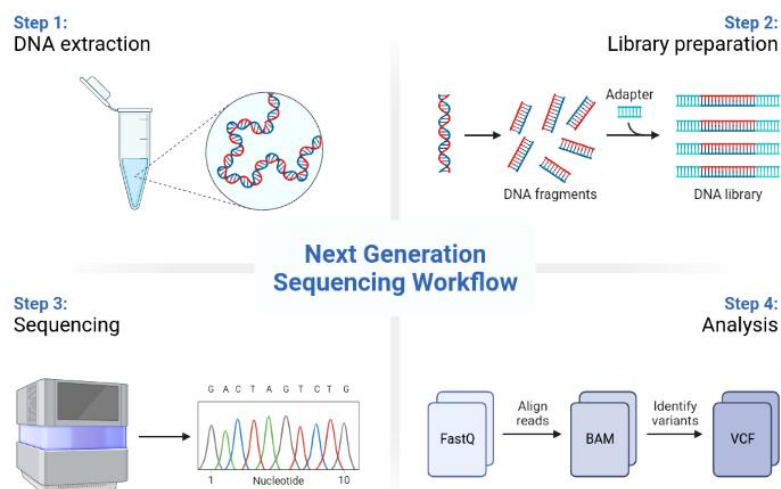


Figure 2.1: Overview of a next-generation sequencing workflow
Created in Biorender.com.

For human-based genomic research, blood or saliva samples are collected from individuals of interest and genomic DNA is then isolated and purified before sequencing. Thereafter, DNA libraries are prepared by shearing the DNA into short fragments and subsequently attaching specialised adapters to either end of the fragments (Behjati and Tarpey, 2013). These adapters are complementary and capable of binding to the NGS flowcell, a substrate where the fragments are immobilised and amplified in parallel in the sequencer (Pereira *et al.*, 2020). Each reaction involves the stepwise incorporation of fluorescently-labelled nucleotides that are attached to the flowcell (Giani *et al.*, 2020). NGS sequencing instruments can convert raw sequencer-generated DNA signals into usable data files containing the correct sequence of nucleotide bases with a corresponding base quality score. This data is then analysed using bioinformatic tools to compile genomic information about the organism being investigated (Oliver *et al.*, 2015).

NGS has allowed for higher sequencing depth and sensitivity, higher novel variant discovery power and, the ability to identify larger mutations and produce larger volumes of data with the same amount of input DNA as compared to Sanger sequencing (Saier, 2019). The triumphant initial sequencing of the human genome consisting of ~3 million bp (using Sanger sequencing) took approximately 14 years whereas, currently it takes about 1-2 days using existing NGS technology (Barba *et al.*, 2014). The drastic improvements in sequencing afforded through the development of NGS have enabled efficient and exponentially accelerated genomic-based research as the cost of sequencing the human genome has decreased significantly (illustrated in **Figure 2.2**).

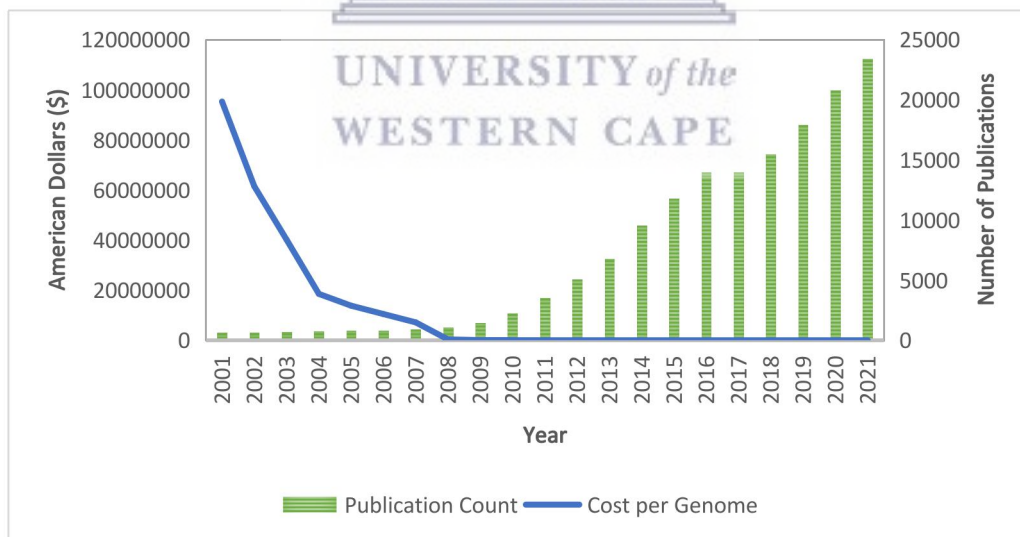


Figure 2.2: The number of publications related to NGS (indexed on PubMed) correlated to the cost of a single sequenced genome (2001-2021)

The PUBMED string search was performed on the 29th of August 2022 and was adapted from Su *et al.*, 2011, and consisted of ('next-generation sequencing' OR 'next-generation sequencing' OR 'next-generation DNA sequencing' OR 'next-generation DNA sequencing' OR 'massively parallel sequencing' OR 'ultrafast DNA sequencing' OR "454 sequencing" OR "deep sequencing") AND (2000[Publication Date]:2021[Publication Date]). The genome cost data (2001 to 2021) was accessed from NHGRI with permission.

2.1.2 Comparing the types of NGS in complex disease research

Briefly, NGS technology includes three sequencing approaches, namely; whole exome sequencing (WES), whole genome sequencing (WGS) and targeted gene panels. For complex Mendelian diseases, these sequencing techniques are typically used to identify variants of significance in both a clinical diagnostic and genetic research setting (Petersen *et al.*, 2017). A comparison of these NGS approaches is provided in **Figure 2.3**.

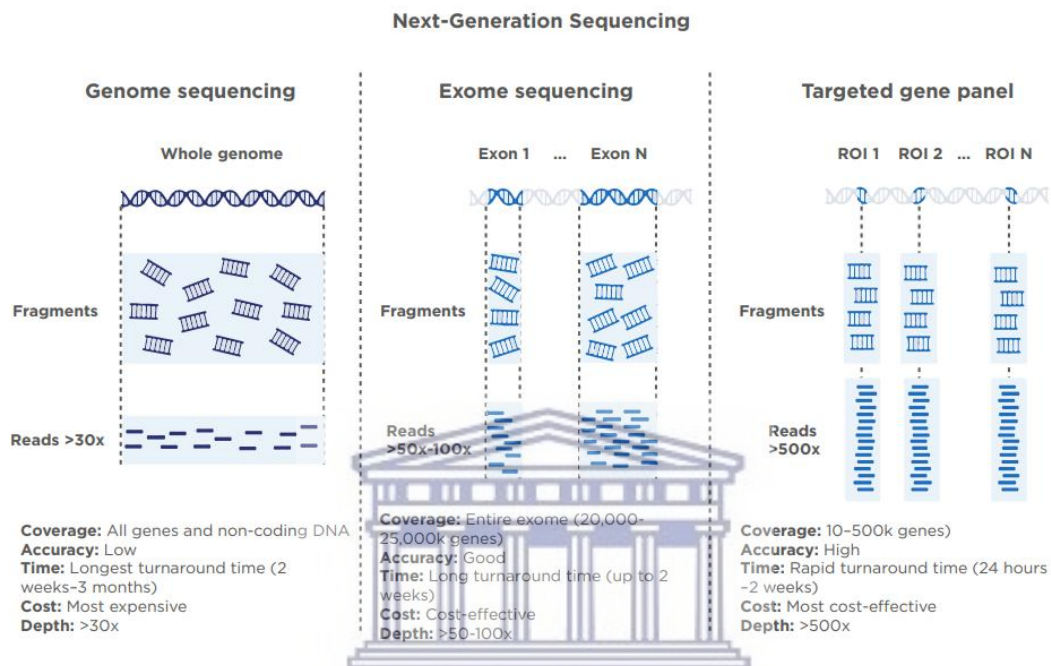


Figure 2.3 A comparison between whole genome sequencing, whole exome sequencing and targeted gene panels

Permission obtained under the Creative Commons Attribution Non-Commercial License 4.0; Duraes *et al.*, 2022.

For clinical diagnostics, the screening of known disease genes is useful for the accurate diagnosis of a particular disorder, allowing for a better prognosis for the individual. This can enable clinicians to be better informed regarding therapeutics as some specific forms of genetic disease respond better to certain treatments that directly target the underlying mechanisms. Targeted gene panels comprise of known disease genes (genes that have already been associated with a particular disease phenotype) that allow for the rapid, accurate identification of specific genetic causes of the disease in an affected individual (Shulskaya *et al.*, 2018). NGS data (e.g., WES data) of the affected individual is filtered through the targeted gene panel containing these disease-relevant genes to confirm the presence of a known variant in included disease genes or potentially find new pathogenic variants in those genes (Reale *et al.*, 2018; Shademan *et al.*, 2021). These gene panels are cost-efficient and provide deeper gene coverage when compared to WES or WGS, thus increasing the likelihood of detecting mutations

in the included genes (Brunelli *et al.*, 2019). These panels can also be updated based on novel disease-gene findings through NGS-based disease research (Reale *et al.*, 2018). However, many Mendelian disorders can be heterogeneous or polygenic, with the underlying genetic causal factors remaining largely undetermined. Targeted gene panels can be useful for the discovery of novel variants or genomic rearrangements in known disease genes though not for the discovery of aberrations in novel genes. Studies using WES or WGS can observe both common and rare variants across an entire genome, thus, making their use optimal for novel gene discovery in rare or complex disease.

WGS, being the more comprehensive choice, cumulatively sequences the entire exome, as well as the intergenic non-coding regions and mitochondrial DNA, resulting in approximately 4,000,000 variants for a single-sequenced human genome (Fernández-Marmiesse *et al.*, 2017). This method also displays relatively even genomic coverage which is necessary for copy number evaluation. However, WGS remains the more expensive method (requiring more sequencing reagents) with an overall lower accuracy than the other NGS methods. WGS also produces a significant amount of sequencing data that requires sufficient storage space and time, as WGS analysis tends to be computationally intensive (Park and Kim, 2016).

NGS, in the form of WES, sequences only the protein-coding (exonic) portion of the individual's genome (1 - 5% of the entire genome) and can provide ~ 20000 - 25000 variants for each exome sequenced (Fernández-Marmiesse *et al.*, 2017; Shulskaya *et al.*, 2018). Notably, as disease-causing variants may be found in the intronic, 5' UTR and 3' UTR regions of the genome, solely using WES may be limiting (Belkadi *et al.*, 2015). WES can also result in skewed coverage due to hybridisation biases, making allele frequency and copy number assessments challenging (Brunelli *et al.*, 2019). However, WES is typically considered the better choice for determining monogenic causes of Mendelian disease. This is because most pathogenic variants (80 - 85%) have been found in the coding region of the genome, there is better coverage of the coding variants (SNVs and indels), it is more cost-effective (at ~20% of the WGS cost) and the sequenced data load is significantly reduced thus allowing for easier computational analysis (Fernández-Marmiesse *et al.*, 2017). WES has been successfully utilised to determine the genetic causes of Mendelian diseases using large families, containing at least two disease-affected family members. Sensitive bioinformatic analyses and co-segregation analysis of familial WES data provide an effective method of reducing the number of relevant variants that may be disease-causing (Shulskaya *et al.*, 2018). Thus, this method of analysis is thought to be more effective for complex disease research targeting affected families in developing countries and it is the NGS approach used in the present study.

2.2 Bioinformatics analysis of NGS data in complex disease research

Over a relatively short time, biomedical research has exponentially geared towards the generation of large datasets (particularly in the form of molecular and genomic data) to unveil new genetic information about disease (Behjati and Tarpey, 2013). However, with the rise in biological data availability, appropriate bioinformatic analysis of this data is considered a significant rate-limiting step for NGS technology. It has now become essential for genomic researchers to be acquainted with the adequate computational tools and skills that allow for robust data assimilation and interpretation (Behjati and Tarpey, 2013).

2.2.1 NGS analysis steps

The final step of the NGS workflow is the analysis of the output data. In the case of WES and WGS data, the basic analytic workflow can be constituted into primary, secondary and tertiary analyses that allow for the prioritisation of genetic factors that may be implicated in the disease under investigation (**Figure 2.4**).

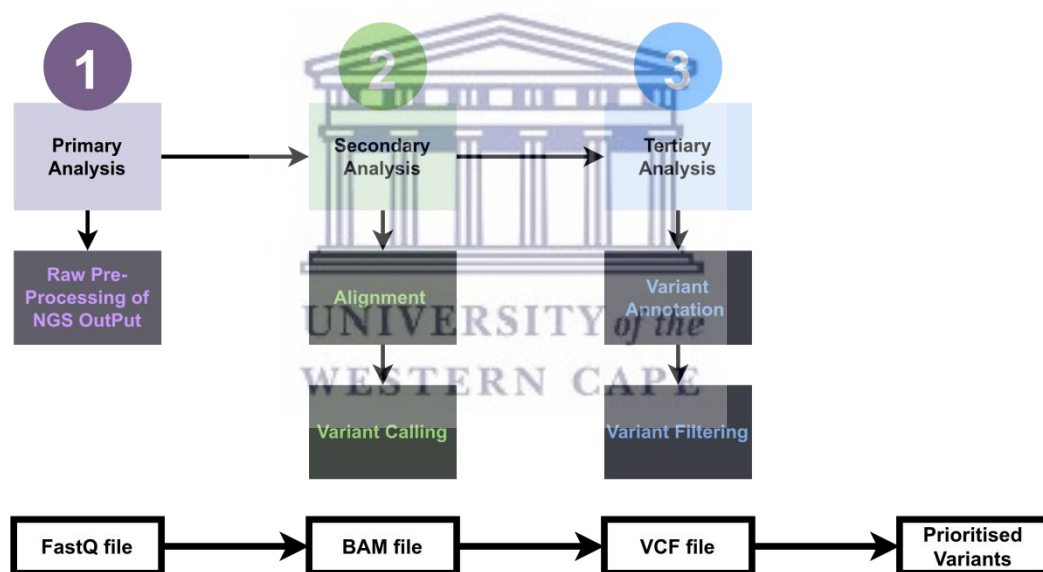


Figure 2.4: Basic steps for NGS analysis

Created in Draw.io.

Primary analysis of the NGS output data is a quality control stage of the workflow. Raw FASTA or FASTQ files include base-calling data with associated quality scores (Phred scores). In this step, bases with low-quality scores are effectively trimmed off. Furthermore, adaptor sequences that are still attached from the sequencing step can be identified and trimmed off, concurrently (Schubert *et al.*, 2019).

Secondary analysis first involves the alignment of the ‘clean’ reads to the latest human reference genome where the alignment output can either be a sequence alignment map (SAM) or binary alignment map (BAM) file. These files contain all the information of the FASTQ file as well as read alignment positions, alignment quality scores and the degree to which the reads matched the reference. Thereafter, the BAM files undergo further processing to remove potential PCR duplicates. Once the BAM files are ready, the aligned reads are then examined to identify positions where the individual possesses genetic variation and to determine the type of variation. This outputs a variant call file (VCF) that, in the case of WES, specifies the genomic position of single nucleotide variants (SNVs) and indels, the statistical probability of the variant call and the accompanying quality score (Bartha and Györfy, 2019).

Tertiary analysis involves the annotation of the VCFs, allowing for an additional layer of information that can be used to later reduce the number of candidate variants. Annotation allows for each called variant to be labelled according to gene-based and functional traits including transcriptional gene regulation, alternative splicing, protein function modifications and evolutionary conservation (Austin-Tse *et al.*, 2022). This step can help describe the variant’s clinical significance, its impact on protein function and the frequency in which the variant is expressed in healthy individuals. Finally, the annotated VCFs are filtered using study-specific criteria to obtain a shorter list of the most significant variants (candidate variants) by eliminating variants according to population frequency, predicted pathogenicity, co-segregation, and further functional filters (Austin-Tse *et al.*, 2022). However, the quality of the final list of variants output after these analytic steps can be highly dependent on the bioinformatic tools used to perform these steps.

2.2.2 NGS analysis using bioinformatic tools

Currently available bioinformatic tools have subtle variations in output, and it is thus imperative to use the best software available that is also appropriate and suited to the disease under investigation. The challenge lies in sifting through the ever-increasing number of software that can be used for each step of NGS analysis (**Figure 2.5**). This is further supported by a study, in which 7 years after the initial analysis, the re-analysis of WES data resulted in an increased diagnostic yield in terms of variants implicated in rare, idiopathic disease cases (Salfati *et al.*, 2019). This improvement was attributed to the use of improved variant classification tools and the use of updated databases (Salfati *et al.*, 2019). This indicates that sequencing technologies tend to develop faster than the tools needed to analyse the assemblies, indicating a rate-limiting step that may hinder the quality and subsequent momentum of scientific findings using these technologies.

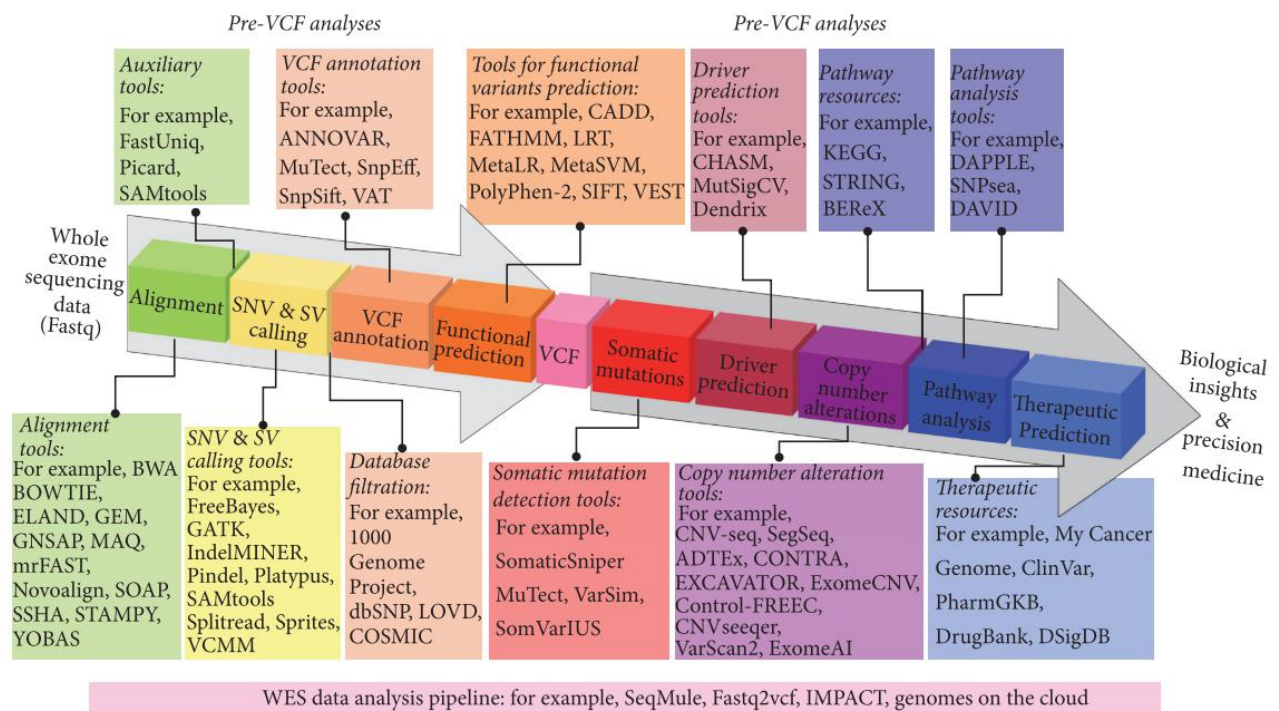


Figure 2.5: Diagrammatic representation of the scope of bioinformatic tools available for each step of WES data analysis

Used with permission from Hindawi International Journal of Genomics under the Creative Commons Attribution License International 4.0; (Hintzsche *et al.*, 2016).

However, an issue arises with the array of tool choices that are available for each step of WES analysis (as seen in **Figure 2.5**). A common encountered issue is the rapid turnover of new tools that are not frequently updated or consistently available, making the reproducibility of WES analysis pipelines virtually impossible. Thus, certain tools tend to be regarded as best-practise choices due to their efficiency, continual updates and availability, and relevance to the type of data or disease being investigated. When it comes to the analysis of WES in complex disease research, the Burrows-Wheeler Aligner BWA-MEM algorithm is commonly employed. However, one study emphasised the importance of using the latest reference genome (GRCh38 over GRCh37) for the alignment, concluding the choice of the aligner is not as important for enhanced genome coverage (Pan *et al.*, 2019).

There are many important caveats to consider when carrying out data analysis on a specific disease or disorder. In the case of rare or complex genetic disease, the experimental design of secondary analysis can greatly improve variant calling by incorporating relevant assumptions. It is important to take into consideration the genetic relationship between related individuals during variant calling, the possible modes of inheritance (autosomal recessive, autosomal dominant, X-linked, and *de novo*), population stratification which needs to be accounted for where variants of interest may be

specific to a particular population group (as opposed to disease-causing) and the locus heterogeneity in complex disease (Hintzsch *et al.*, 2016). The popular choice of variant caller is the Genome Analysis toolkit (GATk) HaplotypeCaller. This is because the HaplotypeCaller has consistently shown the best performance in terms of calling variants from both high and low-depth coverage with the fewest erroneous calls (Andreu-Sánchez *et al.*, 2021). This is particularly important in the study of Mendelian disease where the quality of the genotype call determines the quality of candidate variants. Multiple mini pipelines are readily available for secondary analysis of NGS data however, many lack software that is optimised to deal with heterogeneous rare diseases or genetically diverse, under-represented population groups (Schoonen *et al.*, 2019). Thus, the need for secondary analysis pipelines to be compared or optimised is becoming a necessary part of the analysis, especially with the development of population-specific genomes and upgraded variant caller software.

Tertiary analysis allows for the functional labelling and subsequent filtering of the variant data obtained after secondary analysis. Multiple biological factors surrounding the disease need to be accounted for to reduce the output of clinically irrelevant data (Davis-Turak *et al.*, 2018). Thus, it is imperative to choose an annotator that is consistently updated with the multiple databases that are used to classify the variants. Ensembl's Variant Effect Predictor (VEP) and Annovar tend to be popular choices for the annotation of variants due to their comprehensive selection of biological databases and their consistency with the use of the latest version of the databases used (Cunningham *et al.*, 2022). Specifications that need to be accounted for in tertiary analysis are highly dependent on the molecular understanding of the variant, its genetic interactions and its influence on the biological mechanisms underlying the disease. Thus, the choice of software and the stringency of selected parameters used for secondary and tertiary analysis can 'make or break' the process of novel gene discovery using NGS data.

Importantly, there are several shortfalls to using established bioinformatic pipelines for the data analysis of South African (SA) disease-affected individuals, since most software has been primarily developed for the study of European- or Asian-based datasets (Bentley *et al.*, 2020). This problem primarily occurs during the variant annotation process and recently SA researchers have begun to take this into account (Schoonen *et al.*, 2019). Schoonen *et al.* (2019) incorporated Ensembl-VEP to annotate variants and GENome MINing (GEMINI v0.20) to effectively filter variants according to African allele frequencies, resulting in higher quality output.

2.2.3 Functional analysis of prioritised variants

A common problem encountered when attempting to determine a novel genetic cause of disease is the return of prioritised variants of uncertain significance (VUSs) (Federici and Soddu, 2020). This can be, in part, due to the limited research available on the gene's expression and involvement in

biological mechanisms, and the lack of knowledge regarding the effect of the change on the protein structure and function, thus resulting in the inability to explicitly link the phenotype to the variant. In 2015, the American College of Medical Genetics formed the ‘ACMG guidelines’; a set of stringent criteria incorporating various aspects of a variant’s information including its population data, computational data, functional data and segregation data to classify a variant as either ‘pathogenic’, ‘likely pathogenic’, ‘uncertain significance’, ‘likely benign’, or ‘benign’ (Richards *et al.*, 2015). However, if there is limited information on the variant present in the databases used, the variant remains one of uncertain significance.

Prior to NGS-based analysis, researchers were confined to ‘wet lab’-based functional analysis of a variant of interest. However, this is no longer the most efficient approach due to the influx of VUSs as potential causes of disease, owing to the accessibility of NGS technology (Kwong *et al.*, 2021). Therefore, before time-consuming and expensive methods of functional analyses, it has become easier to potentially determine the functional effect of a variant on a protein using computational approaches first, thus determining whether further functional analysis is needed. Initially, this was a difficult process due to the limited number of experimentally solved protein structures (Waterhouse *et al.*, 2018). However, the development of computational *ab initio* protein modelling that can mimic the qualities of a solved structure using just the protein sequence, as seen with DeepMind’s Alphafold, an initiative to solve the complete protein structure of all proteins (Jumper *et al.*, 2021). Thus, determining the potential functional, structural, and biological effect of a variant on any protein has become easier and necessary, prior to further downstream analysis.

A well-developed analysis workflow is necessary for the optimal analysis of NGS data. The workflow must be capable of thorough data quality control and sensitive variant filtration. This is to reduce the probability of error through the prevalence of false positive/negative variants by using optimum bioinformatic tools and their subsequent parameters (Bayrak and Itan, 2020). Ultimately, the workflow should also be able to prioritise the most-likely disease variants/genes based on a variety of study-specific factors. These include the ethnicity of the population group, co-segregation of the variants within the family and, ultimately, a molecular understanding of the nominated gene/s and disease in question, thereby solidifying the importance of consecutive *in silico* analysis (Pereira *et al.*, 2020). In the present study, we aim to use WES approaches (in conjunction with specialised bioinformatic techniques) and *in silico* pathogenicity analysis to identify potential novel PD variants/genes in a South African family.

2.3 Disease under investigation: Parkinson's disease (PD)

PD is a progressive movement disorder that was first formally recognised in 1817 by Dr James Parkinson and medically described as the 'Shaking Palsy' (Goetz, 2011). This complex neurodegenerative disorder currently affects more than six million individuals, globally (Schneider and Alcalay, 2020). However, this number is considered to be closer to the range of 7-10 million, due to numerous geographical regions with limited statistics and inaccurate reporting of the disease (Selvaraj and Piramanayagam, 2019). Furthermore, the number of known PD cases are expected to rise to approximately 12 million by the year 2040, highlighting the necessity of genetic research into this perplexing disease (Dorsey *et al.*, 2018). Age is considered the largest risk factor of PD, followed by sex, where males are 1.5 - 2 times more likely to develop the disease (Bandres-Ciga *et al.*, 2020, Reekes *et al.*, 2020). The prevalence of PD is around 1% of individuals above the age of 60 years, and approximately 4-5% of individuals over the age of 85 years (Kalinderi *et al.*, 2016).

2.3.1 Clinical features of PD

The principal neuropathological hallmark of PD is the $\sim >70\%$ loss of nigrostriatal dopaminergic neurons (typically present in the *substantia nigra pars compacta* (SNpc) that results in a marked decrease in the amount of available dopamine over time (Figure 2.6) (Kalinderi *et al.*, 2016, Lunati *et al.*, 2019). Lewy bodies are aggregated protein clumps that may be found in the remaining neurons and are another important pathological hallmark of PD, however, it is unknown if these Lewy bodies provide a pathogenic or protective effect (Kalinderi *et al.*, 2016). Braak staging is typically used to determine the movement of Lewy body deposits which can be correlated to the severity of the PD diagnosis (Kouli *et al.*, 2018). In post-mortem studies of the transverse brain stem, loss of pigmentation in the SNpc is recognised as a distinct morphological change in PD-affected individuals. Other neural cell networks are also commonly implicated in PD pathology, adding to their heterogeneous nature and making a definite diagnosis of PD, challenging (Kouli *et al.*, 2018).

Clinical symptoms of PD were first described by Dr. James Parkinson in 1817 and further elucidated by Prof. Jean-Martin Charcot between 1868 and 1881 (Walusinski, 2017). Known motor symptoms consist of bradykinesia, rigidity, resting tremor, postural instability and freezing, however, years prior to the onset of these symptoms, non-motor PD symptoms may manifest during an extended prodromal period (Kouli *et al.*, 2018). These symptoms can include both cognitive and behavioural symptoms, REM sleep disorders, gastrointestinal issues, depression, fatigue and autonomic or sensory dysfunction (Bandres-Ciga *et al.*, 2020, Kalinderi *et al.*, 2016) (Figure 2.7).

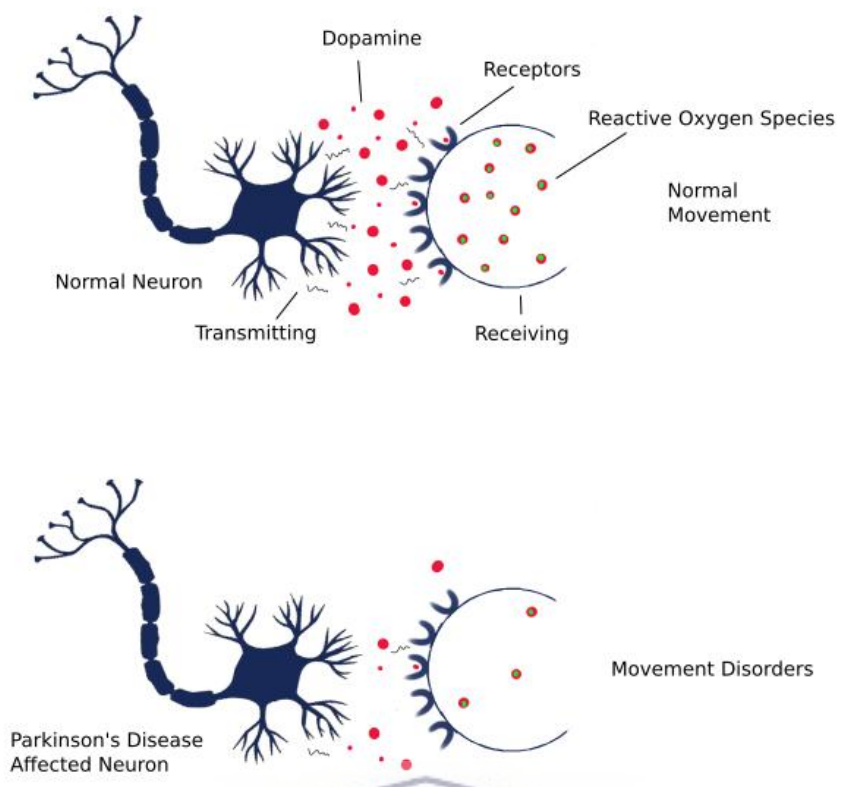


Figure 2.6: Dopamine levels in a normal vs. a PD-affected neuron
 Created using Krita (version 5.1.3) and Inkscape (version 1.2.1) software.

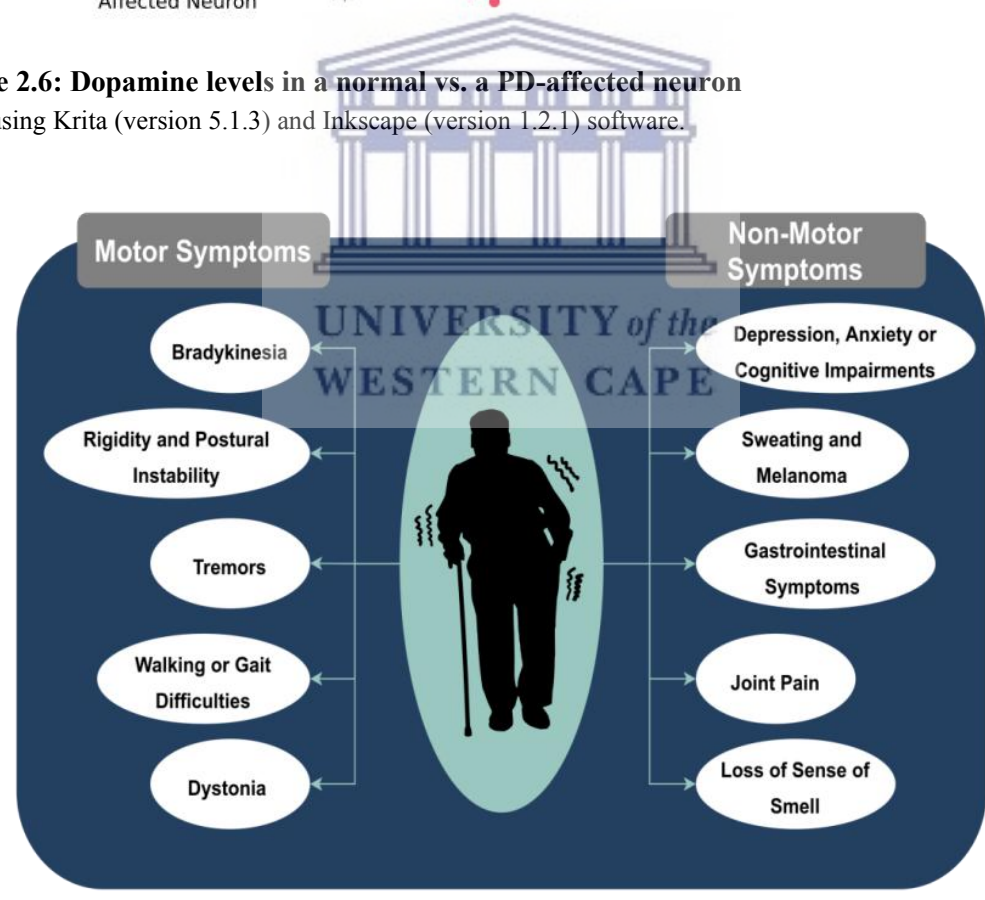


Figure 2.7: Motor and non-motor symptoms present in PD-affected individuals
 Image created using Draw.io and Inkscape (version 1.2.1) software.

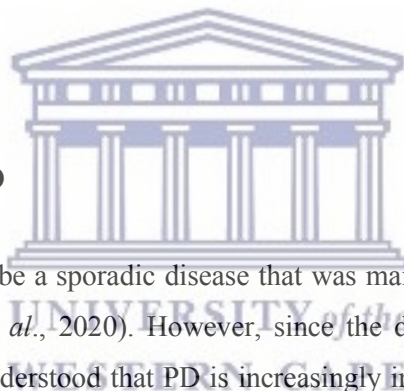
2.3.2 Diagnosis and treatment

Due to the heterogeneous neuropathology of the PD spectrum, the clinical criteria for diagnosis is based on the presence of two of the four common motor symptoms, namely, resting tremor, bradykinesia, rigidity and/or postural instability (Kouli *et al.*, 2018). Definitive diagnoses may require histopathological assessment of post-mortem brain samples to observe the presence of Lewy body deposits and loss of pigmentation in the SNpc. The use of targeted gene panels to confirm the monogenic cause of PD may be implemented if it can improve the treatment plan for the affected individuals (Oertel, 2016). Currently, PD has no known cure and existing treatments aim at alleviating symptoms of the disease, and include antidepressants, anti-tremor medications and cognitive enhancers, with the most common symptomatic treatment aimed at increasing the quantity of dopamine via L-dopa (Bandres-Ciga *et al.*, 2020). Surgical treatment, in the form of deep-brain stimulation or gastrointestinal levodopa implantations, may be encouraged if medicinal treatments do not prove effective (Oertel, 2016). Increasing the understanding of underlying PD genetic aetiology and subsequent biological pathway interferences may eventually lead to the implementation of personalised or targeted therapeutics in the future.

2.4 PD genetics

2.4.1 Identified causes of PD

Initially, PD was assumed to be a sporadic disease that was mainly influenced by environmental factors and age (Bandres-Ciga *et al.*, 2020). However, since the discovery of causal genes within PD-afflicted families, it is now understood that PD is increasingly influenced by genetic factors. The genetic aetiology of PD is quite diverse as it is considered a genetically heterogeneous disorder. PD cases are typically classified as having either familial or sporadic forms of the disorder (Schneider and Alcalay, 2020). Familial PD is rare, only accounting for ~ 2 - 15% of all PD cases. Approximately 5 – 10% of these familial PD cases have been known to reflect classic Mendelian inheritance patterns, whereby the disease is caused by highly penetrant variants (Lesage and Brice, 2009). However, the study of monogenic PD can be complicated by instances in which not all PD-affected family members carry the same pathogenic mutation and present as phenocopies (whereby two affected PD individuals with matching phenotypes in a family have different genotypes possibly due to an environmental risk factor). This phenomenon can easily be confused with intrafamilial heterogeneity (where one affected individual has a different mutation to the family mutation but where this difference maybe due to *de novo* mutations, epigenetic changes, pleiotropy or, in another instance, where multiple rare variants contribute to individual disease risk as seen in oligogenic inheritance (Klein *et al.*, 2011; Farlow *et al.*, 2016; Bentley *et al.*, 2021). True phenocopies in a family may also lead to incorrect conclusions



regarding the inheritance pattern within the family (Klein *et al.*, 2011). These confounding factors are relevant in PD, thus requiring adaptation of inclusion criteria in bioinformatic tools going forward. Established causative genes are subsequently split into three categories; autosomal dominant (AD), autosomal recessive (AR) and an X-linked form (Kalinderi *et al.*, 2016).

The vast majority (~ 85%) of all PD cases are considered sporadic, where the actual cause is unknown and onset is attributed to complex, synergistic interactions between genetic, metabolic, and environmental factors (Gasser, 2015). Epidemiological studies have deduced that certain environmental factors may impart protective effects against the development of PD, such as caffeine or alcohol intake and tobacco exposure (Hancock *et al.*, 2007; Zhang *et al.*, 2014). Adverse risk factors that may promote PD onset include exposure to certain pesticides, air pollution, well water consumption and even head injuries (Jankovic and Tan, 2006). Other factors like depression or gastrointestinal symptoms may also be associated with an increased risk for PD later in life (Bandres-Ciga *et al.*, 2020). However, it is proposed that environmental factors contribute fractionally to the risk of onset and thus, the underlying causative mechanisms for most PD cases remain largely unknown. Thus, newer sequencing technologies may prove insightful when attempting to decipher the ‘missing heritability’ of a disease like PD.

The underlying pathogenesis of PD has been linked to multiple biological mechanisms, including mitochondrial dysfunction, ineffective protein degradation, neuroinflammation and mostly, α -synuclein aggregation (Kouli *et al.*, 2018). Mitochondrial dysfunction is considered a key component in the pathogenesis of both idiopathic and familial PD. PD causal genes such as *PRKN* (an E3 ubiquitin ligase involved in the mitophagy pathway responsible for the removal of damaged mitochondria), *PINK1* (a ligase responsible for the phosphorylation of *PRKN* in the mitophagy pathway), and *DJ-1* (a ubiquitin ligase known to activate *PINK1* transcription) have been found to actively contribute to mitochondrial dysfunction leading to the onset of PD (Bonifati, 2014; Kouli *et al.*, 2018). Putative PD candidate genes including *FBXO7*, *PLA2G6*, *VPS13C* and *CHCHD2* have also been found to play roles in the quality control of mitochondrial systems (Bandres-Ciga *et al.*, 2020). In the case of ineffective protein clearance, the ubiquitin-proteasome system (UPS) has been implicated in PD due to its involvement in neural protein accumulation and deposits, allowing for the accumulation of Lewy bodies. The *PRKN* and *UCH-L1* genes have been linked to ubiquitin-proteasome system function (Mcnaught and Jenner, 2001). Furthermore, malfunctioning or differential expression of proteins involved in the lysosome/autophagic system have also been affected by mutations in genes implicated in PD such as *ATP13A2* and *GBA* (Kouli *et al.*, 2018). Postmortem brain analysis of sporadic PD individuals and most animal models with PD have displayed evidence of endoplasmic reticulum (ER) stress, particularly the upregulation of the unfolded protein response (UPR), a regulatory cascade that promotes homeostasis in the presence of misfolded

proteins or signals autophagy when experiencing chronic ER stress (Mercado *et al.*, 2016). Thus, pathway analysis has increasingly become an area of interest due to the not-yet-understood overlap of biological interactions contributing to the onset of PD.

2.4.2 The discovery of the established PD genes

Before NGS, researchers relied on methods such as chromosomal linkage association within families presenting with extreme disease phenotypes to identify Mendelian diseases. Linkage mapping analysis involving large multi-incident PD families, followed by positional cloning, has also extensively been used to establish the current, known PD genes including α -synuclein (*SNCA*), leucine-rich repeat kinase 2 (*LRRK2*) and vacuolar protein sorting ortholog 35 (*VPS35*), parkin (*PRKN*), PTEN induced putative kinase 1 (*PINK1*) and protein deglycase (*DJ-1*) (**Table 2.1**) (Bonifati, 2014; Klein and Westenberger., 2012). Glucocerebrosidase (*GBA*) has been implicated as a genetic susceptibility factor for PD development (Bandres-Ciga *et al.*, 2020). Genome-wide association studies (GWAS) consist of thousands of markers or single nucleotide polymorphisms (SNPs) scanning across multiple genomes of individuals (with and without the disease) to identify common genetic variation or susceptibility loci linked to a disease. This has led to the adoption of the common-disease-common-variant hypothesis, which has been responsible for the discovery of many PD-susceptibility loci (Hemminki *et al.*, 2008; Nalls *et al.*, 2019; Tam *et al.*, 2019). GWAS has also been able to detect PD-linked common variability in putative candidate gene loci supporting the notion that sporadic and familial PD have a genetic link (Bandres-Ciga *et al.*, 2020). However, GWAS studies face difficulty in terms of statistical power and require a large cohort of cases and controls, thus, having limited potential for the detection of rare variants or those with small effects (Bonifati, 2014; Bayrak and Itan, 2020). The use of these studies has been mostly unsuccessful as they have only been able to explain a small percentage of PD aetiology (Kalinderi *et al.*, 2016). Thus, the introduction of NGS has largely increased the potential for novel gene discovery.

Table 2.1: Genes found to be associated with PD and non-PD Parkinsonism

Locus	Gene Symbol	Chromosome	Method of Discovery	Inheritance	Onset
<i>PARK1</i>	<i>SNCA</i>	4q21-q23	Linkage analysis	AD	EO
<i>PARK2</i>	<i>PRKN</i>	6q25.2-q27	Linkage analysis	AR	EO
<i>PARK3</i>	?	2p13	Linkage analysis	AD	LO
<i>PARK4</i>	<i>SNCA</i>	4q21-q23	Linkage analysis	AD	EO
<i>PARK5**</i>	<i>UCHL1</i>	4p13	Candidate gene approach	AD	LO
<i>PARK6</i>	<i>PINK1</i>	1p35-p36	Linkage analysis	AR	EO
<i>PARK7</i>	<i>DJ-1</i>	1p36	Linkage analysis	AR	EO
<i>PARK8</i>	<i>LRRK2</i>	12p11-q13	Linkage analysis	AD	LO
<i>PARK9*</i>	<i>ATP13A2</i>	1p36	Linkage analysis	AR	EO
<i>PARK10</i>	?	1p32	Linkage analysis	Risk Factor	-
<i>PARK11</i>	<i>GIGYF2</i>	2q37.1	Linkage analysis	AD	EO
<i>PARK12</i>	?	Xq21-q22	Linkage analysis	Risk Factor	-

<i>PARK13</i>	<i>HTRA2</i>	2p12	Candidate gene approach	AD	-
<i>PARK14*</i>	<i>PLA2G6</i>	22q13.1	Linkage analysis	AR	EO
<i>PARK15*</i>	<i>FBXO7</i>	22q12.3	Linkage analysis	AR	EO
<i>PARK16</i>	?	1q32	GWAS	Risk Factor	-
<i>PARK17</i>	<i>VPS35</i>	16q12	WES	AD	LO
<i>PARK18</i>	<i>EIF4G1</i>	3q27.1	Linkage analysis	AD	LO
<i>PARK19</i>	<i>DNAJC6</i>	1p31.3	WES and homozygosity mapping	AR	EO
<i>PARK20</i>	<i>SYNJI</i>	21q22.11	WES	AR	EO
<i>PARK21</i>	<i>DNAJC13</i>	3q22.1	WES	AD	LO
<i>PARK22</i>	<i>CHCHD2</i>	7p11.2	WES	AD	LO/EO
<i>PARK23</i>	<i>VPS13C</i>	15q22.2	WES	AR	EO
<i>PARK24</i>	<i>PSAP</i>	10	Candidate gene approach/WES followed by GWAS	AD	

* Depicts genes responsible for atypical PD or other parkinsonisms.

** No longer considered a significant susceptibility factor for PD.

AR - autosomal recessive; AD - autosomal dominant; EO - early onset; LO - late onset; WES - whole exome sequencing; GWAS - genome-wide association studies.

2.4.3 Strategies for the discovery of novel PD genes and susceptibility factors

It has also been hypothesised that the vast ‘missing heritability’ in complex disorders such as PD, may be attributed to larger penetrant effects of less common variants i.e., the rare-variant-common-disease hypothesis (Gasser *et al.*, 2015; El-Fishawy, 2013; Germer *et al.*, 2019). As described earlier, analysis of NGS data can allow for the identification of rare, highly penetrant variants in multi-incident family pedigrees. The potential of finding a genetic variant with a substantial disease-causing effect is more likely within a PD-affected family than in a sporadic PD individual, due to the added benefit of observing familial co-segregation of the variant of interest. Notably, NGS analysis in PD research has increased exponentially since 2001, probably mainly as a result of the decreasing cost of high-throughput sequencing (**Figure 2.8**). To date, NGS-aided analysis has uncovered several novel genes implicated in PD including AD-inherited genes (*VPS35*, *CHCHD2*, *DNAJC13*) and AR-inherited genes (*DNAJ6*, *VPS13C*, *SYNJI*) (Shulskaya *et al.*, 2018). *VPS35* otherwise referred to as *PARK 17*, is firmly associated with classical PD. However, *DNAJC6* (*PARK 19*), *DNAJC13* (*PARK 21*), *SYNJI* (*PARK 20*), *VPS13C* (*PARK 23*), and *CHCHD2* (*PARK 22*) are also considered pathogenic and viewed as rare genetic contributors to PD (**Table 2.1**) (Olgati *et al.*, 2016; Puschmann, 2017; Schormair *et al.*, 2018; Correia Guedes *et al.*, 2020; Day and Mullin., 2021). The candidate gene approach can then be utilised thereafter to find novel mutations in these putative genes through targeted sequencing or mutational screening.

As a backdrop to the present study, we recently published a Perspective article highlighting all of the studies that used WES and subsequent data analysis to determine novel causes of PD (**Appendix D**). In the article, we highlight the similarities and differences in WES approaches taken in 17 studies that investigated PD-families (of various ancestries) for novel PD genes or susceptibility factors. We

also speculate on the strengths of these approaches (in terms of sequencing and data analysis) and comment on the relevance of our findings for future studies.

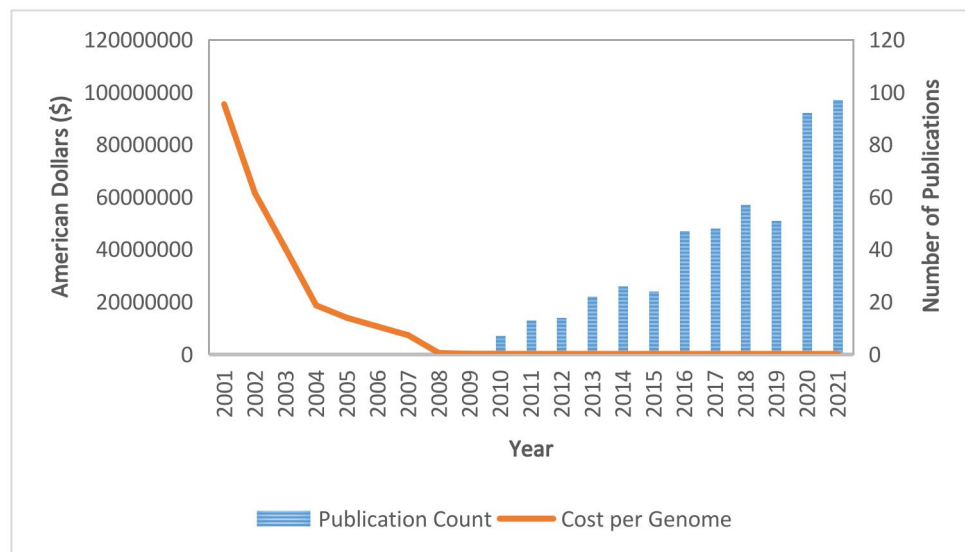


Figure 2.8: The number of publications related to NGS and PD (indexed on PubMed) correlated to the cost of a single sequenced genome (2001-2021)

The PUBMED string search was performed on the 13th of September 2022 and was adapted from Su *et al.*, 2011, and consisted of ('next-generation sequencing' OR 'next-generation sequencing' OR 'next-generation DNA sequencing' OR 'next-generation DNA sequencing' AND 'Parkinson's disease' OR 'PD' OR 'Parkinson's') AND (2000[Publication Date]:2021[Publication Date]). The genome cost data (2001 to 2021) was accessed from NHGRI with permission.

2.5 The state of PD research in sub-Saharan Africa (SSA)

There is currently a major bias in global statistics regarding PD genetics. This is because the majority of PD genomic studies have been conducted on individuals of European and Asian ancestry (Abbas *et al.*, 2017). Estimates of the prevalence of PD in sub-Saharan Africa (SSA) vary widely among studies and range from ~10 - 235 cases/100000 in urban regions (Lekoubou *et al.*, 2014). The low prevalence may be attributed to low life expectancy, the relatively young population, debilitating socio-economic factors, cultural taboos, and the lack of neurological specialists (Dotchin and Walker, 2012).

Critically limiting factors in the diagnosis and subsequent recruitment of PD individuals (of African ancestry specifically) are due to the lack of movement disorder specialists in countries with already limited healthcare and infrastructure and, the beliefs surrounding PD due to the lack of knowledge about the disease. A study analysing the prevalence of movement disorder specialists in Africa indicated a total of 10 in South Africa and 30 in Nigeria, which are concerningly low numbers for the SSA countries with the most research output regarding PD genetics (Hamid *et al.*, 2021). A

cross-sectional survey analysing awareness of PD among South African Xhosa individuals (25 individuals with PD, 98 control individuals and 31 traditional healers) resulted in only 18% of them being able to recognise the disease and almost a third believed the disease was caused by witchcraft and that the affected individual should be removed from the community, indicating there is a significant lack of knowledge regarding the disease among black South Africans (Mokaya *et al.*, 2017). This highlights the limited number of individuals, let alone families, that would be knowledgeable or willing to receive an official PD diagnosis or partake in genetic studies. Furthermore, attempts to bridge the fundamental gaps in African genomics are currently underway. An example is the South African Human Genome Project (SAHGP) initiative to develop a local reference genome based on 24 African ancestry individuals (<https://sahgp.sanbi.ac.za/>). Another initiative is the H3Africa Consortium which aims to develop a pan-African bioinformatics network (H3ABionet) and infrastructure to enhance African genomics research on the continent (Mulder *et al.*, 2017).

Most PD studies in SSA (including Nigeria, Ghana, Zambia, Tanzania and South Africa) to date have focused on targeted genetic screening to determine the frequency of known PD genes in PD individuals. Almost all the available studies have incorporated the use of either multiplex-ligation probe assay (MLPA) or high-resolution melt (HRM) analyses to screen PD individuals for the presence of variants or copy number variations in *LRRK2*, *PRKN*, *PINK1*, *SNCA*, *UCHLI*, *ATP13A2*, *DJ-1*, or *GBA* (Rizig *et al.*, 2021; Okubadejo *et al.*, 2018; Milanowski *et al.*, 2021; Keyser *et al.*, 2010; van der Merwe *et al.*, 2016). However, these studies have reported only a few variants implicated in PD. Another recent study testing South African and Nigerian PD-affected individuals for the presence of common disease-associated variants (using a targeted NGS gene panel) discovered that none of the individuals harboured these common mutations indicating that there may be a host of undiscovered genetic factors influencing the onset of PD in African populations (Oluwole *et al.*, 2020).

2.5.1 Strategies for novel PD gene discovery in SA

Genetic research on complex or rare disease in individuals of African ancestry has now become increasingly relevant, due to their vast genetic diversity as compared to that of Asian or European populations (Sirugo *et al.*, 2019). The use of NGS for the discovery of novel PD genes in families from SSA has been limited to two studies. One study published in 2018 described the WES analysis of an African family affected with juvenile-onset parkinsonism and intellectual disability, resulting in the discovery of a novel homozygous frameshift deletion present in *PTRHDI*, a gene that had previously been implicated in two Iranian families presenting with a similar phenotype (Kuipers *et al.*, 2018). More recently, WES was used for the analysis of an Afrikaner family (a founder population of Dutch, French and German ancestry that are unique to SA) affected with PD, where a novel variant p.G849D in neurexin 2 (*NRXN2*) was prioritised as a candidate disease gene (Sebate *et al.*, 2021).

Our research group has recruited a total of 687 unrelated PD-affected individuals, with only 14.8% of recruits having African ancestry (n = 102). Of these 102 PD recruits, only 6 of those individuals (6.5%) presented with a positive family history (Jansen van Rensburg *et al.*, 2021). This highlights the limitations of the investigation of PD causation in multiplex African families. As important as the recruitment of these ancestry-specific PD families are, it is also important to develop a reproducible analytic workflow (that caters to the genomic diversity in SA PD-affected individuals) to optimise the odds of discovering a disease-associated variant in an unknown gene or to uncover a novel variant in an established PD gene. If this approach is successful, our research could contribute to the understanding of the complex genetic aetiology of PD which could eventually lead to the basis of novel or stratified/personalised modalities of treatment for under-researched ethnic groups afflicted with PD.

Significance of study

This study aims to develop effective bioinformatic methods of NGS analysis for PD-affected individuals in South Africa, thereby allowing analysis to be robust and reproducible. This study also aims to determine a novel genetic factor that may be underlying the cause of PD in a family with South African Xhosa ancestry. Results from this study could provide insight into novel mechanisms that instigate PD onset and progression in individuals of under-researched ethnic populations. Further functional study into novel candidate genes could lead to a formative basis for newer or targeted therapeutic modalities by understanding and manipulating the mutational effects on biological targets associated with the disease. Ultimately, it is a goal that the use of NGS could lead to precision or stratified medicine in complex disease where the treatment and prevention of a particular disease are optimised by considering individual variability at the genetic level. Our study is the first known example of WES analysis in a PD-affected family of Xhosa-African ancestry.

3

CHAPTER 3

Whole exome sequencing analysis of a South African Xhosa family affected with Parkinson's disease

Abstract

Introduction: Parkinson's disease (PD) is a neurodegenerative disorder with complex genetic aetiology. The limited number of mutation screening studies on PD in sub-Saharan African (SSA) populations have not typically identified known genetic causes of the disease. Whole exome sequencing (WES) approaches have previously been successfully utilised to find novel pathogenic mutations or genes in PD families exhibiting Mendelian inheritance patterns. This study aims to identify novel PD susceptibility or pathogenic variants and/or genes through WES and bioinformatic analysis in a Xhosa family (ZA 15) affected with familial PD.

Methods and Results: Initially, WES was performed on two PD-affected siblings and two unaffected siblings from family ZA 15, on the HiSeq 4000 at the Mayo Clinic Core Facility, USA. WES data was analysed using BWA-MEM (GRCh38/hg38 reference alignment), GATk HaplotypeCaller (variant calling) and Ensembl-VEP (variant annotation). Variant call files (VCFs) were scanned for variants in both known (n=21) and putative (n=101) PD genes to eliminate known genetic causes. The VCFs were filtered to include heterozygous, exonic/splice site, non-synonymous variants with a Phred score > 30, present in population databases with a minor allele frequency (MAF) < 0.01 and a CADD > 20. A total of 68 variants were identified, shared between the affected individuals only. These candidate genes were then subjected to gene and protein expression analyses to determine neuro-specific tissue and pathway expression. A total of 24 variants were prioritised and underwent Sanger sequencing to confirm co-segregation within the family. Thereafter, the variants were screened through several private (not publicly available) population cohorts to determine MAFs, resulting in the exclusion of variants with a MAF > 0.01. Following the control population screening, three variants of interest were prioritised, namely, *AHNAK2* p.D1540H, *MANF* p.A13V and *ZDHHC11* p.R276P. These remaining variants were then subjected to Sanger sequencing in 100 South African Xhosa controls, however, none of the variants were found to be present. Lastly, the three remaining variants were re-evaluated based on gene and protein expression data to determine possible correlations to PD pathobiology. A single variant (p.A13V in the mesencephalic astrocyte-derived neurotrophic factor (*MANF*) gene) was prioritised for further study.

Conclusion: Identifying novel PD genes in under-represented population groups may improve clinical diagnoses and treatment options by providing insight into unknown PD molecular mechanisms,

detecting rare PD biomarkers and determining novel drug targets. WES analysis of ZA 15 yielded 3 novel variants that were found to be pathogenic across > 5 *in silico* pathogenicity prediction tools and had a MAF < 0.01 in the available population databases/private PD/non-PD population cohorts, indicating the rarity and potential pathogenicity of these variants. Subsequently, MANF, a protein which exerts a protective effect on dopaminergic neurons in the *substantia nigra pars compacta* (SNpc) (the main neuronal region implicated in PD), was selected as the top candidate. This gene or variant has not been implicated as a genetic cause of PD before and thus, these findings illustrate the usefulness of under-represented populations for providing potentially new insights into disease pathobiology. However, we cannot state with certainty that any one of these variants may have caused the disease in this family, and thus, further *in silico* and ‘wet-lab’ functional analysis of the variants and their impact on protein function, is necessary. Overall, this study illustrates the importance of incorporating understudied populations for novel gene discovery in disease genomics.

Keywords: African Ancestry; Bioinformatic Pipelines; Familial PD; Novel Genes; Parkinson’s Disease; WES



3.1 Introduction

Next-generation sequencing (NGS) approaches have enabled the rapid expansion of genomic-based research due to their ability to perform high-throughput sequencing in a cost-effective and time-saving manner. NGS technology is typically utilised in a clinical setting for diagnostic evaluations of genetic disease, and more recently, in a research setting to determine novel genetic causation of rare or complex diseases, such as PD. When considering NGS for the study of genetic disorders, WES presents as the most suitable choice as most pathogenic variants (80–85%) found to date, are in exonic regions of the genome (Ku *et al.*, 2016). NGS approaches such as whole genome sequencing (WGS) or WES produce a large array of genomic data requiring a robust bioinformatic pipeline (typically comprised of open-source software) that is constructed according to best-practice guidelines to elucidate the most promising candidate variants. However, it is important to employ an analytic workflow that is most relevant and appropriate to the disease of interest, the mode of inheritance and the ancestry group, to optimise the prioritisation of variants.

PD is a neurodegenerative motor disease displaying a diverse, erratic genetic aetiology with very few causal genes having been identified, thus far. To date, several novel gene discovery WES studies in familial PD cases have been published, however, these studies have largely been limited to individuals of European and Asian descent (Bentley *et al.*, 2020). Genetic studies in South Africa on the known PD genes have not yet identified a genetic cause in the majority of PD-affected individuals. These findings hint at the possibility that there are as-yet-undiscovered PD genes. WES approaches provide a practical method to find novel pathogenic mutations in these PD-affected individuals. As PD-affected individuals of African ancestry are considerably understudied, it is imperative that the analysis and filtering of the genomic data, to prioritise variants, be tailored to factor in diverse ancestries. This study aims to evaluate the WES data from a South African Xhosa multi-incident PD family to elucidate novel candidate disease genes/variants using a tailored bioinformatic approach.

3.2 Methods and materials

The methodological approach chosen for the present study was determined after an exhaustive search for all published studies that used WES to identify a novel cause of disease in a PD-affected family (**Figure 3.1**). The workflows and bioinformatic tools for each of the reviewed studies were compiled and compared in a Perspective article which was published (Pillay *et al.*, 2021; **Appendix D**) and this formed the basis for the design of the present study.

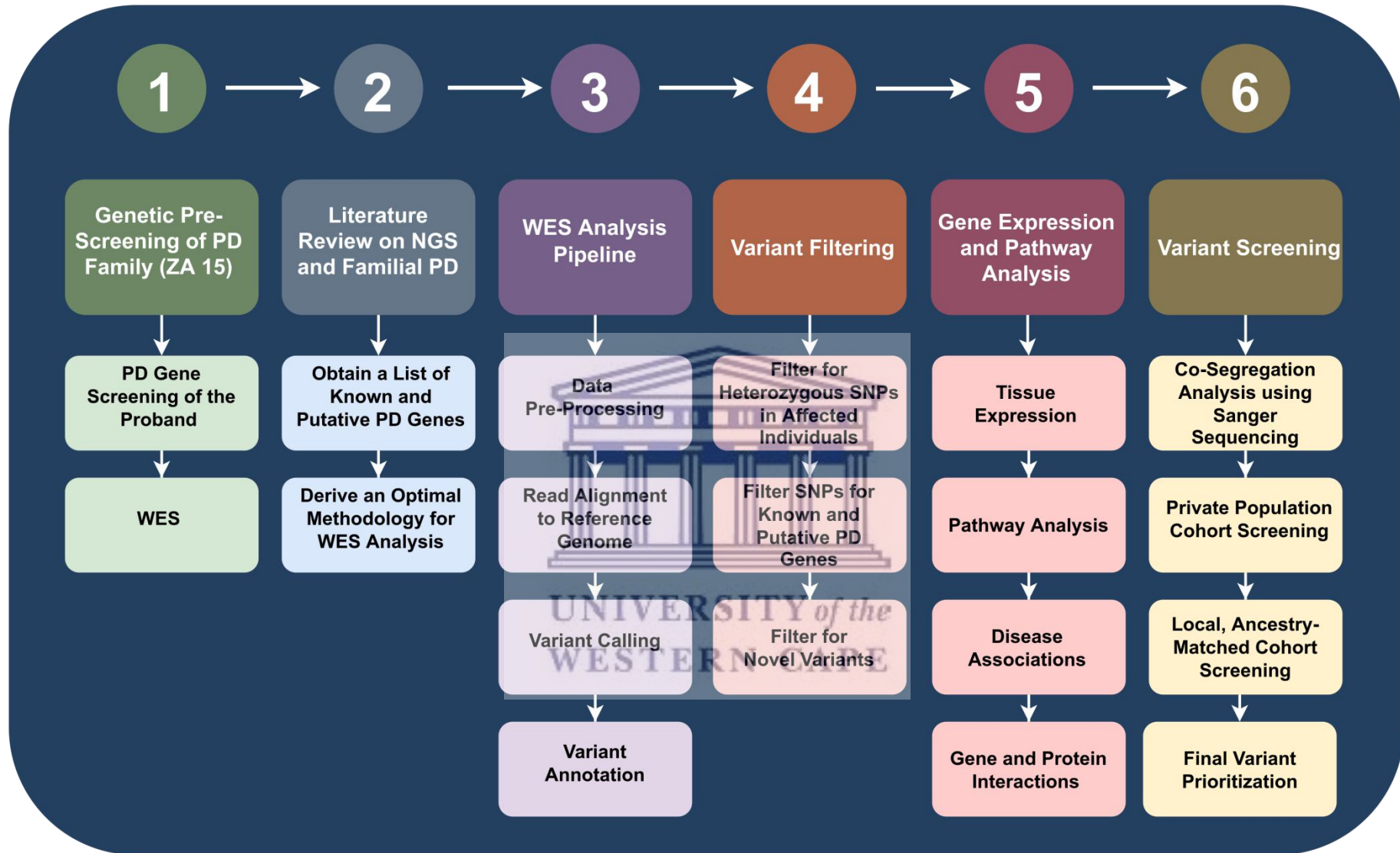


Figure 3.1: Outline of the methodological approach for Aim 1 of the PhD project

3.2.1 Study participants (South African Xhosa family - ZA 15)

3.2.1.1 Ethical considerations for study

Ethics approval for the parent study has been obtained from the Health Research Ethics Committee at Stellenbosch University, South Africa (Reference Number: 2002/C059; **Appendix A**). Ethics approval for this PhD project has also been obtained from the Biomedical Science Research Ethics Committee of the University of the Western Cape, South Africa (Reference Number: BM21/4/13; **Appendix B**). Furthermore, since this project involves a collaboration between two research institutions, a Data Transfer Agreement (DTA) between Stellenbosch University (“Provider”) and the University of the Western Cape (“Recipient”) acknowledging the transfer of NGS data generated from human DNA samples, has been drafted and signed (**Appendix C**). All study participants provided written, informed consent to take part in the study and provide peripheral blood samples for genetic analysis.

3.2.1.2 Selection criteria for study participants

The family selected for WES (designated as ZA 15 since they were the fifteenth family to be recruited) are of South African Xhosa ancestry and consisted of two PD-affected individuals (siblings), their two unaffected siblings and two other unaffected individuals (**Figure 3.2**). They were selected for this study since they are one of the few families of African ancestry for which there is DNA available of >1 affected individual and because these individuals were diagnosed by a movement disorder specialist. The PD family was recruited from the Movement Disorders Clinic at Tygerberg Hospital (Cape Town, Western Cape, South Africa). The affected individuals underwent a standardised neurological examination by a movement disorder specialist - Prof. Jonathan Carr - and were diagnosed according to the UK Parkinson’s Disease Society Brain Bank Diagnostic Criteria (Gibb *et al.*, 1988). The age at onset (AAO) of PD and details regarding family history and lifestyle were also obtained from the patients using a questionnaire by a trained research nurse.

3.2.1.3 Pedigree of South African Xhosa family (ZA 15) affected with Parkinson's disease

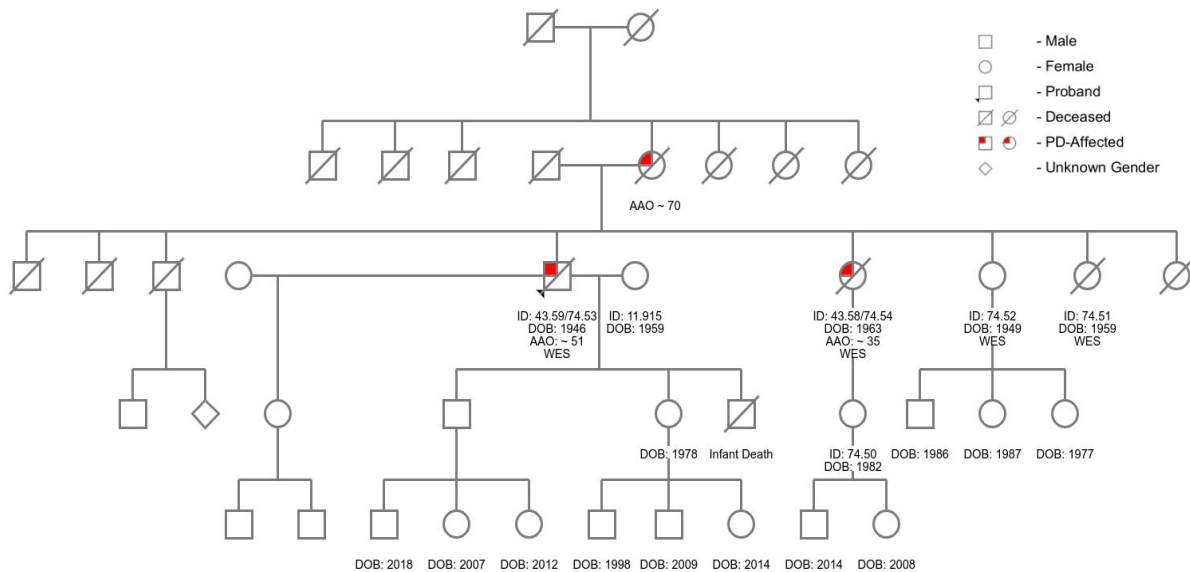


Figure 3.2: Pedigree of Xhosa family ZA 15

Affected family members (PD-affected individuals) are indicated in red. Squares represent males, while circles represent females. Dashed lines through the squares and circles indicate deceased individuals. ID = Lab identification number of individuals that took part in the study, DOB = Date of birth and AAO = Age at onset of PD. (All personal information was removed to allow anonymity of individuals).

3.2.2 WES and data analysis pipeline

3.2.2.1 WES

Four individuals from the PD family ZA 15 were selected for sequencing following the genetic pre-screening of the proband. The proband (74.53), his affected sister (74.54) and their two unaffected siblings (74.52 and 74.51) underwent short-read WES. Furthermore, DNA samples were also collected from the spouse of the proband (ID 11.915) and the daughter of one of the affected siblings (ID 74.50). WES was done by our collaborator, Prof. Owen Ross at the Mayo Core Facility, Mayo Clinic, Florida, USA. Library preparations, using 50ng of sample DNA, were made following the manufacturer's guidelines. Library concentrations were subsequently assessed and enriched using Agilent's SureSelect XT Target Enrichment System V5+UTR (Agilent, Santa Clara, CA, USA). The libraries were then quantified and the enriched exonic regions were sequenced on an Illumina HiSeq 4000 (Illumina, San Diego, CA, USA).

3.2.2.2 WES data analysis workflow

A WES data analysis workflow was constructed for the present study based on best-practice guidelines and tools observed in previously published WES analyses of familial PD cases with an unknown cause of disease (Funayama *et al.*, 2015 Sudhaman *et al.*, 2016, Straniero *et al.*, 2017). A brief overview of all the WES analysis steps can be seen in **Figure 3.3**. The bioinformatic analyses were conducted on the SANBI high-performance computing cluster (South African National Bioinformatics Institute, University of the Western Cape, South Africa).

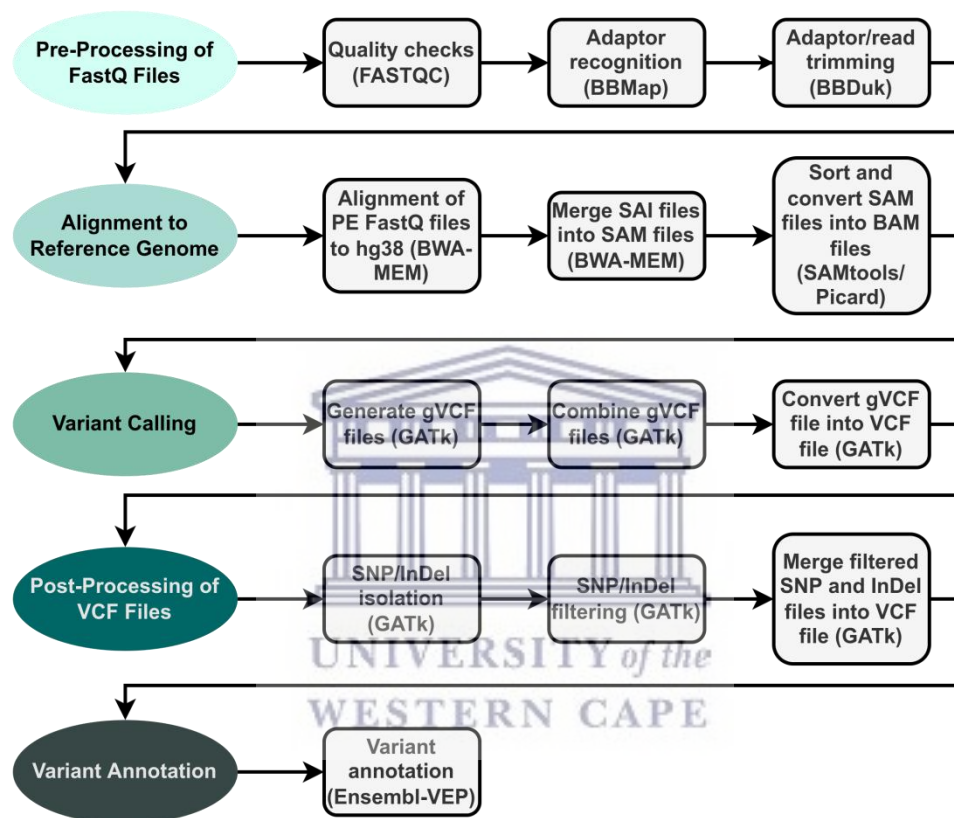


Figure 3.3: WES data analysis workflow used for the analysis of family ZA 15

3.2.2.2.1 Pre-processing of FASTQ files

Raw WES data, in the form of paired-end (PE) FASTQ files, were initially subjected to quality checks using the FASTQC tool, version 0.11.9, (<https://www.bioinformatics.babraham.ac.uk/projects/FASTQC/>). FASTQ files were assessed for per base sequence quality, per read sequence quality and for the presence of remnant adaptor sequences (Leggett *et al.*, 2013). Thereafter, the BMap tool suite, version 38.86 (<https://jgi.doe.gov/data-and-tools/software-tools/bbtools>), was employed to identify the exact adaptor

sequence flanking the reads and to subsequently trim off these identified sequences, as well as, extraneous low-quality reads (Phred < 20) that may affect read alignment (Liao *et al.*, 2017). FASTQC was also used to assess the files during post-processing to ensure the adapters were removed and the reads were trimmed accordingly.

3.2.2.2.2 Read alignment

Burrows-Wheeler Aligner (BWA) (version 0.7.17) was used to index the human reference genome (hg38/GRCh38) assembly and the BWA-MEM algorithm was employed to align the PE FASTQ files to the reference genome, producing a single SAM file per individual. Samtools (version 1.11) was used to convert the SAM file into a workable BAM file. Furthermore, Picard (version 2.20.1) was used to sort the BAM files before variant calling. BAM file summary statistics concerning read alignment rate as well as read length and quality were generated using Samtools (version 1.11).

3.2.2.2.3 Variant calling

The Genome Analysis Toolkit (GATk) HaplotypeCaller (version 4.1.2.0) was utilised for variant calling to create genomic VCF files (gVCFs) for each individual. The gVCFs of all four individuals were then combined into one gVCF also using GATk (version 4.1.2.0), allowing the visualisation of all four genotypes across all genomic sites (Disratthakit *et al.*, 2022). Finally, a VCF file was generated for post-processing using GATk (version 4.1.2.0).

3.2.2.2.4 Post-processing of VCF files

The GATk suite (version 4.1.2.0) was further used to separately isolate SNPs and InDels from the VCF file for further quality control through filtering. The SNPs were filtered to ensure each SNP possessed a quality Phred score > 30.0, a QualByDepth (QD) > 2.0 (used to normalise quality scores preventing raised scores caused by regions of deeper coverage), FisherStrand (FS) < 60.0 (a probabilistic strand bias score determining whether an alternate allele was favoured on either strand), StrandOddsRatio (SOR) < 3.0 (another strand bias estimate calculated using a symmetric odds ratio test), RMSMappingQuality (MQ) > 40.0 (the root mean square mapping quality over all the reads at a particular site), MappingQualityRankSumTest (MQRankSum) > -12.5 (compares the mapping qualities of the reads supporting the reference vs. alternate allele) and a ReadPosRankSum > -8.0 (compares the positions of the reference and alternate alleles on reads). InDels were also filtered to include those with a quality Phred score > 30.0, QD > 2.0, FS < 200.0 and a ReadPosRankSum > -20.0. The filtered SNPs and Indels that contained a PASS flag following filtering were merged into a VCF for annotation.

3.2.2.2.5 Variant annotation

Ensembl's Variant Effect Predictor (VEP), version 104.0, (<https://www.ensembl.org/info/docs/tools/vep/index.html>) was used to perform the variant annotation using the command line interface and cache repository. Typically, variant annotation allows for the addition of auxiliary functional information retrieved from curated databases that allow for individual variant interpretation (Yang *et al.*, 2016). VEP also annotated the VCF against the largest population frequency databases including the 1000 Genomes Project (<https://www.internationalgenome.org/data>), gnomAD (exome and genome) (<https://gnomad.broadinstitute.org/>), NCBI dbSNP, build 155 (<https://www.ncbi.nlm.nih.gov/snp/>) and ExAC (available through <https://gnomad.broadinstitute.org/>). Population-specific minor allele frequencies (MAFs) calculated for each variant in multiple ethnic groups (African/African-American (AFR), Admixed-American/Latino (AMR), East Asian (EAS), Non-Finnish European (EUR) and South Asian (SAS)), were provided, to determine the rarity of the variants in contrast to common polymorphisms. Typically, a variant presenting with a MAF of < 0.01 (1 %) is considered rare in a population group (Bomba *et al.*, 2017).

Furthermore, the variants were also annotated against several *in silico* functional prediction tools to determine pathogenicity scores as well as evolutionary conservation. The Combined Annotation Dependent Depletion (CADD) (<https://cadd.gs.washington.edu/>) pathogenicity predictor scores the deleteriousness of single nucleotide variants by incorporating multiple annotations, as opposed to just sequence homology or conservation used by most pathogenicity prediction scorers (Kircher *et al.*, 2014). SIFT (<https://sift.bii.a-star.edu.sg/>), Poly-Phen2 (<http://genetics.bwh.harvard.edu/pph2/>) and PROVEAN (<http://provean.jcvi.org/index.php>) similarly predict the effect of an amino acid substitution on protein function using a sequence homology approach and simultaneously incorporating the physical properties of the amino acids using probabilistic classifiers (Ng & Henikoff, 2001, Adzhubei *et al.*, 2010 and Mahmood *et al.*, 2017). Mutation Taster (<https://www.genecascade.org/MutationTaster2021>) makes use of Random Forest models incorporating sequence homology for deleterious predictions (Steinhaus *et al.*, 2021) while Mutation Assessor (<http://mutationassessor.org/r3/>) predicts functional impact based on evolutionary conservation of the affected amino acid (AA) in protein homologs (Reva *et al.*, 2011). fathmm (<http://fathmm.biocompute.org.uk/>) predicts functional impact incorporating both sequence conservation with homologous sequences and conserved protein domains with Hidden Markov Models (Rogers *et al.*, 2018). M-CAP (<http://bejerano.stanford.edu/mcap/>) combines a multitude of pathogenicity prediction scores including SIFT, Polyphen-2 and CADD scores using a supervised learning classifier (Jagadeesh *et al.*, 2016). GERP++ (<http://mendel.stanford.edu/sidowlab/downloads/gerp/index.html>) was also incorporated to score the

variants using a novel maximum likelihood rate estimation for determining selectively constrained sites (Davydov *et al.*, 2010).

3.2.3 Variant filtering

Once the VCF file was fully annotated, a stringent variant filtering approach to prioritise a limited number of potentially disease-causing variants was implemented. A brief overview of the workflow for the variant filtering steps is outlined in **Figure 3.4**.

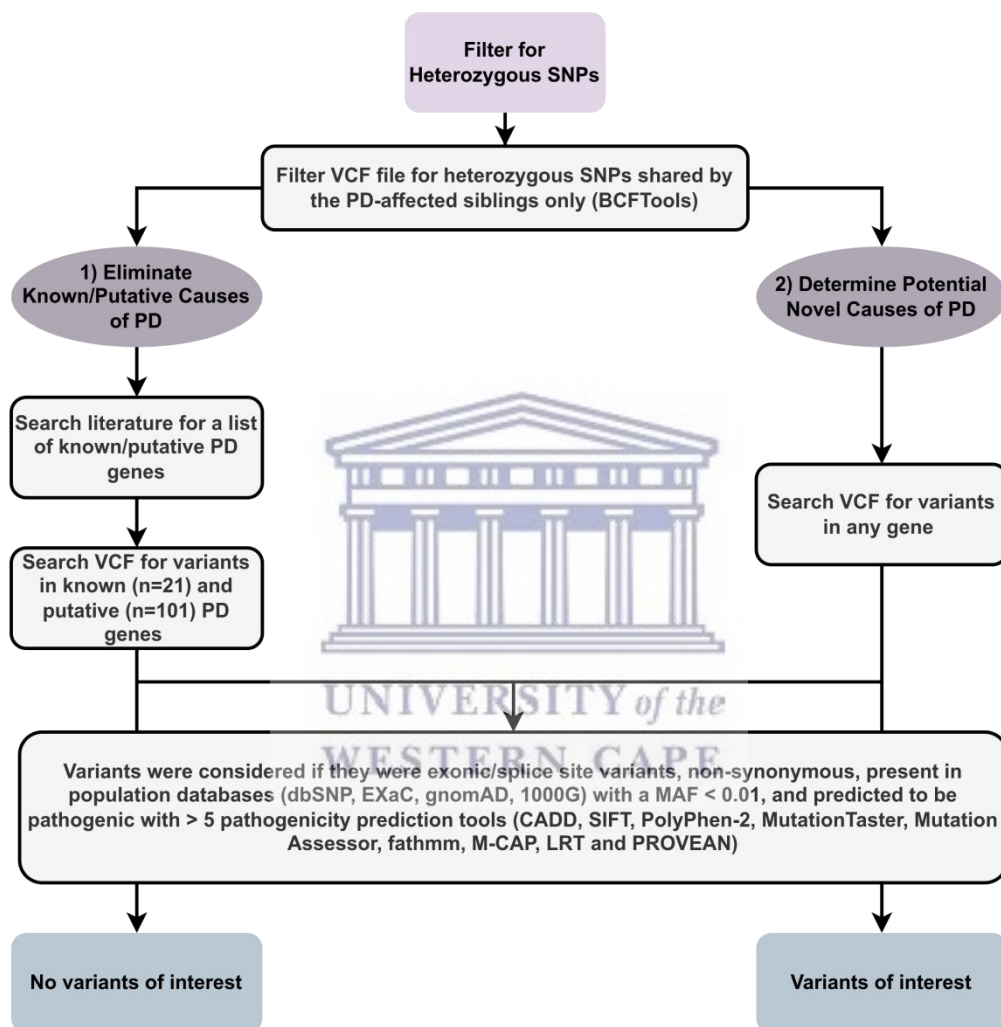


Figure 3.4: Workflow for variant filtering to eliminate known causes of PD and identify novel candidate variants

3.2.3.1 Filter VCF for heterozygous SNPs that are shared between the affected siblings

As a dominant model of inheritance for PD was assumed for ZA 15 (**Figure 3.2**), only heterozygous variants were selected for further analysis. BCFTools (version 1.11) was used to intersect the VCF for

heterozygous, exonic variants that were shared by only the affected individuals and not the unaffected individuals, while all2VCF (version 0.3.1) combined the non-standard genotyping output into a VCF.

3.2.3.2 Filter VCF for known and putative PD genes

A list of both known (n = 21) and putative PD genes (n = 101) (**Appendix E**) was assembled for downstream filtering of the VCF that contained only the shared heterozygous and exonic/splice site region SNPs. The VCF file was scanned for the presence of any SNPs belonging to the known/putative PD genes. This analysis was used to rule out potential known causes of disease or to determine if a novel variant in a known/putative gene was to be investigated further.

3.2.3.3 Filter VCF for novel potential disease variants

Thereafter, the VCF was filtered according to a list of stringent criteria to narrow down the list of novel variants that could be pursued for further investigation. VCF files were then filtered to include only heterozygous, exonic, non-synonymous variants with a Phred QS > 30, present in all population databases with a MAF < 0.01 and a CADD score > 20.

3.2.4 Gene expression and pathway analysis

3.2.4.1 Tissue expression

Variants were analysed against online gene expression databases to determine if any of the variants were expressed in neuro-associated tissue. The Genotype-Tissue Expression (GTEx) Portal (<https://www.gtexportal.org/home/gene>) and the Human Protein Atlas (HPA) (<https://www.proteinatlas.org/>) (Lonsdale *et al.*, 2013, Pontén *et al.*, 2008) were utilised for annotation of the variants. Furthermore, Mouse Genome Informatics (<http://www.informatics.jax.org/>) (Begley *et al.*, 2022) was also utilised to corroborate gene expression profiles of the variants and to determine the phenotype of gene knockout in mouse models (**Appendix G**).

3.2.4.2 Pathway analysis

Kyoto Encyclopaedia of Genes and Genomes (KEGG) Pathways Analyser (<http://www.genome.jp/kegg/pathway.html>) and PANTHER Pathway Analyser (<http://www.pantherdb.org/pathway/>) databases were searched to identify any association between the variants and biological pathways.

3.2.4.3 Gene-disease association

All variants were annotated to determine existing ‘gene-disease’ associations. ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>) provides information on the correlation between human variation and disease phenotypes, with supporting literature (Landrum *et al.*, 2020). Online Mendelian Inheritance in Man (OMIM) (<https://www.omim.org/>) (McKusick *et al.*, 1998) is a curated database that also catalogues human phenotypes with genetic variation.

3.2.4.4 Gene/protein interactions

STRING (<https://string-db.org/cgi/>) produces a network of both known and predicted protein-protein interactions using machine learning (ML) algorithms and text mining (Szklarczyk, 2021). Each of the variant genes was enriched against the list of known/putative PD genes (**Appendix E**) to determine if they are co-expressed or related to known PD genes. All variants were subject to Gene Ontology (GO) (<http://geneontology.org/>) enrichment which allows for the gene annotation of biological processes that were found to be implicated when the gene is expressed in humans.

3.2.5 Variant screening using wet-laboratory techniques

3.2.5.1 Sanger sequencing

3.2.5.1.1 DNA quantification

DNA samples were quantified using a NanoDrop® 2000 spectrophotometer (Thermo Scientific, MA, U.S.A) and diluted to working concentrations of ~30 ng/ul.

3.2.5.1.2 Polymerase chain reaction primer design

Oligonucleotide primers (forward and reverse) for each of the 24 variants were designed using sequence data obtained from NCBI’s Genome Variation Viewer (<https://www.ncbi.nlm.nih.gov/variation/view>). Sequence data was submitted to the Primer-Basic Local Alignment Search Tool (BLAST) (<https://www.ncbi.nlm.nih.gov/tools/primer-blast/>) to generate a list of optimum primer pairs. Primer3 software (version 4.0.0) (<https://primer3.ut.ee/>) (Untergasser *et al.*, 2012) was subsequently used to identify the best primer pair and confirm the size (bp), GC content (%), melting temperature (T_m) (°C) and self-complementarity (**Appendix H**). The primers were synthesised by Inqaba Biotechnical Industries (Pty) Ltd, Pretoria, SA.

3.2.5.1.3 Polymerase chain reaction

The regions of interest were amplified in a 25µl PCR reaction containing ~30ng (1µl) of template genomic DNA. The PCR master mix was composed of 5µl of 5X Green GoTaq® Reaction Buffer (Promega, Madison, Wisconsin, U.S.A), 2.5µl of MgCl₂ (25mM), 0.75µl dNTPs (25mM) (Promega, Madison, Wisconsin, U.S.A), 0.5µl of both forward and reverse primers (20µM), 0.05µl GoTaq® G2 Flexi DNA Polymerase (Promega, Madison, Wisconsin, U.S.A). All PCR reactions included a negative, non-template control (sterile dH₂O) to identify potential PCR contaminants. Amplification was performed in a SimpliAmp thermal cycler (ThermoFisher Scientific, Massachusetts, USA). PCR conditions were as follows: one denaturation cycle of 95°C for 5 minutes, 35 cycles of denaturation at 95°C for 30 seconds, annealing at the calculated optimum temperature (specific to each primer pair) followed by an extension at 72°C for 30 seconds. The final extension cycle was performed at 72°C for 7 minutes before being cooled to 4°C.

3.2.5.1.4 Agarose gel electrophoresis

Gel electrophoresis was performed on the PCR product to confirm the amplification of the target and ensure the amplified product had no contamination or non-specific amplification. A 1% (w/v) agarose gel was prepared using 1.5g agarose powder (Agarose CSL-AG500, Cleaver Scientific Warwickshire, UK), 150mL 1X sodium borate (SB) buffer and 4µL of Conda Safe Nucleic Acid Staining Solution (Condalab, Madrid, Spain). 5 µl of PCR product were loaded into the gel wells with 2µl of DNA molecular weight markers (100 bp and 1000bp) that were used to determine the size of the PCR amplicons. Gel electrophoresis was subsequently performed at 120V for 45 minutes. Gels were viewed using the BioRad GelDoc Go Imaging System (BioRad, Johannesburg, South Africa).

3.2.5.1.5 Post-polymerase chain reaction cleanup and Sanger sequencing

All samples underwent post-PCR enzymatic cleanup and were Sanger sequenced at the DNA Sequencing Unit, Central Analytical Facilities (CAF) (Stellenbosch University, Stellenbosch, South Africa). Capillary electrophoresis was done on a 3130 x1 Genetic Analyser (Applied Biosystems, CA, U.S.A) using the BigDye Terminator Sequence Ready Reaction kit version 3.1 (Applied Biosystems, CA, U.S.A). Primer pairs used for the previous PCR reactions were diluted to 1.1 pmol/ul for Sanger sequencing. Analysis of Sanger sequencing data was carried out using Cütepeaks (version 0.2.3) (Schutz *et al.*, 2021). Sanger sequences were then analysed to confirm the correct targeting of the region of interest using Ensembl's Blast Like Alignment Tool (BLAT) (<https://www.ensembl.org/Tools/Blast>).

3.2.5.2 Screening of variants in private cohorts

To further reduce the number of candidate variants, the variants were screened through private (not publicly available) population cohorts, by contacting authors of published articles or our collaborators. This was done to determine the MAFs of our variants of interest since they appeared to be rare or non-existent in public databases e.g. gnomAD. The cohorts selected were either ancestry-matched (Xhosa or African), or had neurological disorders, including PD. The first cohort of interest was a large Xhosa-ancestry cohort that consisted of WES data for both cases and controls, with cases including schizophrenia-diagnosed individuals ($n = 909$) and controls that had been shown to not have neurological conditions ($n = 917$) (Gulsuner *et al.*, 2020). The 24 prioritised variants were screened through the cohort and variants exhibiting a $MAF > 0.01$, were excluded. Thereafter, the prioritised variants were subjected to screening in smaller cohorts including a study group on tuberculosis (ResisTB) comprising 161 self-identified Xhosa individuals that had undergone WGS (unpublished data). Another African-ancestry-based control cohort incorporated into our study included the H3Africa Baylor Dataset (excluding participants from Mali and the South African Human Genome Project) consisting of 386 WGS samples (Choudhury *et al.*, 2021).

We also incorporated three PD-specific cohorts to determine if our variants had been found in any other individuals with PD or other movement disorder. The Queensland Parkinson's Project (QPP) consisting of 66 individuals (47 PD cases, 5 'other movement disorder' cases and 14 family controls) who had undergone WES (Bentley *et al.*, 2021), the Mayo Clinic cohort consisting of familial PD and Lewy Body disease brain cases, and the French and Mediterranean Parkinson's Disease Genetic Study group (FMPD cohort) consisting of 1319 PD cases of European and North African ancestry (Fevga *et al.*, 2022).

3.3 Results

3.3.1 Study participants (South African Xhosa family - ZA 15)

3.3.1.1 Descriptive overview of the family

The family ZA 15 has Xhosa-African ancestry; a population group that is descended from an admixture of the Northern-African Bantu and Southern-African San population groups (Newman, 1995). They are the second largest Bantu population group residing in SA, specifically concentrated in the Eastern Cape (the region where the family was recruited).

The mother of the four siblings was stated to have had PD, and thus a dominant model of inheritance was assumed in this family. However, it should be noted that she was deceased at the time of this study and so her affected status could not be confirmed. The two affected and two unaffected siblings were earmarked for WES analysis. The proband had a typical age of onset at 51 years old and presented with typical features of PD including bradykinesia and a marked resting tremor. He also had non-motor features including insomnia and autonomic involvement, and subsequently developed mild psychosis with visual and auditory hallucinations. His affected sister presented with a much earlier age of onset of 35 years, where she showed initial symptoms of tremor and difficulty with walking. However, no abnormalities were detected with a magnetic resonance imaging (MRI) brain scan. Eight years after onset, she exhibited no autonomic complaints with no cognitive disturbances. Both siblings were found to be levodopa responsive.

3.3.1.2 Pre-screening of the proband resulted in no variant/s of significance in known or putative PD genes

Before WES, the proband had been subjected to genetic screening of the common PD genes to determine whether he had a pathogenic variant in one of those genes (**Table 3.1**). A single heterozygous variant was found in *LRRK2* (rs148113070, p.E899D).

Table 3.1: Pre-screening results for known PD genes in the proband (ID 43.59/74.53) using various mutation screening techniques

Parkinson's Disease Genes Screened	Screening Method	Variants/Copy Number Changes/Present in Proband
Parkin	MLPA*	No
751 genes associated with neurological diseases	Ion AmpliSeq™ Neurological Research panel**.	No
Copy number changes in <i>PARK2, DJ-1, SNCA, LRRK2, PINK1, GCH1, UCHL1, ATP13A2, LPA, TNFRSF9, CAV2 & CAV1</i>	MLPA	No
<i>LRRK2</i> (G2019S)	MLPA	No
<i>SNCA</i> (A30P)	MLPA	No
<i>DJ-1</i>	Sanger Sequencing	No
<i>LRRK2</i>	Sanger Sequencing	Heterozygous p.E899D
<i>SNCA</i>	MLPA	No
<i>PINK1</i>	Sanger Sequencing	No
<i>EIF4G1</i>	Sanger Sequencing	No
<i>VPS35</i>	Sanger Sequencing	No
<i>SCA</i>	Sanger Sequencing	No
<i>JPH3</i>	Sanger Sequencing	No
<i>GBA</i>	Sanger Sequencing	No

*MLPA - multiplex ligation-dependent probe amplification.

**Ion AmpliSeq™ Neurological Research panel - a commercially available panel containing 751 genes affecting brain and nervous system function.

MAFs for this variant, in all the population databases considered, were less than 0.01 indicating the rarity of the alternate allele (orange cells, **Table 3.2**) (Bomba *et al.*, 2017). All pathogenicity prediction programs determined that the variant had little to no impact on protein function and it was considered benign or tolerated (**Table 3.2**). Furthermore, with analysis of the WES data it was found that the variant is present in both affected siblings, as well as, an unaffected sibling, and thus, the variant was excluded from further analysis. No other likely pathogenic variants were found in the PD genes that had been screened.

Table 3.2: Population database MAFs and pathogenicity prediction scores for *LRRK2* p.E889D

	Population Databases			Pathogenicity Prediction Tools					
	1000G (ALL)	ExAC (ALL)	GnomAD (ALL)	CADD	PP-2	SIFT	MA	fathmm	PROVEAN
<i>LRRK2</i> p.E899D	0.0006 rare	0.00024 rare	0.000385 rare	18.62 likely benign	0.101 benign	0.68 tolerated	0.496 low	-0.63 tolerated	-0.147 neutral

1000G = 1000 Genomes Project; ExAC = Exome Aggregation Consortium; GnomAD = Genome Aggregation Database; CADD = Combined Annotation Dependent Depletion score; PP-2 = Polyphen-2 score; SIFT = Sorting Intolerant From Tolerant score; MA = Mutation Assessor score; fathmm = Functional Analysis through Hidden Markov Models score and PROVEAN = Protein Variation Effect analyser. Orange cells indicate that the variant was considered rare due to the observed MAF.

3.3.2 WES analysis was performed on 4 individuals in ZA 15

Based on these findings, it was decided that an NGS approach would be needed to find a possible novel PD-causing gene in family ZA 15. Four individuals were selected for WES: the proband (74.53), his affected sister (74.54) and their two unaffected siblings (74.52 and 74.51) (**Figure 3.2**). As mentioned previously, after analysing the published literature, the best practice/most appropriate approach to use for novel gene discovery in PD was determined and this informed the subsequent flow of steps used in this study.

3.3.2.1 Summary statistics of BAM and VCF files depicted a high rate of sequence alignment and read quality

Summary statistics obtained for the BAM files indicated a high coverage rate (> 99%) for each individual, with all reads possessing a quality score (QS) > 30. The generated gVCF containing the variants across all four individuals revealed a total of 2,486,448 SNPs and 317,459 InDels (**Table 3.3**). Following annotation, the multi-sample VCF was intersected to isolate heterozygous variants that

were only shared between the affected siblings and not present in either of the unaffected siblings, as a dominant model of PD inheritance was assumed for this family. This resulted in a total of 1,785 heterozygous exonic variants that were unique to the two affected siblings, 780 of which were non-synonymous variants (the remaining variants were synonymous variants, which indicate the SNPs did not result in a change in the protein sequence).

Table 3.3: Summary statistics produced for each sample that was whole exome sequenced (BAM and gVCF files)

	Sample ID			
	74.53	74.54	74.52	74.51
Total number of reads (n)	116,499,548	108,867,292	115,197,926	108,794,328
Total number of mapped and paired reads (n)	116,054,378	108,706,426	114,726,244	108,517,790
Read mapping alignment rate (%)	99.62	99.85	99.59	99.75
Average read length (bp)	151	151	151	151
Average quality score (Phred)	38.1	38.3	37.4	38.3
Total number of SNPs (n)	2,486,448			
Total number of INdels (n)	317459			

The 780 variants were then subjected to further downstream filtering using a list of known and putative PD genes (**Appendix E**) to eliminate potential known causes of PD that had not been screened for, previously. Eight variants were detected in these genes (**Table 3.4**). The MAFs and pathogenicity prediction scores for each of the variants were analysed to determine their validity as being potentially pathogenic, however, none of the variants were found to be rare across all population databases or deleterious, thus they were excluded from further analysis.

Table 3.4: Non-synonymous, exonic variants found in the known/putative PD genes that are shared by the affected individuals (74.53 and 74.54) only

Variants	Population Databases			Pathogenicity Prediction Tools					
	1000G (ALL)	EXAC (ALL)	gnomAD (ALL)	CADD	PP-2	SIFT	MA	fathmm	PROVEAN
<i>GBA</i> p.K13R	0.024 common	0.0071 rare	0.0224 common	0.003 B	0.0 B	0.387 T	0.205 N	3.62 T	-0.06 N
<i>FAM83</i> p.V11L	0.4 common	0.3516 common	0.3635 common	0.356 B	0.0 B	1.0 T	-0.92 N	2.87 T	-0.2 N
<i>ELOA2</i> p.A446T	0.53 common	0.506 common	0.4761 common	19.87 LB	0.999 D	0.096 T	2.095 M	3.16 T	-1.37 N
<i>ELOA2</i> p.C254F	0.53 common	0.5058 common	0.4758 common	0.003 B	0.115 B	0.715 T	0 N	3.31 T	-0.95 N
<i>ELOA2</i> p.R179P	0.56 common	0.5614 common	0.5377 common	0.249 B	0.0 B	0.353 T	0 N	3.24 T	1.37 N
<i>OR8B3</i> p.M114I	0.027 common	0.2767 common	0.3001 common	3.52 B	0.038 B	0.555 T	-1.1 N	7.61 T	-0.05 N

<i>ZNF543</i> p.N489S	0 rare	0.01498 common	0 rare	0.001 B	0.02 B	1.0 T	-1.26 N	2.37 T	-1.32 N
<i>ZNF543</i> p.G559R	0.027 common	0.0141 common	0.0265 common	21.8 LP	0.171 B	0.201 T	2.077 M	3.15 T	-1.66 N

1000G = 1000 Genomes Project; ExAC = Exome Aggregation Consortium; GnomAD = Genome Aggregation Database.

CADD = Combined Annotation Dependent Depletion score; PP-2 = Polyphen-2 score; SIFT = Sorting Intolerant From Tolerant score; MA = Mutation Assessor score; fathmm = Functional Analysis through Hidden Markov Models score and PROVEAN = Protein Variation Effect analyser.

B = benign; D = Deleterious; T = Tolerated; N = Neutral; M = Moderate; LB = Likely benign; LP = Likely pathogenic.

'Common' indicates a MAF > 0.01 and 'rare' indicates a MAF < 0.01 in the population databases.

Orange cells are indicative of results of significance.

3.3.3 Variant filtering and prioritisation yielded 68 variants of interest

Since all the possible known genetic causes of PD were effectively excluded, the remaining 772 variants were then filtered to include only exonic, non-synonymous, heterozygous variants that appeared in the population databases with a MAF < 0.01 (across the entire population and in the African population) (**Table 3.6**) and had a CADD score > 20. Notably, it is imperative to examine both the entire population and the ancestry-matched population, as a variant that seems rare (< 0.01 MAF) across the entire population, may be quite common in a specific population such as those with African-ancestry. The CADD score was chosen as a prioritisation parameter as it is a cumulative predictive score that does not only rely on the scores produced by other algorithms, but also, employs less biased, larger training sets to improve pathogenicity prediction accuracy (Rentzsch *et al.*, 2019). A scaled CADD score of > 20 represents the top 1% of deleterious variants in the genome (Rentzsch *et al.*, 2019). This filtering approach yielded a total of 68 variants.

3.3.3.1 Gene expression and pathway analysis revealed that 24 variants were expressed in the brain

To further identify the best candidate variants for PD, they were then subjected to gene expression analysis to identify variants that are in ubiquitously-expressed genes that are also highly expressed in neurological tissue. Of the 68 genes, only 24 variants of interest were in genes found to be expressed in the brain/nervous system. Subsequently, it was also determined if any of the 24 genes (i) were implicated in neuro-related pathways or diseases, (ii) were involved in protein-protein networks involving PD genes or (iii) were implicated in biological processes of interest. Some of these genes were involved in relevant pathways and diseases, and these findings are presented in **Appendix G**.

3.3.3.2 In-silico pathogenicity prediction scores found 13 variants to be pathogenic across >5 pathogenicity prediction tools

Thereafter, the 24 variants were subjected to a variety of *in silico* pathogenicity prediction scoring to determine the impact of the mutation on protein function and to determine if the mutation occurs on a highly conserved residue (**Table 3.5**). No variant was found to be deleterious across all the pathogenicity predictors, highlighting the necessity of using multiple predictive algorithms to assess the variants. The Mutation Assessor tool found no variants to be considered highly deleterious, even among the three variants with a CADD score > 30 (representing the top 0.1% of deleterious variants). Fifteen of the 24 variants had a GERP++ score of > 4, indicating the variant resides in a region where fewer substitutions are occurring than usual, thus indicating higher evolutionary constraint (Huber *et al.*, 2020). A few studies have prioritised the deleteriousness of a variant if it is found to be pathogenic by 5 or more predictors (Quadri *et al.*, 2018, Ruis-Martinez *et al.*, 2017). Thirteen of the 24 variants fulfilled these criteria (highlighted in grey in **Table 3.5**), however, no variants were eliminated prior to Sanger sequencing validation and private cohort screening.

The MAF for these 24 variants (across multiple publicly available databases) is shown in **Table 3.6**. None of the variants appeared in any of the population databases with a MAF > 0.01. This indicated a high level of rarity in both the global and African-specific subdivisions of the population databases. Three of the variants in *EIF2A*, *KLHL35* and *MZFI*, were not found in any of the databases.

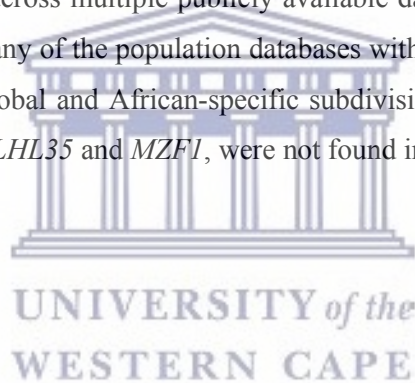


Table 3.5: *In silico* pathogenicity prediction scores across the prioritised 24 variants with a CADD score > 20 and expressed in neuro-specific tissue

Variant Identifiers			<i>In Silico</i> Pathogenicity Prediction Scores																
Gene Symbol	rsID	Amino acid change	CADD Score	SIFT	PP-2	PROVEAN	MT	MA	M-CAP	LRT	fathmm	gERP++							
<i>DNAH5</i>	rs113742238	p.R3077Q	34,0	0.011 D	0.998 D	-3.98 D	1 D	3.05 M	0.075 D	0.000 D	0.75 T	5.79							
<i>NPHP3</i>	rs111727307	p.R1167H	33,0	0.001 D	0.991 D	-2.99 D	1.000 D	1.99 M	0.098 D	0.000 D	-3.58 D	5.6							
<i>DNAH10</i>	rs186639935	p.E698G	32,0	0.006 D	0.952 P	-5.42 D	1.000 D	3.005 M	0.035 D	0.013 N	0.36 T	5.53							
<i>FRMD4B</i>	rs144459338	p.R360W	26,5	0.0 D	1.0 D	-5.82 D	0.999 D	1.955 M	0.302 D	0.000 D	-2.44 D	2.86							
<i>DAAM2</i>	rs375083979	p.R209G	25,0	0.405 T	0.615 P	-0.95 N	0.993 D	0.69 N	0.070 D	0.000 D	-2.19 D	5.52							
<i>CLSTN2</i>	rs147617850	p.D289N	24,4	0.115 T	0.984 D	-3.21 D	1.000 D	1.67 L	0.044 D	0.000 D	0.08 T	5.2							
<i>STAC</i>	rs111403865	p.P103L	23,9	0.075 T	0.906 P	-2.7 D	1.000 D	1.295 L	0.021 T	0.000 D	-0.94 T	4.63							
<i>CLSTN2</i>	rs140202819	p.E910K	23,8	0.067 T	0.804 P	-1.76 N	1.000 D	1.735 L	0.009 T	0.004 N	1.3 T	5.24							
<i>MRE11</i>	.N/A	p.E406K	23,8	0.536 T	0.003 B	-0.75 N	1.000 D	0.74 N	0.034 D	0.000 D	-1.01 T	4.96							
<i>CD47</i>	rs761086667	p.A252S	23,8	0.25 T	0.294 B	-1.05 N	0.854 N	0.775 N	0.010 T	0.037 N	. .	1.7							
<i>ZDHHC11</i>	rs528116435	p.R276P	23,7	0.002 D	1.0 D	-4.1 D	1.000 D	0 N	. .	0.006 D	1.41 T	-0.417							
<i>KNTC1</i>	rs141767241	p.A1083T	23,6	0.006 D	0.912 P	-2.31 N	1.000 D	2.57 M	0.031 D	0.000 D	2.15 T	5.91							
<i>MANF</i>	rs545661735	p.A13V	23,5	0.02 D	0.005 B	. .	0.897 D	0.895 L	0.239 D	0.000 D	. .	4.08							
<i>AHNAK2</i>	rs776830611	p.D1540H	23,4	0.005 D	1.0 D	-3.69 D	1 D	2.64 M	0.004 T	. .	5.03 T	4.16							
<i>EIF2A</i>	rs561839835	p.A143V	23,3	0.132 T	0.038 B	-2.63 D	1.000 D	1.095 L	0.004 T	0.000 D	1.01 T	6.17							
<i>KLHL35</i>	.	p.R179C	23,3	0.121 T	. .	0.57 N	0.953 D	. .	0.902 D	0.132 N	-0.39 T	3.02							
<i>FAM149B1</i>	rs377021877	p.I149M	23,2	0.003 D	0.991 D	-1.67 N	0.977 N	2.25 M	0.039 D	0.000 D	0.96 T	-0.081							

<i>CX3CR1</i>	rs137947370	p.A313V	23,1	0.017	D	0.001	B	-1.5	N	1	N	2.06	M	0.006	T	0.080	N	1.24	T	5.91
<i>SALL3</i>	rs150707152	p.R1012Q	23,0	0.004	D	0.996	D	-1.68	N	1	D	1.355	L	0.021	T	0.000	D	2.18	T	5.1
<i>MZF1</i>	.N/A	p.S721R	23,0	1.0	T	0.43	B	-0.21	N	0.822	N	1.295	L	0.009	T	0.001	D	1.62	T	-0.46 1
<i>AHNAK2</i>	rs11852016	p.P1711L	23,0	0.007	D	1.0	D	-6.26	D	1	N	2.995	M	0.003	T	.	.	4.42	T	2.86
<i>ZNF418</i>	rs201309448	p.R667G	23,0	0.009	D	0.99	D	-3.96	D	1	N	2.72	M	0.003	T	.	.	2.93	T	1.7
<i>IL3RA</i>	rs776812933	p.S91C	22,9	0.117	T	0.998	D	-1.75	N	1	N	0.975	L	0.002	T	0.055	U	1.33	T	0.364
<i>NPHP3</i>	rs113364886	p.F1324S	22.5	0.081	T	0.013	B	-0.75	N	0.902	D	1.7	L	0.054	D	0.000	D	-2.97	D	5.93

Variants are provided in order of the CADD score (highest to lowest).

CADD = Combined Annotation Dependent Depletion score; SIFT = Sorting Intolerant From Tolerant score; PP-2 = Polyphen-2 score; PROVEAN = Protein Variation Effect analyser; MT: Mutation Taster score; MA = Mutation Assessor score; M-CAP = Mendelian Clinically Applicable Pathogenicity; LRT = Likelihood Ratios Test; fathmm = Functional Analysis through Hidden Markov Models score and gERP++ = Genome Evolutionary Rate Profiling.

B = Benign; D = Deleterious; T = Tolerated; N = Neutral; M = Moderate; L = Low; LB = Likely benign; LP = Likely pathogenic.

Genes highlighted in grey are considered to be deleterious across > 5 pathogenicity prediction tools.

Table 3.6: MAF population frequencies, in public databases, for the 24 prioritised variants with a CADD score > 20 and expressed in neuro-specific tissue

Variant Identifiers			Database Population Frequencies							
Gene Symbol	rsID	Amino acid change	1000G (all)	1000G (AFR)	ExAC (all)	ExAC (AFR)	gnomAD (exome_ALL)	gnomAD (exome_AFR)	gnomAD (genome_all)	gnomAD (genome_AFR)
<i>DNAH5</i>	rs113742238	p.R3077Q	0.0006	0.0023	0.0002	0.0017	0.0001	0.0018	0.0005	0.0018
<i>NPHP3</i>	rs111727307	p.R1167H	0.003	0.0091	0.0004	0.0038	0.0003	0.0041	0.0010	0.0033
<i>DNAH10</i>	rs186639935	p.E698G	0.0002	0.0008	0.008323	0.0001	0	0	0.0056	0.0002
<i>FRMD4B</i>	rs144459338	p.R360W	0.0016	0.0061	0.0002	0.0014	0.0002	0.0012	0.0003	0.0008
<i>DAAM2</i>	rs375083979	p.R209G	0.0008	0.0023	0.0001	0.0007	0.0002	0.0008	0.0002	0.0006
<i>CLSTN2</i>	rs147617850	p.D289N	0	0	0.00411	0.00961	0,005693	0,00353	0,00645	0.0002

<i>STAC</i>	rs111403865	p.P103L	0	0	0.004946	0.0006	0,00243	0.0004	0.0002	0.0005
<i>CLSTN2</i>	rs140202819	p.E910K	0.0022	0.0083	0.0006	0.0070	0.0005	0.0065	0.0017	0.0060
<i>MRE11</i>	.	p.E406K	0	0	0	0	0,000000	0,000000	0,00660	0
<i>CD47</i>	rs761086667	p.A252S	0	0	0.00386	0.0003	0.0058	0,00970	0,00323	0.0001
<i>ZDHHC11</i>	rs528116435	p.R276P	0.0012	0	0.0102	0.0058	0.0009	0,006390	0.0027	0.0004
<i>KNTC1</i>	rs141767241	p.A1083T	0.0016	0.0061	0.0005	0.0066	0.0005	0.0067	0.0021	0.0075
<i>MANF</i>	rs545661735	p.A13V	0.0002	0.0008	0	0	0	0	0	0
<i>AHNAK2</i>	rs776830611	p.D1540H	0	0	0.0002	0.0004	0,00579	0.0003	0.0005	0.0018
<i>EIF2A</i>	rs561839835	p.A143V	0	0	0	0	0	0	0	0
<i>KLHL35</i>	.	p.R179C	0	0	0	0	0	0	0	0
<i>FAM149B</i> <i>1</i>	rs377021877	p.I149M	0	0	0.0004	0	0.0003	0,000000	0,00322	0.0001
<i>CX3CR1</i>	rs137947370	p.A313V	0.0014	0.0053	0.0001	0.0014	0.0002	0.0022	0.0004	0.0013
<i>SALL3</i>	rs150707152	p.R1012Q	0	0	0.00248	0.0002	0,001219	0,006545	0,006467	0.0002
<i>MZF1</i>	.	p.S721R	0	0	0	0	0	0	0	0
<i>AHNAK2</i>	rs11852016	p.P1711L	0.0012	0.0045	0.0001	0.0012	0,00827	0.0012	0.0004	0.0015
<i>ZNF418</i>	rs201309448	p.R667G	0.0014	0.0045	0.0004	0.0025	0.0003	0.0022	0.0005	0.0016
<i>IL3RA</i>	rs776812933	p.S91C	0	0	0.00411	0.0005	0,002030	0,00640	0,0060	0.0002
<i>NPHP3</i>	rs113364886	p.F1324S	0.003	0.0091	0.0004	0.0036	0.0003	0.0039	0.0010	0.0033

1000G = 1000 Genomes Project; ExAC = Exome Aggregation Consortium; GnomAD = Genome Aggregation Database.
 ALL = All individuals sequenced in population; AFR = All African/African American individuals sequenced in population.
 Genes highlighted in orange were not present in any of the population databases.

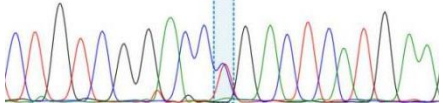
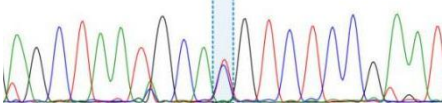
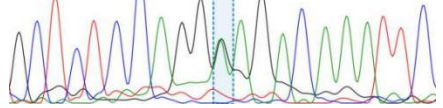
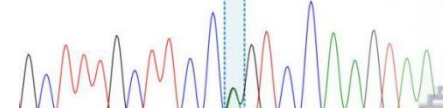
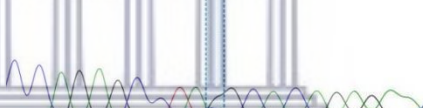
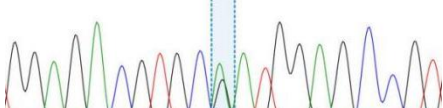
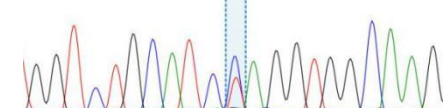
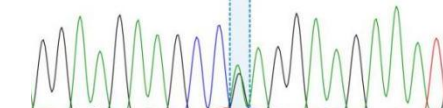
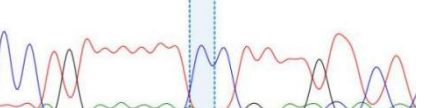
3.3.4 Co-segregation analysis in family ZA 15

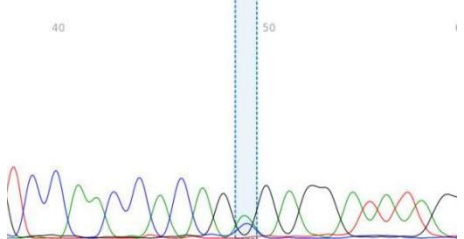
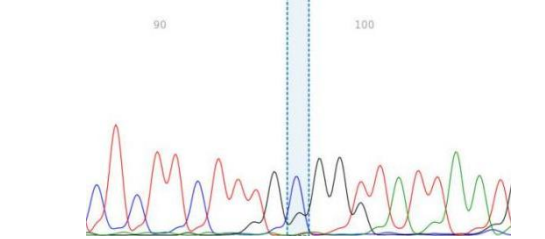
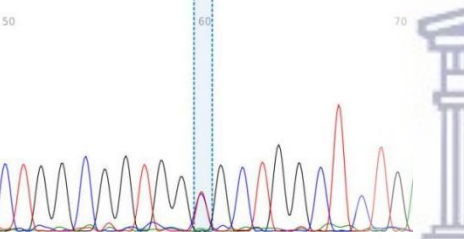
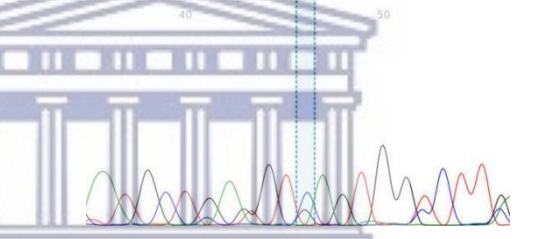
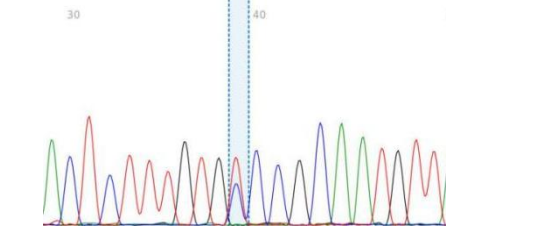
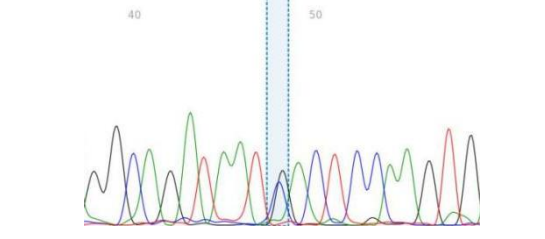
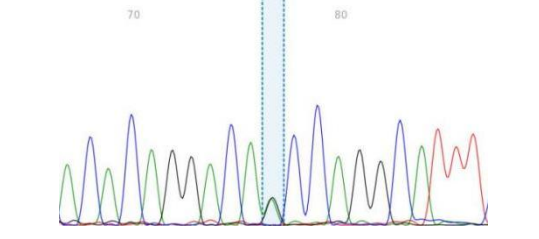
3.3.4.1 Sanger sequencing revealed that 20 variants co-segregated in family ZA 15

To further confirm our NGS results and the co-segregation of the variants within this family, Sanger sequencing was performed on the 24 variants for all 6 family members. The sequencing chromatograms generated for the proband (74.53) are shown in **Table 3.7**. Three of the variants were unable to be sequenced and may be attributed to the presence of repeat expansions around the variant, preventing primer hybridisation (in *KNTC1* and *MZF1*) (Hommelsheim *et al.*, 2014), or primer design failure due to the presence of several pseudogenes, as confirmed after blasting the sequence against the reference genome (for *KLHL35*) (Chen *et al.*, 2011). Notably, the variant in *MRE11* was not found to be present in the proband or any of the family members and was excluded as an NGS artefact which could occur due to DNA deamination, amplification and sequencing error (Deans *et al.*, 2017). The remaining 20 variants were all validated by Sanger sequencing, as illustrated in **Table 3.7**.

In the co-segregation analysis in the family members, 8 variants (in the *DNAH5*, *FRMD4B*, *DAAM2*, *CD47*, *ZDHHC11*, *EIF2A*, *ZNF418*, *IL3RA* genes) were found to be present in the affected individuals only, while 12 of the variants (in *NPHP3*, *DNAH10*, *CLSTN2*, *STAC*, *CLSTN2*, *MANF*, *AHNAK2*, *FAM149B1*, *CX3CR1*, *SALL3*, *AHNAK2*, *NPHP3A*) were found to be present in both the affected individuals, as well as, the affected female's daughter (**Table 3.8**). However, as the daughter has not shown any symptoms of early-onset PD and may develop PD later on (she is currently 40 years old), we cannot use this finding as a criterion to exclude variants. One of these variants (*CLSTN2* p.E910K) was also found to be present in the proband's wife who is an unaffected individual of Xhosa descent, thus this variant was excluded.

Table 3.7: Sanger sequencing of the 24 prioritised variants in the proband (ID 43.59/74.53)

<p>Gene Nucleotide Change: Reference Sequence: Chromatogram:</p>	<p><i>DNAH5</i> C>T CTGTCGGACCCGACTCATGAA C T G T C G G A C C Y G A C T C A T G A A 50 60</p> 	<p><i>NPHP3</i> C>T AGCTAATGCACGTCTCCGAAT A G C T A A T G C A Y G T C T C C G A A T 130 140 150</p> 	<p><i>DNAH10</i> A>G GCTCTCCAGGAAGACAAATTC G C T C T C C A G G Y A G A C A A A T T C 100 110</p> 
<p>Gene Nucleotide Change: Reference Sequence: Chromatogram:</p>	<p><i>FRMD4B</i> G>A GCTTTGCTTCCGGTGTCCAAGTAA G C T T T G C T T C C Y G T C C A A G T A A 50 60 70</p> 	<p><i>DAAM2</i> C>G CCAGAGCCTACGCACAGAGAA C C A G A G C C T A Y G C A C A G A G A A 70 80</p> 	<p><i>CLSTN2</i> G>A CTGGAGACGTGCGATGGAGCCGTGTC G G A G A C G T G C Y A T G G A G C C G T 120 130</p> 
<p>Gene Nucleotide Change: Reference Sequence: Chromatogram:</p>	<p><i>STAC</i> C>T CTGGTCTGCATCCAGGTGGCAAGGCT G G T C T G C A T C Y A G G T G G C A A G 120 130</p> 	<p><i>CLSTN2</i> G>A GGAGGAAGAAGCCGAGGAAGAAATGA G G A A G A A G C C Y A G G A A G A A A T 100 110</p> 	<p><i>MRE11</i> Wildtype variant present ACCTGTTTTTTCCTTTTGTCTCTATG C C T G T T T T T C C T T T T G T C T A T G 140 150</p> 

<p>Gene</p> <p>Nucleotide Change:</p> <p>Reference Sequence:</p> <p>Chromatogram:</p>	<p>CD47</p> <p>C>A</p> <p>TCCAACCACAGGGAGGATATAGG</p> <p>T C C A A C C A C A G K G A G G A T A T A G</p> 	<p>ZDHHC11</p> <p>C>G</p> <p>AACTCTCTTCTTTGCGGTTATTAATGA</p> <p>C T C T T C T T T S G G M T A T T A A T</p> 	<p>KNTC1</p> <p>Sequencing did not work</p> <p>GAACATCAAAAACAGCACTGAAAAAATG</p> <p>N/A</p>
<p>Gene</p> <p>Nucleotide Change:</p> <p>Reference Sequence:</p> <p>Chromatogram:</p>	<p>MANF</p> <p>C>T</p> <p>GGCTGGCGGTGGCTGGCTCTGAGCG</p> <p>C T G G C G G T G G R G C T G G C T C T G</p> 	<p>AHNAK2_A</p> <p>C>G</p> <p>TGTATGCTCAGGTAGTGGCCTTGAGG</p> <p>A T G C T G A K G T S A G T G G Y C T T R</p> 	<p>EIF2A</p> <p>C>T</p> <p>TGAAACTCTTTGTGCCCAATGTTAAC</p> <p>A C T C T T T G T R C C G C A A T G T T</p> 
<p>Gene</p> <p>Nucleotide Change:</p> <p>Reference Sequence:</p> <p>Chromatogram:</p>	<p>KLHL35</p> <p>Sequencing did not work</p> <p>AAGGCCTGACGCAGGACGCGGC</p> <p>N/A</p>	<p>FAM149B1</p> <p>C>G</p> <p>GTAGGCAGATAATCACTCCAAGTGAAG</p> <p>G G C A G A T A A T S A C T C C A A G T G</p> 	<p>CX3CR1</p> <p>G>A</p> <p>ACACAGGACAGCCAGGCATTT</p> <p>A C A C A G G A C A Y C C A G G C A T T T</p> 

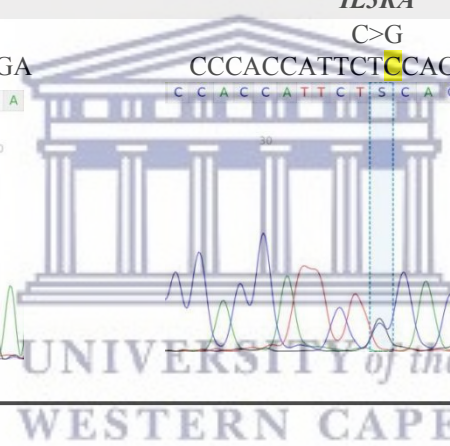
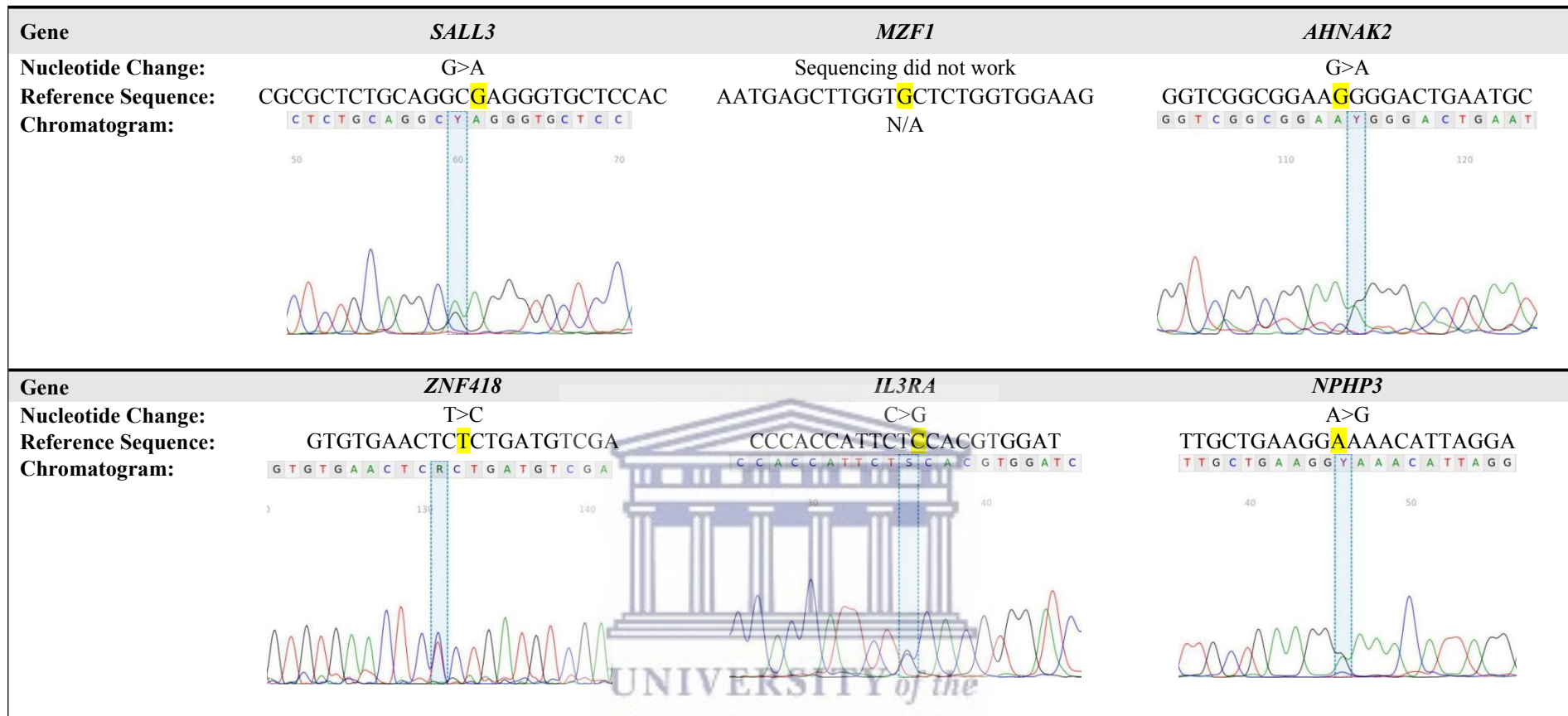


Table 3.8: Co-segregation of the variants in the ZA 15 family members

Gene Symbol	ZA 15 Family Members (Lab IDs)						Gene Symbol	ZA 15 Family Members (Lab IDs)					
	74.53 (Affected)	74.54 (Affected)	74.52	74.51	74.50	11.915 (Spouse)		74.53 (Affected)	74.54 (Affected)	74.52	74.51	74.50	11.915 (Spouse)
<i>DNAH5</i>	Yes	Yes	No	No	No	No	<i>MANF</i>	Yes	Yes	No	No	Yes	No
<i>NPHP3</i>	Yes	Yes	No	No	Yes	No	<i>AHNAK2</i>	Yes	Yes	No	No	Yes	No
<i>DNAH10</i>	Yes	Yes	No	No	Yes	No	<i>EIF2A</i>	Yes	Yes	No	No	No	No
<i>FRMD4B</i>	Yes	Yes	No	No	No	No	<i>KLHL35</i>	N/A	N/A	N/A	N/A	N/A	N/A
<i>DAAM2</i>	Yes	Yes	No	No	No	No	<i>FAM149B1</i>	Yes	Yes	No	No	Yes	No
<i>CLSTN2</i>	Yes	Yes	No	No	Yes	No	<i>CX3CR1</i>	Yes	Yes	No	No	Yes	No
<i>STAC</i>	Yes	Yes	No	No	Yes	No	<i>SALL3</i>	Yes	Yes	No	No	Yes	No
<i>CLSTN2</i>	Yes	Yes	No	No	Yes	Yes	<i>MZF1</i>	N/A	N/A	N/A	N/A	N/A	N/A
<i>MRE11</i>	No	No	No	No	No	No	<i>AHNAK2</i>	Yes	Yes	No	No	Yes	No
<i>CD47</i>	Yes	Yes	No	No	No	No	<i>ZNF418</i>	Yes	Yes	No	No	No	No
<i>ZDHHC11</i>	Yes	Yes	No	No	No	No	<i>IL3RA</i>	Yes	Yes	No	No	No	No
<i>KNTC1</i>	N/A	N/A	N/A	N/A	N/A	N/A	<i>NPHP3</i>	Yes	Yes	No	No	Yes	No

Green cells indicate the presence of the variant in the family member; Orange cells indicate the absence of a variant in an affected family member.

3.3.5 Allele frequencies of variants in private PD and non-PD cohorts

3.3.5.1 Private cohort screening further reduced the number of candidates to 3 variants

The 23 variants (excluding the *CLSTN2* (rs140202819) variant found in the proband's spouse) were screened through a large Schizophrenia Xhosa cohort and variants exhibiting a MAF > 0.01 were excluded. Thirteen of the variants (in *CX3CR1*, *DAAM2*, *DNAH10*, *DNAH5*, *FAM149B1*, *FRMD4B*, *IL3RA*, *KNTC1*, *MZF1*, *NPHP3*, *NPHP3*, *SALL3* and *ZNF418*) had MAFs > 0.01 either in cases or controls (first row of **Table 3.9**), and were thus excluded from further analysis. Notably, the results from this initial screening indicated that although these variants were considered rare across all population databases, about half of the variants could be considered commonly occurring among the Xhosa-ancestry population group. This further highlights the need for expanded public population allele frequency databases to include more non-European populations.

Thereafter, the final 10 variants were subjected to screening in two smaller cohorts including; (i) the TB study consisting of 161 Xhosa individuals (unpublished data) and (ii) the H3Africa Baylor Dataset comprising 386 African (non-Xhosa) individuals (Choudhury *et al.*, 2021). Notably, in the TB study cohort, 6 of the variants (in *AHNAK2*, *CD47*, *CLSTN2*, *EIF2A*, *KLHL35*, *STAC*) were found to be prevalent with MAFs > 0.5 (**Table 3.9**). This finding further illustrates the notion of ancestry-specific variants that may skew variant filtering analysis, if not considered. For the H3Africa study cohort, only three of the variants were found with MAFs < 0.01 and were still considered rare, possibly due to the lack of Xhosa-ancestry individuals in this collection.

However, in addition to the cohort screening above, we also screened the 10 variants in three PD-specific cohorts to determine if any variants were found in any other PD/movement disorder patients. However, none of our variants was found in these cohorts (**Table 3.10**). As the Queensland Parkinson's Project and the Mayo Clinic PD and LB cohorts consisted of individuals of European descent, it was unlikely for these variants to have been common in either group. In the third PD cohort, the French and Mediterranean Parkinson's Disease Genetics Study group (FMPD cohort) which consisted of both European and North African ancestries, 2 PD patients presented with the rs776830611 variant) in the *AHNAK2* gene. However, it was found that the gene is polymorphic, as a large number of variants were found, but it could not yet be removed as a variant of interest due to the low MAF.

In summary, from the cohort screening (**Table 3.9 and 3.10**), all variants with $MAF < 0.01$ in non-PD (population) cohorts and that were present in other individuals with PD, were selected for further study. Thus, 4 variants in *AHNAK2*, *MANF* and *ZDHHC11* and *MRE11* met these criteria. However, since the *MRE11* variant was not found in the proband during Sanger sequencing, this variant was omitted from subsequent analysis. Interestingly, further co-segregation analyses on *MANF* and *ZDHHC11* (using families recruited to the FMPD cohort) both indicated the presence of variants (not the variants found in this study) in PD-affected individuals that belonged to PD-affected families, thus, indicating a potential role of these genes in relation to PD onset.



Table 3.9: Variant screening in various ‘non-PD’ private population cohorts

Cohort	Number of Study Participants	Origin of Study Participants	Sequencing Method	Recruitment criteria	Variant Screening Results		Reference
					Gene Symbol	Allele Frequency (AF)	
Schizophrenia Xhosa Cohort	n = 909 cases/ n = 917 controls	Western Cape and Eastern Cape, South Africa	WES	Cases included individuals diagnosed with schizophrenia; controls did not have neurological conditions Cases and controls were matched for age, gender, education, and region of recruitment	<i>AHNAK2</i> (rs776830611)	Not present	(Gulsuner <i>et al.</i> , 2020)
					<i>AHNAK2</i> (rs11852016)	Not present	
					<i>CD47</i>	Cases: 0.0050/Controls: 0.0093	
					<i>CLSTN2</i> (rs147617850)	Cases: 0.0066/Controls: 0.0076	
					<i>CX3CRI</i>	Cases: 0.0242/Controls: 0.0251	
					<i>DAAM2</i>	Cases: 0.0088/Controls: 0.0109	
					<i>DNAH10</i>	Cases: 0.0127/Controls: 0.0087	
					<i>DNAH5</i>	Cases: 0.0149/Controls: 0.0093	
					<i>EIF2A</i>	Not present	
					<i>FAM149B1</i>	Cases: 0.0435/Controls: 0.0153	
					<i>FRMD4B</i>	Cases: 0.0132/Controls: 0.0153	
					<i>IL3RA</i>	Cases: 0.0937/Controls: 0.0523	
					<i>KLHL35</i>	Not present	
					<i>KNTC1</i>	Cases: 0.0143/Controls: 0.0115	
					<i>MANF</i>	Cases: 0.0017/Controls: 0.0005	
					<i>MRE11</i>	Not present	
					TB Xhosa Cohort (ResisTB study)	n = 161 (Mostly individuals who self-identified as Xhosa)	
<i>AHNAK2</i> (rs11852016)	G:0 A:1						
<i>CD47</i>	C:0.5 A:0.5						
<i>CLSTN2</i> (rs147617850)	G:0.5 A:0.5						
<i>MZF1</i>	Cases: 0.0193/Controls: 0.0262						
<i>NPHP3</i>	Cases: 0.0666/Controls: 0.0747						
<i>NPHP3</i>	Cases: 0.0677/Controls: 0.0763						
<i>SALL3</i>	Cases: 0.0462/Controls: 0.0540						
<i>STAC</i>	Cases: 0.0077/Controls: 0.0082						
<i>ZDHHC11</i>	Not present						
<i>ZNF418</i>	Cases: 0.0292/Controls: 0.0284						

	non-South African individuals have also been included in this cohort		No post-WGS screening was done	<i>EIF2A</i> <i>KLHL35</i> <i>MANF</i> <i>MRE11</i> <i>STAC</i> <i>ZDHHC11</i>	C:0.5 T:0.5 G:0.5 A:0.5 Not present Not present C:0.5 T:0.5 Not present	
African cohort n= 386 (H3Africa Baylor Dataset without Mali and SAHGP)	H3Africa Biobank samples - no provincial level information available	WGS (high and medium coverage)	All samples from Benin were all sickle cell positive Half of the samples from Cameroon were sickle cell carriers The rest did not report any disease	<i>AHNAK2</i> (rs776830611) <i>AHNAK2</i> (rs11852016) <i>CD47</i> <i>CLSTN2</i> (rs147617850) <i>EIF2A</i> <i>KLHL35</i> <i>MANF</i> <i>MRE11</i> <i>STAC</i> <i>ZDHHC11</i>	Not present Not present Not present Not present AC=1; AN=772; AF=0.00129534 Not present Not present AC=2; AN=772; AF=0.00259067 AC=4; AN=772; AF=0.00518135	(Choudhury <i>et al.</i> , 2021)

Rows highlighted in green indicate that the variant in the gene presented with a MAF < 0.01 within the cohort.
Rows highlighted in orange indicate that the variant was found to be rare (MAF < 0.01) across all cohorts.

Table 3.10: Variant screening in various PD-specific private population cohorts

Cohort	Number of Study Participants	Origin of Study Participants	Sequencing Method	Recruitment criteria	Variant Screening Results		Reference
					Gene Symbol	Allele Frequency (AF)	
Queensland Parkinson's Project	n = 66	Queensland, Australia	WES	23 separate kindreds, containing 47 PD cases, 5 family cases with 'other' movement disorders (e.g. dystonia), and 14 family controls (Average Age last seen – 64 yrs).	<i>AHNAK2</i> <i>AHNAK2</i> <i>CD47</i> <i>CLSTN2</i> <i>EIF2A</i> <i>KLHL35</i> <i>MANF</i> <i>MRE11</i> <i>STAC</i> <i>ZDHHC11</i>	Not present Not present Not present Not present Not present Not present Not present Not present Not present Not present	(Bentley <i>et al.</i> , 2021)
Mayo Clinic PD and LB Exomes			WES	Familial PD and Lewy body disease cases	<i>AHNAK2</i> <i>AHNAK2</i> <i>CD47</i> <i>CLSTN2</i> <i>EIF2A</i> <i>KLHL35</i> <i>MANF</i> <i>MRE11</i> <i>STAC</i> <i>ZDHHC11</i>	Not present Not present Not present Not present Not present Not present Not present Not present Not present Not present	Owen Ross (personal communication)
French and Mediterranean Parkinson's Disease Genetics 4 Study group (FMPD cohort)	n = 1319 (Multi-ethnicities, particularly Caucasians (66%) and North-Africans (10%))	Europe, North Africa, South Africa	WES	Probands with familial PD or parkinsonism, patients with sporadic PD	<i>AHNAK2</i> <i>AHNAK2</i> <i>CD47</i> <i>CLSTN2</i> <i>EIF2A</i> <i>KLHL35</i> <i>MANF</i> <i>MRE11</i> <i>STAC</i> <i>ZDHHC11</i>	Present in 2 patients Not present Not present Not present Not present Not present Not present Not present Not present Not present	Fevga <i>et al.</i> , 2022

The row highlighted in yellow indicates the presence of the variant in the cohort.

3.3.5.1.1 Screening of ethnic-matched controls revealed that the top three candidate variants were not present in the Xhosa population

Finally, the top 3 prioritised variants were screened in 100 Xhosa controls from our South African PD study collection, using Sanger sequencing. A control group of 100 allows for the removal of common polymorphisms at a frequency of 0.5% or more (typically polymorphisms are defined at a frequency of 1% or more). These controls had been recruited from the same region and ancestry-group in South Africa as the family, and would therefore be a more appropriate ethnic match. None of the 3 variants was present in these controls, indicating the rarity of these variants in an ancestry-matched population.

3.3.6 Prioritising a single variant in family ZA 15 for further *in-silico* protein analysis resulted in the nomination of the p.A13V variant in MANF

The three variants were then analysed to select the best candidate for further follow-up studies. Each of the three variants (**Table 3.11**) was re-examined according to their gene expression annotations to determine if there was an association with PD. Previously annotated gene expression data from the HPA, UniProt and GO terms were examined. Furthermore, information on animal knockout effects for each gene was retrieved from the Flybase and MGI gene knockout servers.

All 3 of the variants were previously found to have functional effects across more than 5 pathogenicity predictors (including CADD, SIFT, PROVEAN, MutationTaster, LRT and GERP++ and were considered ‘likely pathogenic’ post-ACMG-guidelines classification and co-segregation analysis on wIntervar (<https://wintervar.wglab.org/>) (**Appendix J**; p.A13V in *MANF*). Thus, it was necessary to determine whether any of the variants had been associated with PD-specific gene expression regions, pathways and biological processes.

The variants were compared and although all the genes were expressed in the brain, one gene in particular, *MANF*, was found to be expressed in the SNpc (**Table 3.11**), a key neurological region implicated in PD (Surmeier, 2018). UniProt also provides a series of functional information that is corroborated by actual functional studies. *ZDHHC11* was found to ‘influence endoplasmic reticulum-localised palmitoyltransferase that could catalyse the addition of palmitate onto various protein substrates and potentially play a role in cell proliferation’. *AHNAK2* lacked any functional information due to the lack of studies on the gene. *MANF* was found to ‘selectively promote the survival of dopaminergic neurons of the ventral midbrain, ‘modulate GABAergic transmission to the dopaminergic neurons of the substantia nigra’ and ‘inhibits cell proliferation and endoplasmic reticulum (ER) stress-induced cell death’.

Animal models depicting gene knockout can provide information on the phenotype the loss of a protein may cause. Mouse knockout models indicated that the *AHNAK2* gene knockout affected the nervous system while the *MANF* gene knockout affected behaviour/and the neurological and nervous system. Furthermore, *MANF* was the only gene that had been studied on the *D. melanogaster* model in the Flybase database and was found to be involved in dopamine metabolic processes and synaptic transmission, neuronal projection development and cellular homeostasis (**Table 3.11**). This information is of importance as it is specifically the loss of dopaminergic neurons that decreases dopamine, a pathological hallmark of PD pathobiology (Surmeier, 2018). Based on this information and its expression in PD-specific brain regions, the *MANF* variant was prioritised for further *in silico* analysis.

Table 3.11: Comparison of the gene expression profiles and functional processes for each gene of interest

Gene Symbol	Gene Expression in the Brain	UniProt	MGI	Flybase
<i>AHNAK2</i> p.D1540H (rs776830611)	Cerebral cortex	N/A	Nervous system	N/A
<i>MANF</i> p.A13V (rs545661735)	Cerebellum, hippocampus, caudate, substantia nigra	Selectively promote the survival of dopaminergic neurons of the ventral mid-brain, modulates GABAergic transmission to the dopaminergic neurons of the substantia nigra and inhibits cell proliferation and endoplasmic reticulum (ER) stress-induced cell death	Behaviour/neurological, homeostasis metabolism, nervous system, growth/size/body, cellular	Dopamine metabolic process, neuron cellular homeostasis, neuron projection development, dopaminergic synaptic transmission
<i>ZDHHC11</i> p.R276P (rs528116435)	Cerebellum	Endoplasmic reticulum-localised palmitoyltransferase could catalyse the addition of palmitate onto various protein substrates and potentially play a role in cell proliferation	Homeostasis/metabolism, immune system, hearing/vestibular	N/A

3.4 Conclusion

WES analysis in a South African Xhosa family affected with PD ultimately yielded 24 variants in 22 genes that were expressed in the brain, considered to be deleterious and predicted to be rare in all population databases (**Figure 3.5**). Sanger sequencing confirmed 20 of the 24 variants, indicating the variants were co-segregating within the family and were not sequencing artefacts. Subsequent screening of these variants through private, and some ancestry-matched cohorts, found several variants to have high MAFs (> 0.01), thus allowing for the exclusion of these variants from further analysis. This emphasised the importance of observing the allele frequencies of rare variants in ethnically-matched cohorts. Based on MANF's function and gene expression profile in relation to PD, the p.A13V variant was prioritised, and will undergo further *in silico* analysis to determine the functional impact of this variant on the protein and if it could be potentially disease-causing.



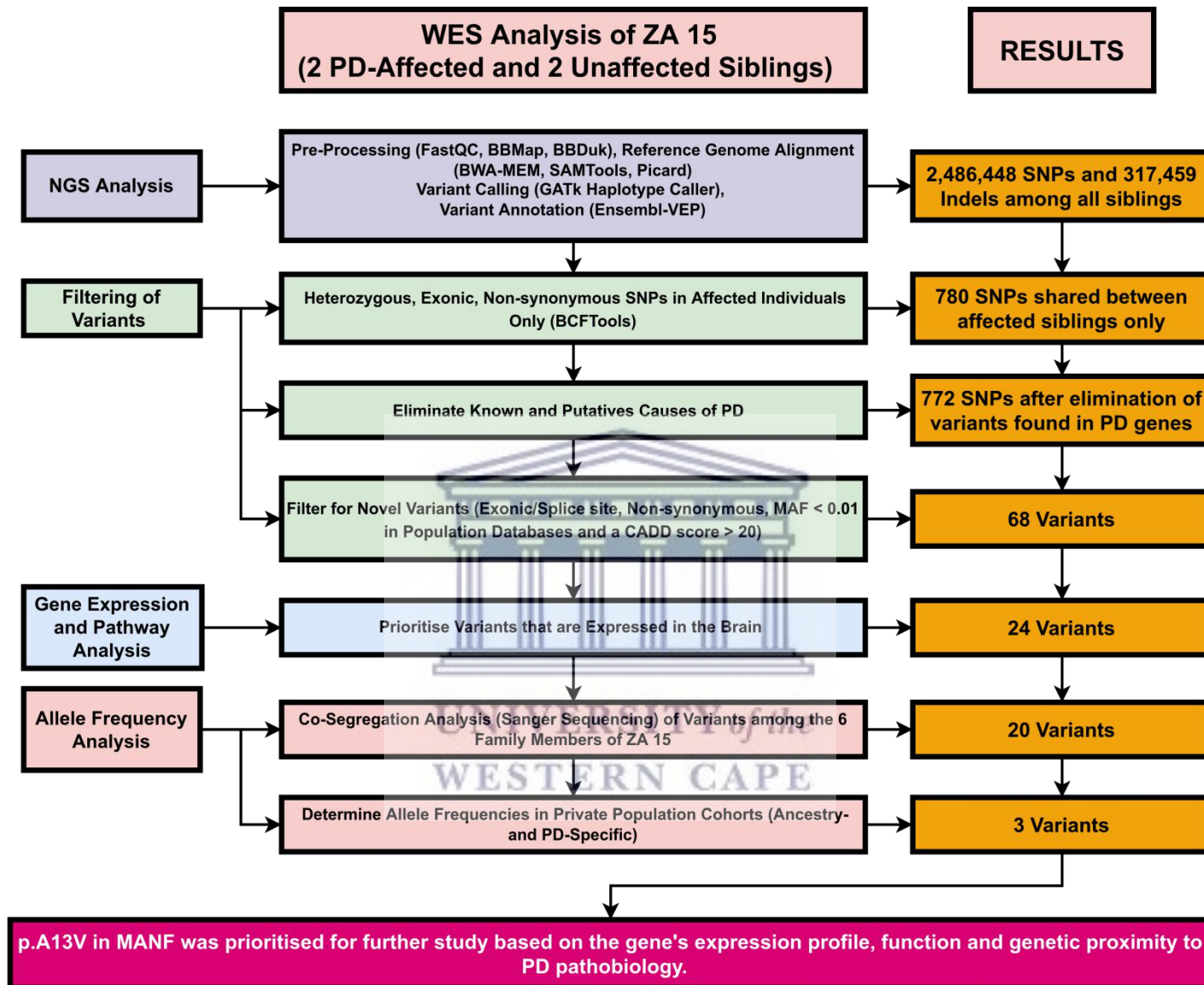


Figure 3.5: Summary of the main findings of the present study based on WES analysis of ZA 15

In-silico mutation analysis of p.A13V in mesencephalic astrocyte-derived neurotrophic factor (MANF)

Abstract

Introduction: Parkinson's disease is a multi-genic neurodegenerative disease that is characterised by the loss of dopaminergic neurons in the *substantia nigra pars compacta* (SNpc). This deterioration results in a progressive decrease of available dopamine that presents as a host of both motor and non-motor symptoms. We investigated a PD-affected family with African Xhosa ancestry to determine a potential novel genetic cause of the disease using next-generation sequencing (NGS). Whole exome sequencing (WES) analysis and subsequent population screening of the PD-affected family yielded variants that were further prioritised based on various filtering criteria. These criteria also included analysis of gene and protein expression information to determine involvement with PD-linked brain regions. Ultimately, a single variant, p.A13V in mesencephalic astrocyte-derived neurotrophic factor (MANF) was earmarked for further analysis after specific filtering. In lieu of (and as a precursor to) traditional 'wet-lab' functional analysis, the aim of this study was to determine p.A13V's impact on MANF's protein structure/function through bioinformatic *in silico* analysis using readily available online tools.

Methods and Results: Evolutionary conservation analysis of the MANF protein's amino acids (AA) was performed using the ConSurf web server revealing that the p.A13V variant occurs as a buried residue (i.e. occurring on the interior of the protein and is not exposed to the surrounding solvent). The protein sequence was then analysed to determine the effect of the variant using a predictive approach through Project HOPE, I-Mutant 3.0 and MUpro. The protein sequence containing the variant was then subjected to secondary structure analysis using PredictProtein, SignalP 3.0 and Phobius. Secondary structure analysis confirmed that the variant was a buried residue. Importantly, this analysis also showed the variant falls within the signal peptide of the protein and specifically, within the hydrophobic core of the signal peptide of the protein. Furthermore, the variant was predicted to be destabilising at the sequence level with a change in Gibbs free energy ($\Delta\Delta G$) of - 0.2 and - 0.21 obtained from MuPro and I-Mutant 3.0, respectively. Thereafter, theoretical 3-D structural models of the wildtype and mutant proteins were created using I-TASSER, DeepPotential and Robetta. Consequently, these models were scored on TM-Align to select the best ones for downstream analysis. Robetta was found to produce the best theoretical models of all the servers used, according to the scores generated by TM-Align. The models were further quality-checked using Verify3D, Q-MEAN, PROCHECK and ERRAT where they were scored and deemed appropriate for further structural

analysis. Thereafter, the wildtype model was uploaded to DUET, DynaMut and MaestroWeb to determine the effect of p.A13V on the complete theoretical structure. It was predicted that the variant had a destabilising effect on the wildtype structure, while MaestroWeb also indicated an increase in rigidity of the signal peptide, close to the cleavage site. Pymol was used to determine a difference in polar contacts between the wildtype and mutant structure. As valine is larger than alanine, due to the addition of an alkyl group, a decrease in polarity results which may potentially disrupt the function of the hydrophobic core. Molecular dynamics (MD) simulations were performed using the validated theoretical wildtype and mutant models. GROMACS was used to perform the MD simulations and produce root mean square deviations (RMSD), root mean square fluctuations (RMSF) and principal component analysis (PCA) outputs which indicated a deviation in structural conformation and flexibility between the wildtype and mutant models.

Conclusion: *In silico* analysis of the p.A13V variant found it to be destabilising, across all the algorithms used to detect a change in stability, through both sequence-based and structural-based analysis. This indicates that there may be disruption to the hydrophobic core of the signal peptide, as well as the cleavage site and C-terminal of the protein, which may affect protein function due to decreased translocation and incomplete expressivity of the protein. MANF has been found to have neuroprotective properties through the regulation of endoplasmic reticulum (ER) stress, which can promote neuroinflammation and the subsequent death of dopaminergic neurons if left to chronically persist. It is postulated that a damaging variant in MANF may modulate its neuroprotective properties which could result in the deterioration of dopaminergic neurons, as is seen in PD. Our findings indicate that the impact of the p.A13V variant on the MANF protein is potentially significant and may be worth further investigation in wet-laboratory-based analysis to validate the findings from this *in-silico* analysis.

Keywords: *In Silico*; *Ab Initio* Protein Modelling; Signal Peptide; MANF; Hydrophobic Core

4.1 Introduction

NGS analysis can be a useful tool that aids in the discovery of novel sequence variants which could contribute to the onset of the disease of interest. This approach typically yields a large number of variants which are usually condensed after stringent filtering, usually in relation to the investigated disease and the individual. However, these prioritised variants often end up being regarded as variants of unknown significance (VUSs). Before a variant could be regarded as ‘pathogenic’ in a disease, it should be examined through functional-based assays in a ‘wet-laboratory’ to determine whether the variant is capable of producing a biological effect that coincides with the disease phenotype under investigation (Fatkin and Johnson, 2020). The downside to this approach is that these types of functional studies can be time-consuming, expensive and sometimes not feasible. This is especially notable when dealing with many VUSs or when researchers have no definitive evidence as to the variant’s potential effect on protein function (Sosnay and Cutting, 2013).

Missense variants cause alterations in the protein sequence which can have an impact on the charge of the protein, its hydrophobic nature, folding, translation, dynamics, as well as, protein-protein interactions (Iqbal *et al.*, 2020). Thus, computational *in silico* analysis is used as a precursory assessment, prior to functional laboratory methods of analysis. This approach is warranted to study the effects on the protein, thereby delineating variants requiring further analysis. This method often involves the use of machine learning-based methods that are trained on existing, experimentally solved protein structures. Typically, a combination of both sequence and structural analysis should be undertaken to gain a more complete understanding of ways the variant may impact both the stability and function of the protein.

The variant of interest prioritised for *in silico* analysis in this study is the p.A13V non-synonymous variant present in MANF, a hormonal secretory protein. It is a conserved neurotrophic factor protein that displays a protective role on mid-brain dopaminergic neurons (Yu *et al.*, 2021). Typical neurotrophic factors, including brain-derived neurotrophic factor (BDNF; an orthologue of MANF), play an important role in regulating the synthesis, growth, survival and plasticity of neurons while MANF also regulates ER stress (Jääntti and Harvey, 2020). The gene is expressed in various tissues and is abundantly expressed in several brain regions including the SNpc. The SNpc is the main brain region involved in PD due to the death of localised dopaminergic neurons that cause a severe decrease in dopamine synthesis (Petrova *et al.*, 2004).

By incorporating both a sequence and structure-based approach to the *in-silico* analysis of the variant on the protein of interest, it may be possible to infer potential functional impacts that inform targeted lab-based functional studies. Thus, this chapter aimed to perform secondary and tertiary structural

analysis of the p.A13V variant and its effect on the MANF protein using a variety of ‘best practice’ open source *in silico* tools.

4.2 Methods and materials

In silico analysis allows us to determine the effect of the variant on the protein structure and how this may translate regarding protein function. The methodology was determined after analysing the literature on similar studies that highlighted the role of singular variants that may be implicated in disease through *in silico* analysis. A two-fold approach through the analysis of the secondary structure and thereafter, through the tertiary structure allows for a more comprehensive evaluation of the potential effect on the protein. The methodological approach used for this study is outlined in **Figure 4.1**.



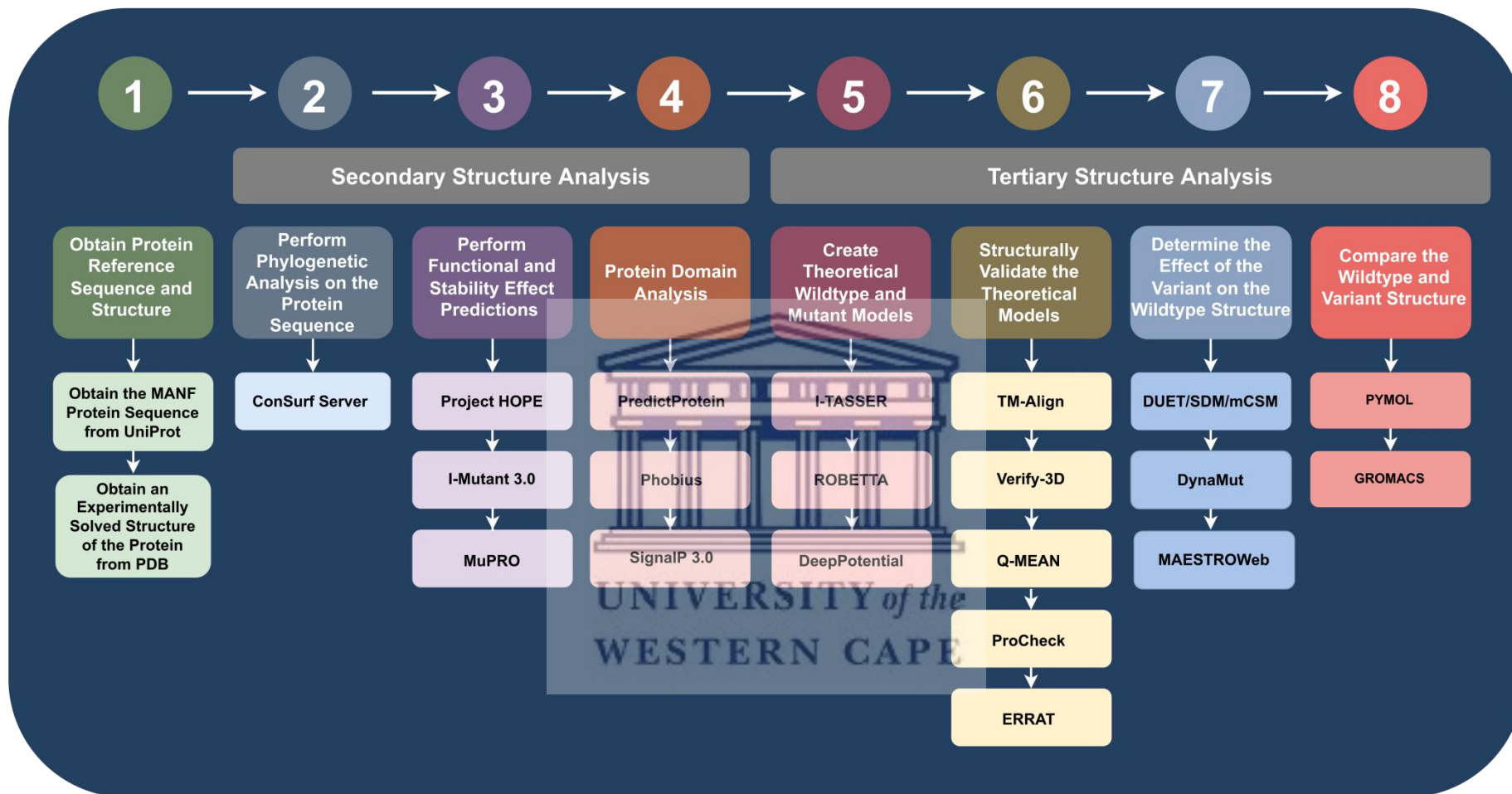


Figure 4.1: A brief overview of the methodology for *in silico* analysis of the p.A13V variant in MANF

4.2.1 Dataset identifiers

The protein sequence for human MANF was obtained from the UniProt (UniProt ID: P55145) (<https://www.uniprot.org/>), while additional information for the biological role and function of the protein was obtained from PUBMED (<https://pubmed.ncbi.nlm.nih.gov/>) (**Appendix I; A**). The experimentally solved crystallographic protein structure of MANF was obtained from the Protein Data Bank (PDB) (<https://www.rcsb.org/>). The structure with the highest scoring resolution (closest to 2Å which is the median resolution of most crystallographic structures) (PDB ID: 2W51) was chosen for comparative downstream analysis (**Appendix I; A**).

4.2.2 Secondary structure analysis

4.2.2.1 Phylogenetic analysis of the protein sequence

Phylogenetic evolutionary constraint analysis and the determination of solvent accessibility of each amino acid in the sequence were performed using the ConSurf server (<https://consurf.tau.ac.il/>). ConSurf uses a novel, cumulative method of phylogenetic analysis across homologous sequences while incorporating localised AA quality scores (Ashkenazy *et al.*, 2016). The parameters determined were based on the recommended guidelines for the analysis of a singular protein and included:

- Multiple Sequence Alignment is built using MAFFT;
- all homologues are collected from UNIREF90;
- the homolog search algorithm: HMMER;
- HMMER E-value: 0.0001;
- number of HMMER iterations: 3;
- maximal % ID between sequences: 95;
- minimal % ID For homologs: 35;
- 150 sequences that sample the list of homologues to the query;
- method of calculation: Bayesian and
- model of substitution for proteins: best fit.

4.2.2.2 Functional and stability effects prediction analysis

The Project HOPE server (<https://www3.cmbi.umcn.nl/hope>) was utilised to determine the potential functional effect of the variant on the protein sequence. Furthermore, the possible effect of the variant on protein stability was determined using MuPro (<https://mupro.proteomics.ics.uci.edu/>) and I-Mutant 3.0 (<http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi>) which both make use

of a support vector machine-based analysis to determine Gibb's free energy change ($\Delta\Delta G$) representing a change in stability.

4.2.2.3 Protein domain analysis

Once stability prediction and functional information had been derived, the protein sequence was used to determine the secondary structure of the protein and to elucidate whether the variant fell in a protein domain of interest. The PredictProtein server (<https://predictprotein.org/>) was incorporated to provide topological transmembrane information of the MANF protein based on the sequence input, as well as to determine the exact domain in which the variant of interest was present. Thereafter, Phobius (<https://www.ebi.ac.uk/Tools/pfa/phobius/>) and SignalP 3.0 (<https://services.healthtech.dtu.dk/service.php?SignalP>) were both utilised to confirm the presence of the signal peptide within the protein and to determine its domains and cleavage site.

4.2.3 Tertiary structure analysis

4.2.3.1 Ab initio structural modelling of the wildtype and variant proteins

Complete theoretical models of both the wildtype and mutant proteins were determined via an *ab initio* modelling method using the DeepPotential (<https://zhanggroup.org/DeepPotential/>), I-TASSER (<https://zhanggroup.org/I-TASSER/>), and Robetta (<https://robetta.bakerlab.org/>) servers. These servers were chosen for their ability to model a protein using only its protein sequence.

4.2.3.2 Structural validation of protein models

Initially, the resulting theoretical models were subjected to structural validation using the TM-Align (<https://zhanggroup.org/TM-align/>) server. The models were uploaded and analysed against the PDB experimental structure (2W51). The generated RMSD and Tm-Scores were then compared. The structures (both wildtype and mutant) chosen for further analysis possessed an RMSD closest to 2 Å (high-resolution structure) and a TM-Score closest to 1 (indicating the accuracy of the structural alignment when compared with the protein sequence).

The two prioritised theoretical models were then subjected to further structural analyses to confirm the quality of the structure using alternate quality scores, to corroborate the results obtained above. Verify3D (<https://www.doe-mbi.ucla.edu/verify3d/>) was used to determine the compatibility between the 3D theoretical structure and its protein sequence. Q-MEAN (<https://swissmodel.expasy.org/qmean/>) was used to assign an overall composite score that incorporated both local and global protein structure quality estimates. ProCheck (<http://www.csb.yale.edu/userguides/datamanip/procheck/manual/index.html>) was used to generate

Ramachandran plots to determine if the amino acids in the protein models fell in favoured regions, and ERRAT (<https://saves.mbi.ucla.edu/>) was used to determine the localised error rate values based on non-bonded atom-atom interactions within the entire structure.

4.2.3.3 Determining the effect of the variant on the wildtype structure

Once the wildtype structure was fully validated, *in silico* web tools that determined the effect of the variant on the theoretical model (as opposed to the protein sequence alone), were utilised. The ‘optimum’ wildtype protein was uploaded to the DUET (<http://biosig.unimelb.edu.au/duet/stability>), DynaMut (<https://biosig.lab.uq.edu.au/dynamut/>) and MAESTROWeb (<https://pbwww.services.came.sbg.ac.at/maestro/web>) servers to predict the functional effect of the missense mutation in MANF using complementary consensus prediction algorithms.

4.2.3.4 Comparing the theoretical wildtype and mutant structures

Initially, Pymol was used to identify polar contacts of the 13th residue relative to neighbouring residues in both the mutant and wildtype models. A fluctuation in the number of polar contacts can be indicative of an increase or decrease in stability based on the AA’s interaction with water.

MD simulations for the wildtype and variant models was run through GROMACS (Galaxy Version 2022+galaxy) on the Galaxy web server (<https://usegalaxy.eu/>). The initial setup required an input PDB file to generate a topology, a GRO and a position restraint file for downstream molecular dynamics analysis. The topology file consists of important descriptors of the protein including charges, bond lengths and angles, and the masses of atoms. The topology file is produced after the selection of a force field and water mode. The GRO file contains all the information about the protein’s structural co-ordinates. For this analysis, the chosen force field was an OPLS/AA force field with a TIP3 water model. Next, the structural configuration was incorporated to outline the parameters for the simulation box. The box dimensions were set at 1.0 nm, as a rectangular box with all sides equal. Thereafter, the protein is solvated where water molecules are added to the structure and topology files to fill the unit cell (Bellissent-Funel *et al.*, 2016). At this step, sodium or chloride ions are also added to neutralise the charge of the system. To remove any steric clashes or unusual geometry which would artificially raise the energy of the system, we must relax the structure by running an energy minimisation (EM) step. The following parameters were specified for the EM of the protein: Choice of integrator”: Steepest descent algorithm (a most common choice for EM)

“Neighbor searching”: Generate a pair list with buffering (the ‘Verlet scheme’)

“Electrostatics”: Fast smooth Particle-Mesh Ewald (SPME) electrostatics

“Distance for the Coulomb cut-off”: 1.0

“Cut-off distance for the short-range neighbour list”: 1.0 (but irrelevant as we are using the Verlet scheme)

“Short range van der Waals cutoff”: 1.0

“Number of steps for the MD simulation”: 50000

“EM tolerance”: 1000

“Maximum step size”: 0.01

At this point of the analysis, equilibration of the solvent around the protein is necessary. This is performed in two stages: equilibration under an NVT ensemble, followed by an NPT ensemble. Use of the NVT ensemble entails maintaining a constant number of particles, volume and temperature, while the NPT ensemble maintains a constant number of particles, pressure and temperature. (The NVT ensemble is also known as the isothermal-isochoric ensemble, while the NPT ensemble is also known as the isothermal-isobaric ensemble) (Bellissent-Funel *et al.*, 2016). During the first equilibration step (NVT), the protein must be held in place while the solvent is allowed to move freely around it. This is achieved using the position restraint file that was created in the system setup. During the second NPT step, the positional restraints are removed. The following parameters were followed for both the NVT and NPT steps.

“Choice of integrator”: A leap-frog algorithm for integrating Newton’s equations of motion (A basic leap-frog integrator)

“Bond constraints”: Bonds with H-atoms (bonds involving H are constrained)

“Neighbor searching”: Generate a pair list with buffering (the ‘Verlet scheme’)

“Electrostatics”: Fast smooth Particle-Mesh Ewald (SPME) electrostatics

“Temperature”: 300

“Step length in ps”: 0.002

“Number of steps that elapse between saving data points (velocities, forces, energies)”: 5000

“Distance for the Coulomb cut-off”: 1.0

“Cut-off distance for the short-range neighbour list”: 1.0

“Short range van der Waals cutoff”: 1.0

“Number of steps for the NVT simulation”: 50000

Finally, the product simulation is run using the checkpoint file obtained after the equilibration steps to obtain a trajectory and final GRO file using the following parameters:

“Choice of integrator”: A leap-frog algorithm for integrating Newton’s equations of motion (A basic leap-frog integrator)

“Bond constraints”: Bonds with H-atoms (bonds involving H are constrained)

“Neighbor searching”: Generate a pair list with buffering (the ‘Verlet scheme’)

“Electrostatics”: Fast smooth Particle-Mesh Ewald (SPME) electrostatics

“Temperature”: 300

“Step length in ps”: 0.002

“Number of steps that elapse between saving data points (velocities, forces, energies)”: 5000

“Distance for the Coulomb cut-off”: 1.0

“Cut-off distance for the short-range neighbour list”: 1.0 (but irrelevant as we are using the Verlet scheme)

“Short range van der Waals cutoff”: 1.0

“Number of steps for the simulation”: 500000

Thereafter, Bio3D on the Galaxy server was used to produce the RMSD, RMSF and PCA analysis plots using the trajectories and GRO files obtained after the product simulation.

4.3 Results

4.3.1 Secondary structure analysis

4.3.1.1 Phylogenetic conservation analysis reveals that p.A13V is a buried residue with variable conservation

The ConSurf server provides an estimate of evolutionary conservation and solvent accessibility across the amino acids in a protein sequence, which helps ascertain whether a variant at a particular site may be considered more ‘harmful’. The p.A13V variant returned a normalised score of 1.014 indicating variable conservation (i.e. the AA site may be in a conserved region with some accepted AA substitutions), however, the multiple sequence alignment of the MANF protein with homologues indicated that valine is not a common substitution at position 13 (**Appendix I; B**). The output also showed that it is in a buried residue (which occurs on the interior of the protein and play a significant role in its structural stability) (Aftabuddin and Kundu, 2007). Previous studies have shown that a variant with low (or variable) conservation can have a considerable effect on protein function (Coulthurst *et al.*, 2012).

4.3.1.2 Functional and stability effects prediction analysis indicated that the variant is situated in the signal peptide and is destabilising

Project Hope is a webserver that identifies the structural effects of the point mutations that are provided in the protein sequence of interest. It is also able to provide a basic analysis of whether the AA change causes an alteration to the protein structure. The output outlined that the mutant residue (valine) is larger than the wildtype residue (alanine) due to the addition of a single side chain which

may introduce ‘bumps’ or spatial adjustments at the site of the variant in the protein structure. The analysis also noted the variant to be located within the protein’s signal peptide. This is important as signal peptides are often recognised by other proteins and cleaved off to generate the mature protein. Notably, the new residue (valine) that is introduced in the signal peptide differs in its properties from the original one (alanine) thus the variant may disturb the recognition of the signal peptide. Although the alanine-to-valine change is not considered a large change, it has been noted to cause aberrations in protein structure and ultimately its function (Bough and Dayan, 2022).

As the stability of a protein governs its structural conformation, any change in stability can have a significant impact on its subsequent folding, as well as potential degradation *in vivo*. MuPro and I-Mutant 3.0 stability prediction analyses use both SVM and NN algorithms. In MuPro, a confidence score between -1 and 1 is typically computed alongside the $\Delta\Delta G$ score. A score < 0 means the variant decreases the protein stability whereas a score > 0 means the variant increases the protein stability. I-Mutant 3.0 produces a similar $\Delta\Delta G$ score to MuPro, as well as a Reliability Index (RI) score where 0 and 10 represent the lowest and highest reliability, respectively. MuPro depicted a decrease in stability (-0.2 kcal/mol) using its recommended analysis (SVM) and also with the use of a NN (-0.7 kcal/mol), while proposing an increase in stability (0.16 kcal/mol) when calculated using just an SVM and a shorter sequence window as seen in **Table 4.1**. Furthermore, the I-Mutant 3.0 server also depicted a decrease in stability (-0.21 kcal/mol) after the introduction of the variant, with an RI of 4 indicating moderate reliability. Thus, the indication of decreased stability of the variant using these tools warrants further analysis of the complete protein structure of MANF.

Table 4.1: Predicted effect of the p.A13V variant on the MANF protein using MuPro and I-Mutant 3.0

Gene	MuPro	I-Mutant 3.0
MANF	1. Predicted both value and sign of energy change using SVM and sequence information only (Recommended)	SVM2 Prediction Effect:
	$\Delta\Delta G = -0.20007335$ kcal/mol (DECREASE stability)	Decrease RI: 4
	2. Prediction of the sign (direction) of energy change using SVM and neural network with a smaller sequence window	$\Delta\Delta G$ Value Prediction:
	<i>Method 1: Support Vector Machine, uses sequence information only.</i> Effect: INCREASE the stability of protein structure. Confidence Score: 0.16348451 kcal/mol	Decrease -0.21 kcal/mol
	<i>Method 2: Neural Network, uses sequence information only.</i> Effect: DECREASE the stability of protein structure. Confidence Score: -0.74377979768028 kcal/mol	

4.3.1.3 Protein domain analysis confirmed the presence of the p.A13V variant in the hydrophobic core of the signal peptide

The protein sequence was run through PredictProtein to obtain a detailed analysis of the secondary structure of the MANF protein. The output confirmed the presence of a signal peptide spanning between residues 1 and 24 (**Figure 4.2**). Furthermore, the program depicted that the variant fell within a helix in the signal peptide and was considered a buried residue - further confirming the information provided by the Project HOPE server. This is important as mutations in the signal peptide, particularly the hydrophobic core, have been found to reduce the translocation of a protein thereby reducing its expression (Rajpar *et al.*, 2002; Pidasheva *et al.*, 2005).



Figure 4.2: Secondary structure prediction of the MANF protein using the PredictProtein server

The red bar indicates the region of the protein sequence in which the p.A13V variant occurs in MANF.

It is important to note that this portion of the structure is not included in the experimentally solved structures present in Protein Data Bank (PDB), as signal peptides are typically cleaved off before protein secretion, and are therefore not part of the mature/ final protein. These short sequences (18-30 AA) typically form a structure that constitutes an N-terminal and a C-terminal, a hydrophobic helical core and a cleavage site as seen in **Figure 4.3**.

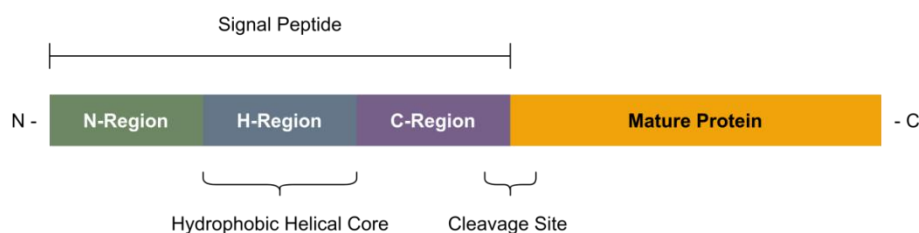


Figure 4.3: Simplified structure of a signal peptide

Figure created using Draw.io.

In subsequent analysis to confirm the presence of the signal peptide, determine its domains and confirm that the variant fell within the hydrophobic core, the protein sequence was run on the Phobius server. The server output a 2-Dimensional visualisation of the peptide structure indicating the location of each of the signal regions as seen in **Figure 4.4**. The H-region or hydrophobic core of the signal peptide spans across AAs 9-19 of the protein sequence confirming that the p.A13V variant is present in a ‘critical’ region. Mutations in this core helix have been previously linked to protein dysfunction resulting in disease (Kamp and Daggett, 2010). Furthermore, the peptide cleavage site is allocated between AA positions 24 to 25.

Phobius prediction

Prediction of sp|P55145|MANF_HUMAN

```

ID   sp|P55145|MANF_HUMAN
FT   SIGNAL          1    24
FT   REGION         1    8      N-REGION.
FT   REGION         9   19      H-REGION.
FT   REGION        20   24      C-REGION.
FT   TOPO_DOM       25   182     NON CYTOPLASMIC.
//

```

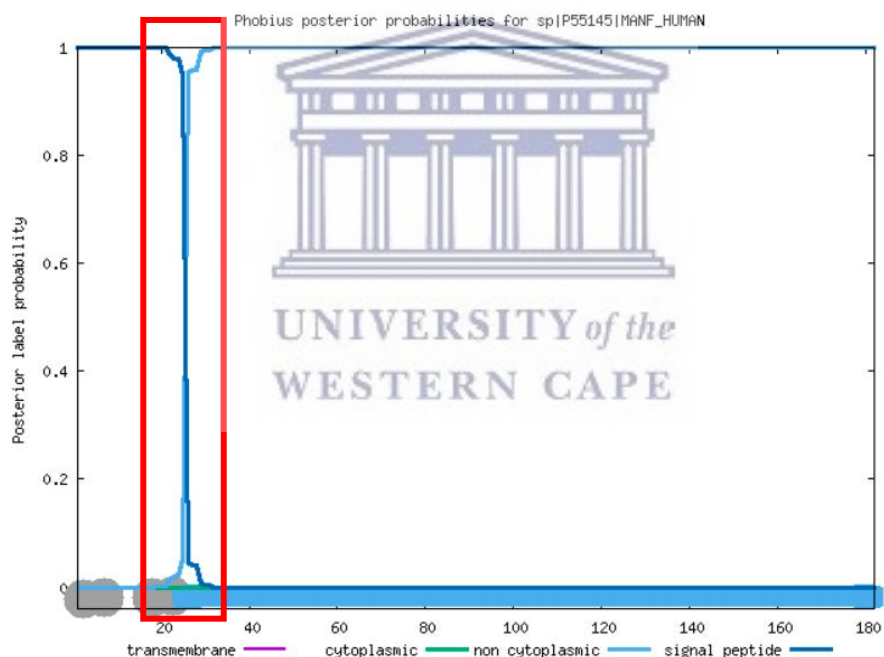


Figure 4.4: Signal peptide secondary structure prediction for the MANF protein using the Phobius server

The red bar indicates the cleavage region of the signal peptide. The grey regions are indicative of the N- and C-regions of the signal peptide.

The SignalP-Neural Network (NN) is composed of two independent NNs, one based on the confidence that the protein sequence is a signal peptide sequence (the S score), and the other NN is based on the confidence that a given position in the sequence is the cleavage site (the C score). When

the SignalP-NN is run, it integrates the output of both NN to generate a D score between 0 and 1. D scores greater than a threshold yields a prediction that the sequence is a recognisable signal peptide. The uploading of both the wildtype (D-score = 0.903) and variant-containing (D-score = 0.867) protein sequences yielded slightly different peptide predictions as seen in **Figure 4.5**. A study analysing the difference in D-scores (produced by SignalP 3.0) among variants found in the signal peptide and linked to various diseases, noted that the range of change in D-score among the wildtype and mutant alleles ranged from 1.6 to 28.6%, with 70% of the variants residing in the hydrophobic core (Jarjanazi *et al.*, 2008). Our wildtype and variant D-score difference was calculated as 3.99%. Variants with smaller changes (<5%) in the study were found to display impaired co-translational processing and protein secretion deficiency (Fingerhut *et al.*, 2004) and incomplete protein glycosylation (Anjos *et al.*, 2002). Although this is considered a smaller change in terms of probability prediction of the signal peptide, it does not take into account the allosteric effects (alteration of the protein conformation resulting in a change of function) of the variant on different regions of the protein and may still have a significant effect and thus, warrants further analysis.

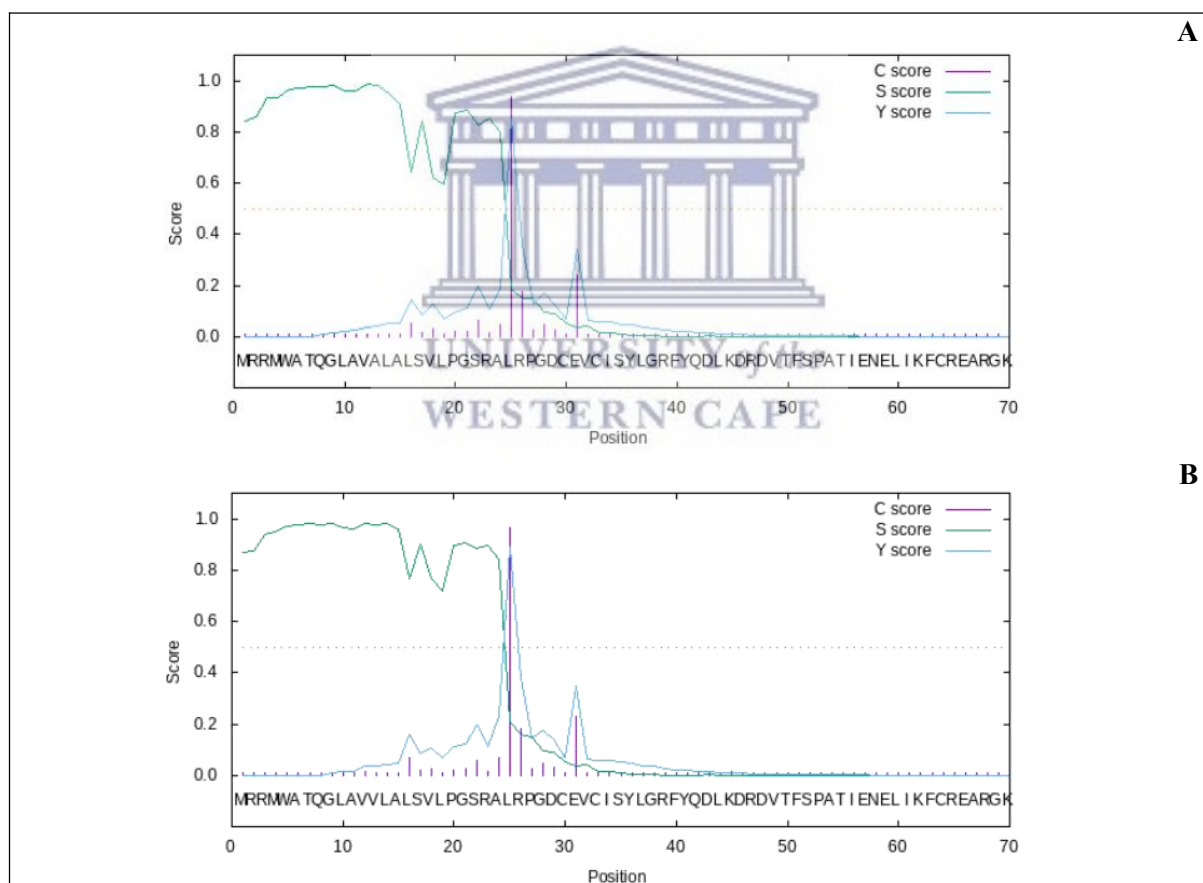


Figure 4.5: Signal peptide secondary structure prediction for the MANF protein using the SignalP 3.0 server

[A]- wildtype protein and [B]- mutant protein

C-Score (purple lines): confidence score that a given position in the sequence is the cleavage site; S-Score (green lines): confidence that the protein sequence is a signal peptide sequence.

4.3.2 Tertiary Structure Analysis

Part 1 (Ab Initio Protein Modelling and Validation of the MANF Protein)

4.3.2.1 Ab initio structural modelling of the wildtype and variant protein successfully introduced the signal peptide domain onto the protein structures

Protein structures on PDB tend to lack the signal peptide since this portion of the protein is typically cleaved before the mature protein has fully translocated. This was also the case for the MANF protein, thus, for the present study, the helical structure of the signal peptide needed to be included in the 3-D model for tertiary structure analysis. The programs I-TASSER, Robetta and DeepPotential were chosen for their ability to model the ‘full length’ protein (incorporating the signal peptide) using the protein sequence only (*ab initio* modelling) as opposed to comparative modelling since the signal peptide was not present on any of the solved structures. Each of the programs was able to output probable models that included the signal peptide structure. These were then evaluated to determine the best models representing the complete wildtype (Alanine at position 13) and the complete mutant model (Valine at position 13).

4.3.2.2 Structural validation of protein models indicated that Robetta produced the highest quality theoretical structures

When compared to an experimentally solved structure, accurate theoretical models present with RMSD scores $< 2.0 \text{ \AA}$ and TM-scores that are approaching 1. The RMSD score allows one to observe the resolution of the model, while the TM-score produces a structural alignment score that is normalised against larger differences between the theoretical and solved protein structures (Zhang *et al.*, 2022). To obtain the best theoretical wildtype and variant structure of MANF, each model generated was aligned against the crystallographic solved structure in PDB (2W51; PDB) using TM-Align to obtain individual RMSD and TM-scores.

As seen in **Table 4.2**, where the structural alignment scores for all the models are presented, none of the predicted models obtained an RMSD of $< 2.0 \text{ \AA}$. A potential reason for the higher scores may be due to the non-alignment of the entire signal peptide that is omitted from the reference structure, however, TM-Align does provide scores that are normalised against the reference protein to take this into account. TM-Align is typically utilised to evaluate the overall quality of the protein structure by comparing not only the sequence alignment but also the folding of the protein when compared to the reference (Zhang *et al.*, 2022). Overall, both the Robetta wildtype and variant models produced the highest TM-scores (> 0.74653) across the 3 software programs used to generate the models. The

TM-score tends to be a more sensitive scoring method as opposed to the RMSD score alone, as it is more heavily weighted to smaller distance errors in the alignment and accounts for global fold similarity rather than just local structural variations (Zhang *et al.*, 2022). This is particularly helpful as most theoretical models are created to analyse a region that is not accounted for in experimentally solved structures. Thus, due to the consistently higher TM-scores across all the Robetta-generated models, a wildtype and variant model was chosen from Robetta, with structural scores of RMSD = 2.88Å / TM-Score = 0.75190 and RMSD = 2.90 / TM-Score = 0.75231, respectively (highlighted in **Table 4.2**). The optimum models generated by Robetta were then subjected to further structural validation to ensure appropriate structure quality for further variant analysis (**Appendix I; C**).

Table 4.2: Structural alignment and validation of the generated models of wildtype and mutant MANF protein

Modelling Algorithm	Model	Wildtype Models		Mutant Models	
		RMSD (Å)	TM-Score	RMSD (Å)	TM-Score
Robetta	1	2.94	0.75472	2.90	0.75231
	2	2.88	0.75190	2.95	0.74867
	3	2.93	0.74699	2.93	0.74772
	4	2.94	0.76302	3.23	0.75233
	5	2.92	0.75060	2.95	0.74653
Deep Potential	1	2.80	0.68056	3.02	0.68991
I-TASSER	1	2.80	0.63104	2.81	0.64786
	2	2.35	0.67449	2.85	0.63083
	3	2.91	0.54822	2.45	0.66491
	4	2.67	0.63065	3.00	0.54804
	5	2.82	0.70670	3.74	0.55001

Orange cells indicate the best quality scores for the wildtype and mutant theoretical structures.

Verify-3D calculates the quality of each of the AA acids in the structure concerning their position in the protein sequence to depict the overall quality of the model. In the wildtype model, 90.11% of the residues had averaged a 3D-1D score ≥ 2 and had scored a pass score, while the mutant model depicted that 86.26% of the residues had averaged a 3D-1D score ≥ 2 and had scored a pass score (**Figure 4.6: A and B**, respectively). A score above 80% represents a good quality model due to the accurate correlation between the predicted secondary structure and the theoretical 3D structure. The Q-MEAN server allocates a Z-score that is indicative of how well a model is structured by comparing it to the hundreds of experimentally solved protein structures on PDB. A Z-score less than 1 indicates the theoretical structure is of good quality and comparable to experimental structures. The wildtype model produced a Z-score of - 0.6 while the mutant model was scored at 0.10. Thus, both models fell in the region of high-quality protein structures as seen by the presence of the model (**red star**) in the dark grey region of panels **A and B** in **Figure 4.7**. Furthermore, Q-MEAN also analyses the structure according to the similar protein structures found on PDB (**Figure 4.8: A and B**) hence the decrease in

quality of the peptide region for both models as these parts of the structure are not present in the template structures.

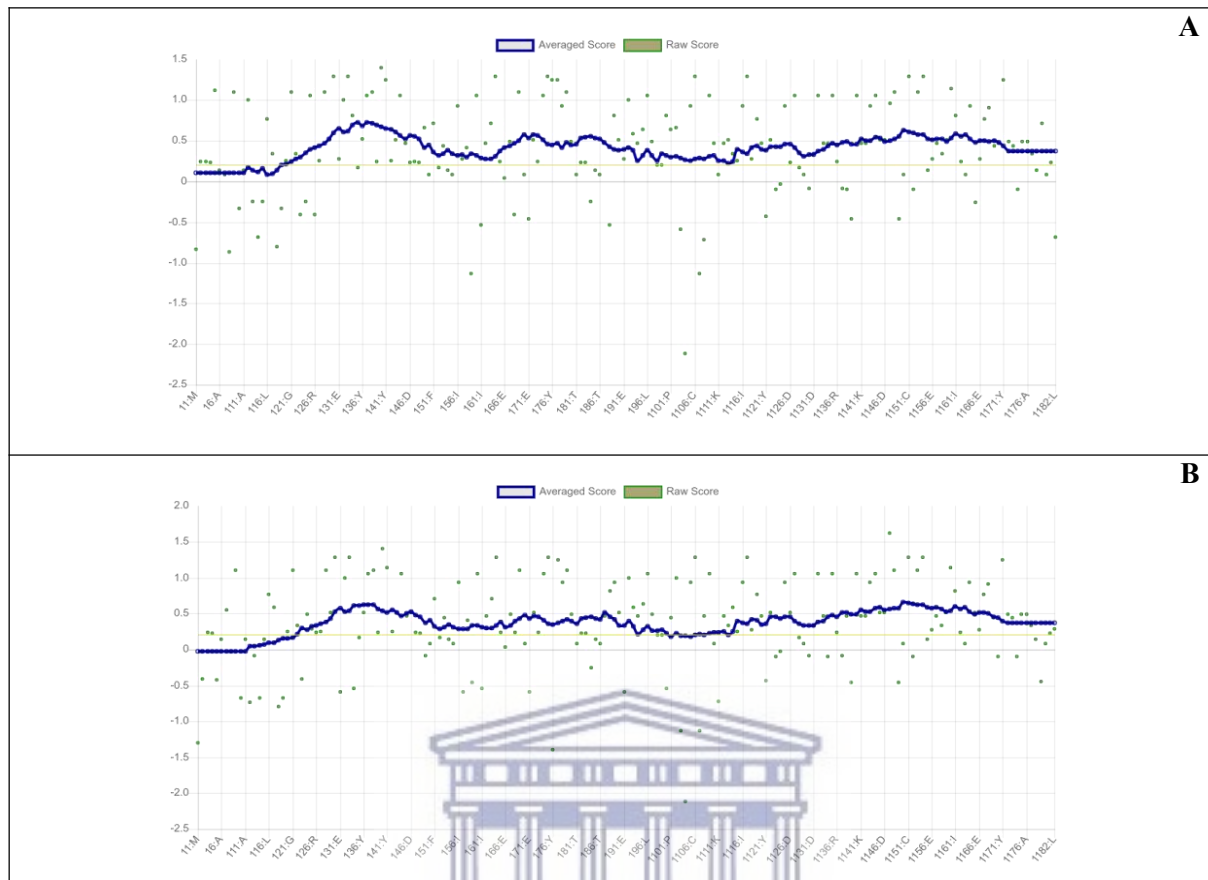


Figure 4.6: AA quality plots generated for the theoretical protein structures by Verify3D

[A]- Wildtype protein and [B]- Mutant protein; X axis shows the positions of the protein residues while the Y axis shows the score value; the blue line indicates the averaged score for each AA of the protein; the green scatter shows raw scores for each AA in the databank.

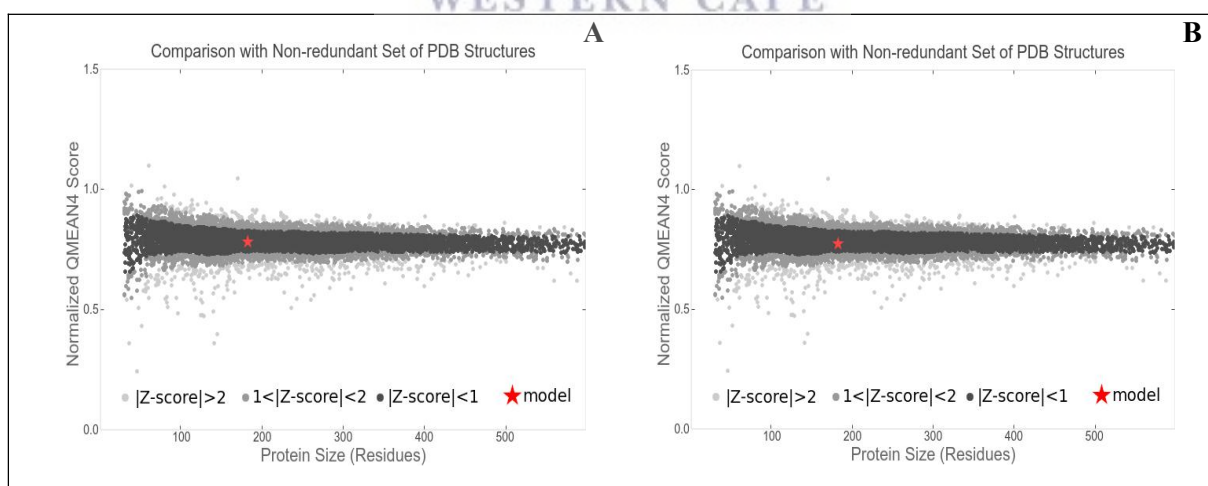


Figure 4.7: Q-MEAN PDB structure comparison plots generated for the theoretical protein structures

[A]- Wildtype protein and [B]- Mutant protein; X-axis: Protein size (residues) of all the proteins in the PDB; Y-axis: Normalised Q-MEAN score for each of the proteins; the red star indicates the quality of your protein in comparison to the experimentally solved protein structures.

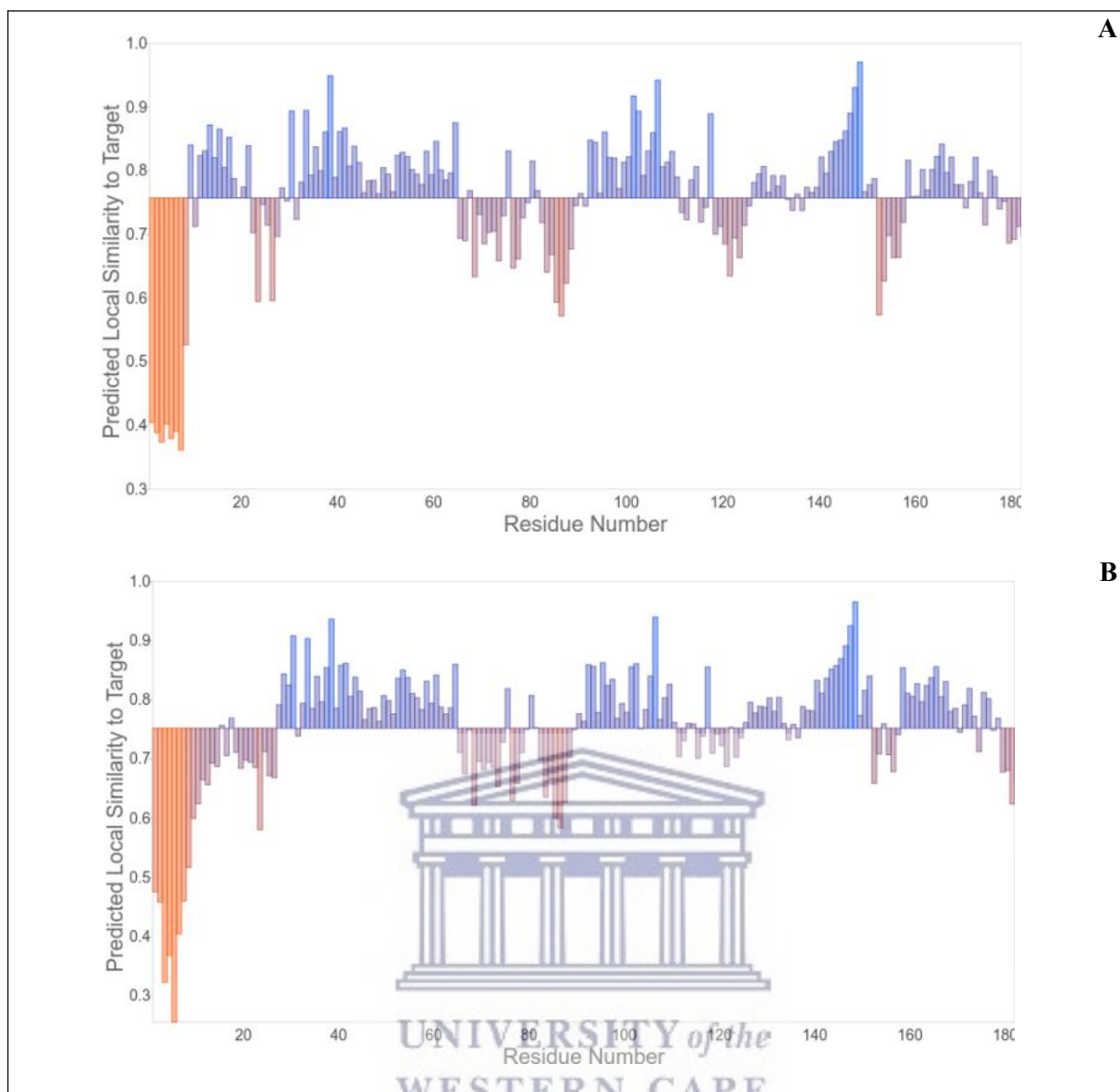


Figure 4.8: Q-MEAN local similarity plots generated for the theoretical protein structures

[A]- Wildtype protein and [B]- Mutant protein; X-axis: Residue number; Y-axis: Predicted similarity to target score. Orange bars indicate a score less than 0 (lower quality) while the blue bars indicate good quality between local residues.

ProCheck was then used to create Ramachandran plots for each of the models (**Figure 4.9: A & B**). These plots are used to determine whether or not the AAs in the theoretical model fall in favoured regions, thus indicating a high-quality model. The wildtype model was scored having 94.6% of its AAs in favoured regions with 5.3% of AAs falling in additionally allowed regions and 0.1% fall in ‘generously allowed’ regions. The mutant model depicted 90.3% of its AAs in favoured regions with 7.3% of AAs falling in additionally allowed regions and 2.4% fall in ‘generously allowed’ regions. Thus, both models were considered high-quality models.

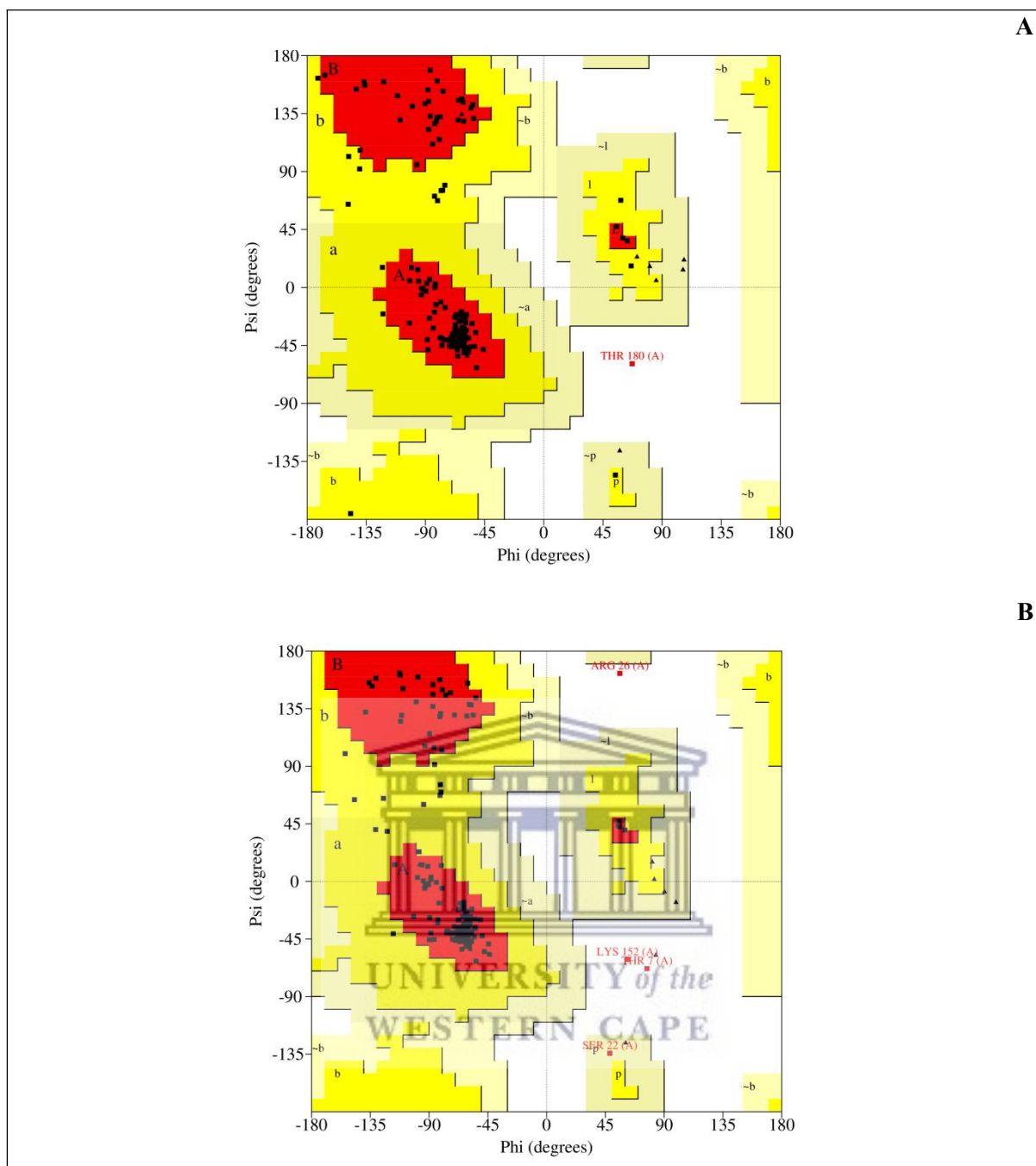


Figure 4.9: Ramachandran plots generated using ProCheck

[A]- Wildtype protein and [B]- Mutant protein.

The Overall Quality Factor provided by ERRAT on the wildtype structure was 100% and on the mutant structure 96.5% which can be seen in **Figure 4.10**. The 100 represents the percentage of protein with a calculated error value that falls below the 95% rejection limit. According to the ERRAT server, structures with good resolution typically produce values around 95% or higher. Thus, this indicates that both structures were predicted to be of good quality based on localised AA evaluation. It is to be noted that a proportion of the error values in the mutant fall in the region of AA

84, a helical region of the protein. Another deviation in the error values occur at the end of the sequence in the mutant, a region containing a conserved signalling sequence, ‘RTDL’.

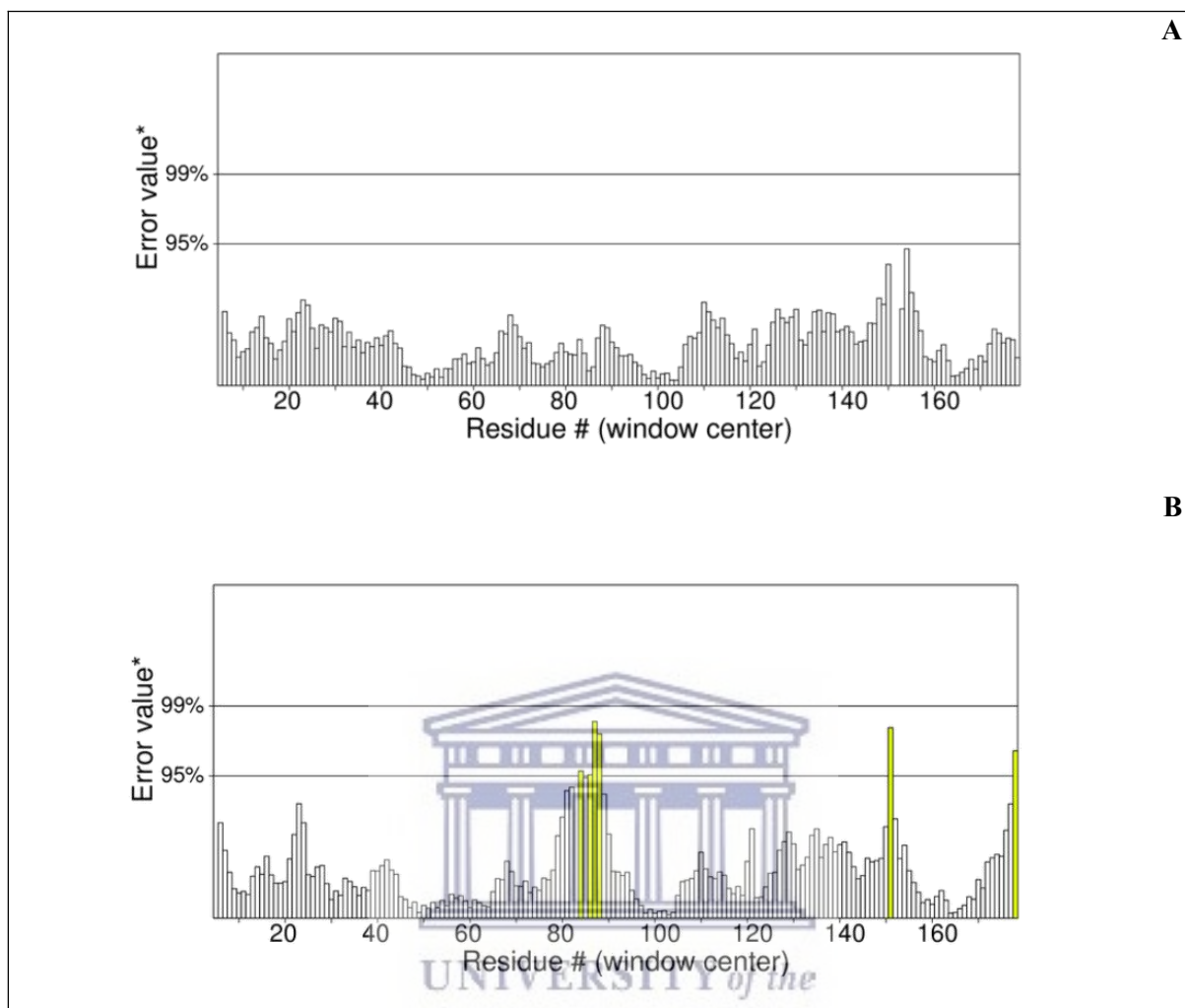


Figure 4.10: Error values calculated for each AA using the ERRAT server

[A]- Wildtype protein and [B]- Mutant protein; Yellow highlighted areas show a significant deviation in error values.

Part 2 (Mutation Analysis of the Theoretical Protein Structures)

4.3.2.3 Determining the effect of the mutation on the wildtype structure indicated that the variant is destabilising

DUET predicts the change in protein stability by ($\Delta\Delta G$) upon the introduction of a variant using a 3-Dimensional model of a protein, as opposed to solely the protein sequence. $\Delta\Delta G$ is a measure of the change in energy between the folded and unfolded protein states and the change in folding when a point mutation is introduced (Park *et al.*, 2016). It does this by combining two separate approaches, namely SDM and mCSM which represents a statistical potential energy function and a predictive model that makes use of graph-based signatures to represent the protein structure (Pires *et al.*, 2016).

These methods in tandem provide a more comprehensive analysis of the effect of the variant on the protein. The individual predictions, as well as the combined DUET prediction, are portrayed in **Table 4.3**. All three algorithms indicated that the variant was destabilising the protein with an aggregate DUET $\Delta\Delta G$ score of -0.108 kcal/mol.

Table 4.3: $\Delta\Delta G$ scores for the p.A13V variant on the wildtype protein structure using DUET/SDM/mCSM

Stability Effect Algorithm	$\Delta\Delta G$	Predicted Effect on Protein
mCSM Predicted Stability Change	-0.263 kcal/mol	Destabilising
SDM Predicted Stability Change	-1.03 kcal/mol	Destabilising
DUET Predicted Stability Change	-0.108 kcal/mol	Destabilising

Dynamut incorporates both graph-based signatures as well as a normal mode dynamic analysis to produce a consensus prediction of the impact of a variant on a wildtype protein structure (Rodrigues *et al.*, 2018). Again, a decrease in the $\Delta\Delta G$ (-0.309 kcal/mol) was detected resulting in a potentially destabilising effect on the protein. As seen in **Figure 4.11**, the vibrational entropy was calculated to determine if the variant affects the flexibility of the structure. There was a decrease in the change in vibrational entropy (-0.076 kcal/mol) and the blue colouring on the helical structure represents a rigidification upon the introduction of the mutation. Here, the signal peptide, towards the cleavage site, is seen to change flexibility.

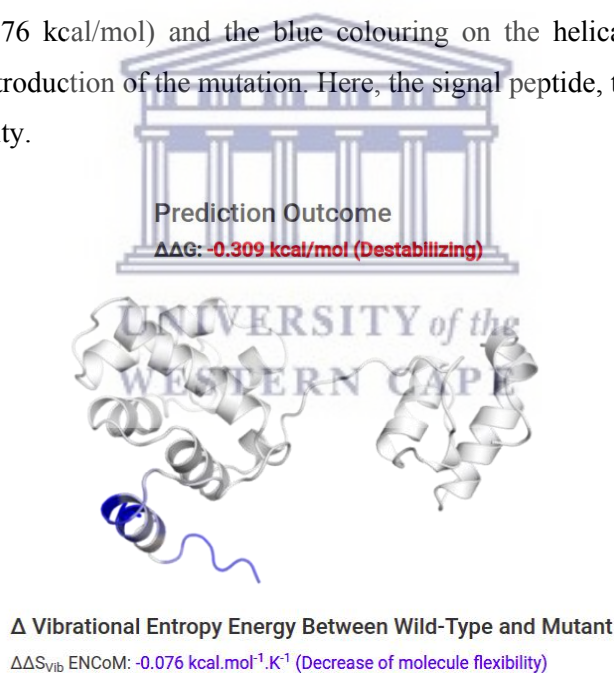


Figure 4.11: Output depicting the $\Delta\Delta G$ and molecular flexibility of the p.A13V variant on the wildtype structure using Dynamut

The figure depicts the outcome of superimposing the variant onto the wildtype structure. The dark blue portion of the protein represents a the structural change to the C-terminal of the signal peptide.

The MAESTROWeb server also incorporates the use of the wildtype model and creates mutation sensitivity profiles and evaluates potential disulfide bonds to determine the effect of a variant on the 3D wildtype structure. However, the output from this server indicated a $\Delta\Delta G$ of -0.183 with a

confidence score of 0.904 (1 being the most accurate). This indicates a predicted destabilising effect of the variant on the protein.

4.3.2.4 Comparing the wildtype and mutant structures using Pymol and MD simulations indicated a difference in polarity, potential flexibility and conformation

Pymol was used to determine the difference in intramolecular polar contacts (H-bonds and/or salt bridges) between the wildtype and mutant structure. The wildtype structure was shown to have 4 polar contacts to neighbouring residues, as compared the mutant which only had 2 (**Appendix I; D**). This result coincides with earlier analysis findings indicating that valine would decrease polarity of the molecule due to its extra alkyl group, which could cause spatial disturbances in the hydrophobic core of the protein. The loss in polarity seen among amino acid substitutions has previously been associated with a decrease in stability due to a difference in solvation, affecting the fold of the domain, and potentially the overall function of the protein (Worth and Blundell, 2010)

MD simulations are typically performed to determine the flexibility of a protein, though can be computationally extensive. The RMSD time series is indicative of the stability of the protein conformation through the simulation frames/time. The RMSD is typically a calculation of the carbon atoms of the protein backbone and their change in conformation throughout the simulation. Ideally, a protein indicating thermal stability will depict fewer deviations throughout the simulation. The wildtype protein indicated there were fluctuations in the RMSD graph, though the values were found to have plateaued at around 2.9 Å as confirmed by the RMSD histogram plot (**Figure 4.10 A**). This is expected for a stable protein and indicates no large conformational changes occurring during the simulation (Bray *et al.*, 2020). The wildtype protein however indicates a steady deviation away from the protein's original conformation. The presence of the 2 distinct peaks seen in the RMSD histogram plot represents conformational changes that were observed during the trajectory (Batut *et al.*, 2018).

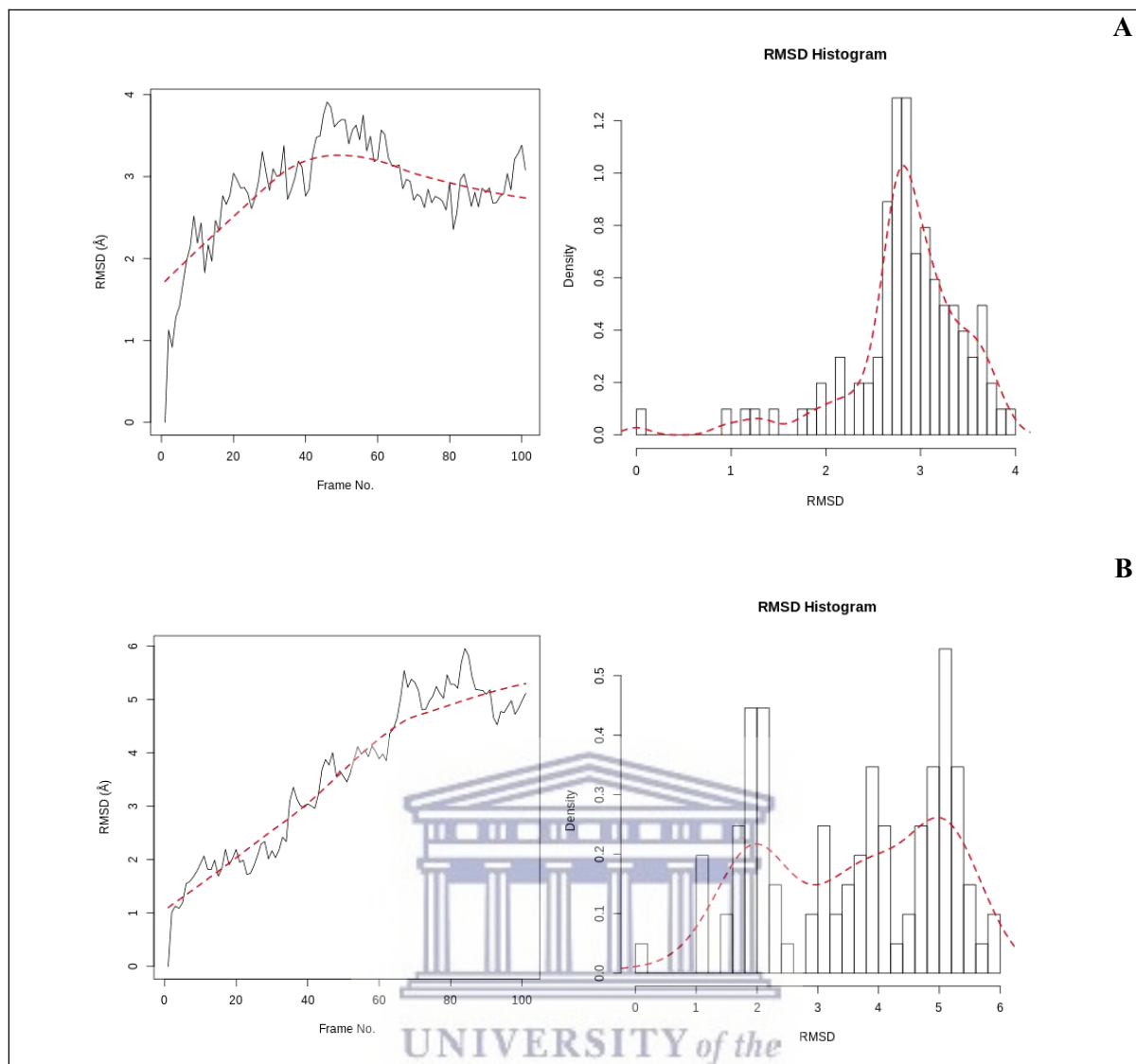


Figure 4.12: Comparing the RMSD between the wildtype and variant structure
 [A]- RMSD plot and histogram for the wildtype protein; [B]- RMSD plot and histogram for the variant protein.

The RMSF produced from an MD simulation measures the average deviation of a particle (e.g. a protein residue) over time from a reference position (typically the time-averaged position of the particle). Thus, RMSF analyses the portions of the structure that are fluctuating from their mean structure the most (or least) (Barazorda-Ccahuana *et al.*, 2018). The significantly higher difference in RMSF values ($>1\text{\AA}$) was seen for the mutant structure at sites 12 to 24, a region predicted to be the cleavage site in this protein (**Figure 4.13**). Also, the increase in RMSF values at AA sites ~ 130 and ~ 150 (regions that contain a helix) is seen in the mutant. There is also a large RMSF fluctuation seen at the C-terminal of the protein. An increase in RMSF value corresponds to the higher flexibility of a residue while a decrease can indicate increased rigidity to the structure). Potentially, a perturbation that interferes with the flexibility of a protein can interfere with the function of the protein (Guo *et al.*, 2022; Teilum *et al.*, 2011). As the cleavage site and helical regions depict the most significant

deviation in terms of RMSF values, it could be hypothesised that the variant's impact on the protein could potentially prevent adequate cleavage of the signal peptide thereby preventing translocation to the ER. Also, the interference noted in the helical structure that is assumed to induce flexibility within the structure could prevent proper folding of the protein, thus preventing normal trafficking.

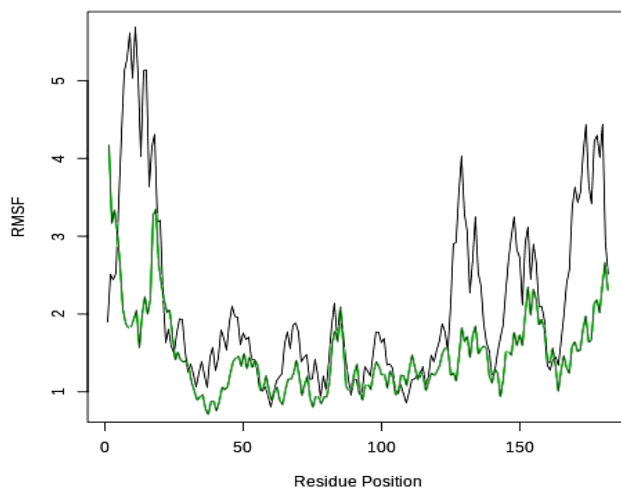


Figure 4.13: Comparison of the root mean square fluctuation between the wildtype and variant structure

RMSF values (Å) of the wildtype protein (green) and RMSF values (Å) of the variant protein (black).

Principal component analysis is typically used to derive evident variation from large datasets, such as the trajectories produced during the MD simulation. The analysis converts the observed correlations (i.e. movement of the carbon atoms in the protein backbone) to an uncorrelated set of principal components (PCs). This is done to determine the interactions between important conformations in the protein. In the wildtype protein, the first three PCs are found to be responsible for 69.3% of the variance, as seen in the eigenvalue plot (**Figure 4.14**). In the variant protein, the first three PCs are found to be responsible for 85.8% of the variance, as seen in the eigenvalue plot. The red and black dots each represent a conformational cluster state. The distributions of these conformational clusters are seen to differ entirely between the wildtype and the mutant protein. where the position of the clusters is reversed after the introduction of the variant, indicating a change in allosteric conformation. Allosteric conformations within the hydrophobic core can have downstream structural impacts on the protein structure and subsequent fold.

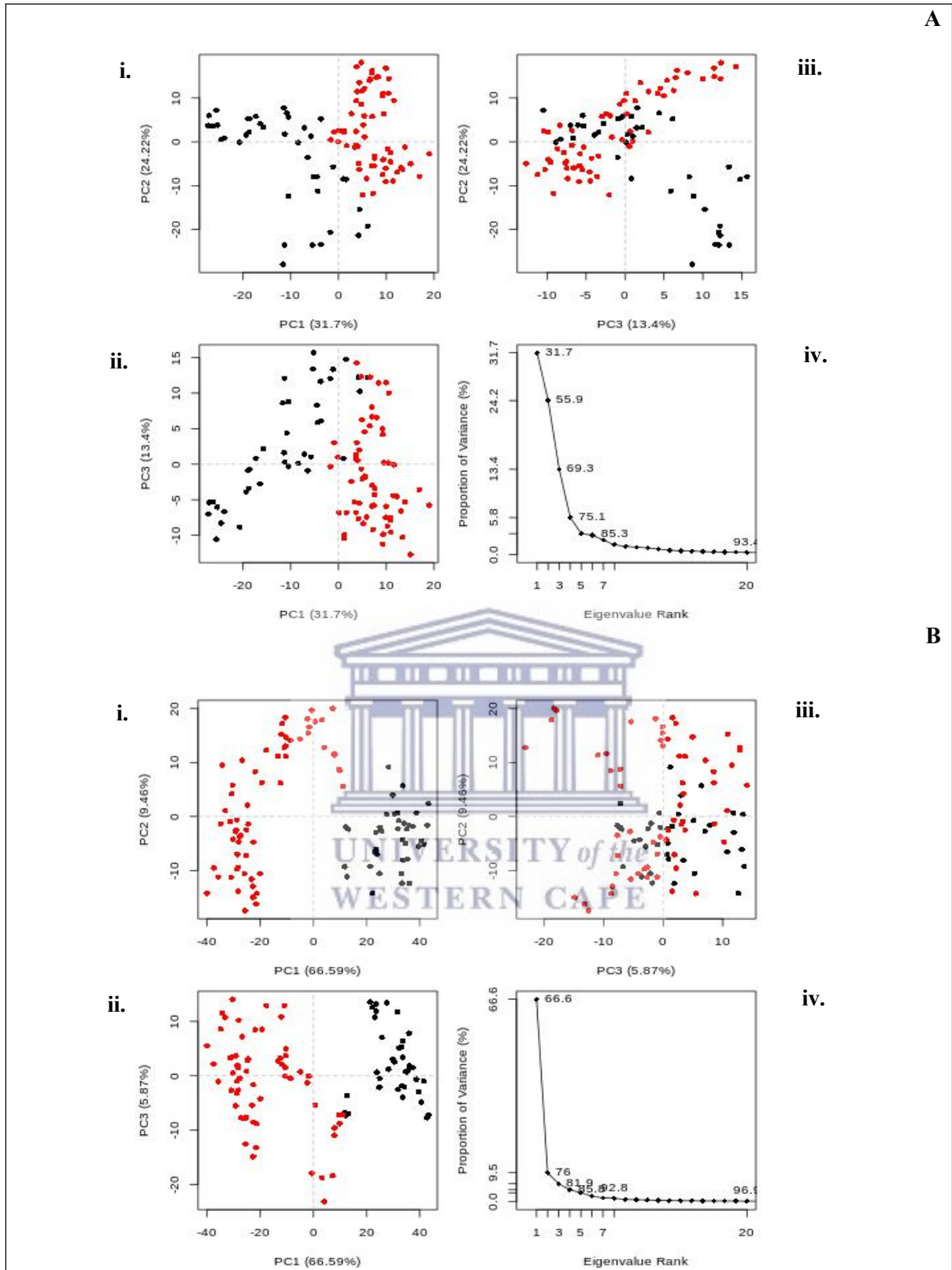


Figure 4.14: PCA of the wildtype and variant models of MANF

[A]- Wildtype protein: **i.** Principal components (PC) 1 and 2. **ii.** PC1 and PC3. **iii.** PC2 and PC3. **iv.** Percentage of variance explained by the first 20 PC. [B]- Variant protein: **i.** Principal components (PC) 1 and 2. **ii.** PC1 and PC3. **iii.** PC2 and PC3. **iv.** Percentage of variance explained by the first 20 PC.

In summary, the combined approach of using multiple tools to determine the impact of the variant on the protein supports the hypothesis that the protein structure is altered by the variant which may impact downstream function. Using both a sequence-based and complete theoretical structural approach, it is evident that the variant may cause steric destabilisation. The destabilisation of a protein can cause misfolding, degradation and, in this case, interference with the signal peptide's hydrophobic core and cleavage site. This can affect a variety of factors including translocation and protein expressivity. If the MANF protein is unable to be expressed, the resulting lack of dopaminergic neuron protection and the limited response to ER and oxidative stress could be linked to the development of PD (**Figure 4.15**).

4.4 Conclusion

Understanding the role of an amino acid change concerning the protein structure using computational methods is useful to determine whether or not functional studies would be worth the time, cost and effort. In this chapter, we have elucidated the secondary structure of MANF and revealed that the variant is located within the h-region or helix of the signal peptide. Several variants in the signal peptide regions of different proteins have been previously associated with disease and confirmed via functional analysis (Karamyshev *et al.*, 2020). In the present study, using the protein sequence only, stability prediction tools found the variant to have an impact on the protein by reducing the $\Delta\Delta G$ and thereby destabilising the protein. Tertiary structure analysis using the complete theoretical wildtype model also revealed destabilisation of the protein upon the introduction of the variant. The variant was also found to cause rigidity of the protein around the cleavage region of the signal peptide. Furthermore, MD analyses that compared the wildtype and mutant structures indicated that there was a change in the conformational stability and flexibility of the protein after the introduction of the variant. This could potentially involve non-cleavage of the signal peptide which may render the protein inactive or cause improper folding of the protein if the peptide is not cleaved off.

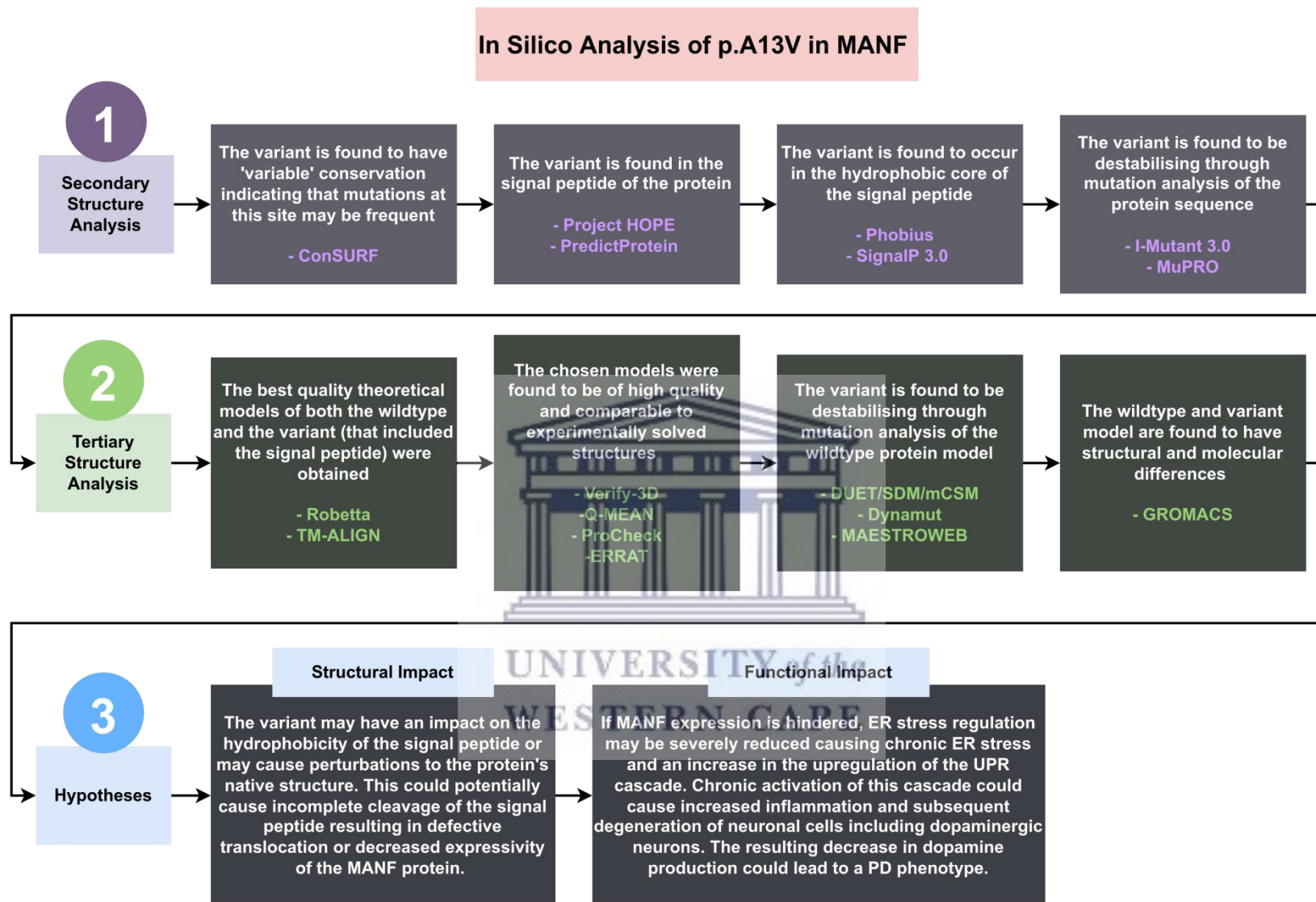


Figure 4.15: Summary of the results obtained through *in-silico* analysis of p.A13V in MANF

5

CHAPTER 5

Discussion and conclusions

This chapter briefly elucidates and discusses the main findings of the research study and their potential implications. Furthermore, the strengths and limitations of the study are explored and recommendations for future work for this study, as well as Parkinson's disease (PD) genomics in sub-Saharan Africa (SSA), are proposed.

It is important to note that the prevalence of PD is set to double in the next 20 years (GBD 2016 Parkinson's Disease Collaborators, 2016). This is a pressing concern, particularly among SSA countries due to increased life expectancies, as well as the limited information available regarding the cause of disease in understudied populations. Also, due to the recent Coronavirus disease 2019 (COVID-19) pandemic, there is speculation as to the associated risk of PD onset in individuals who have been infected by SARS-CoV-2, due to secondary neuroinflammation that can cause neurodegeneration and deterioration of nigrostriatal pathways (Leta *et al.*, 2022; Lippi *et al.*, 2020). Together, the occurrences of newly ageing populations and the possibility of viral-induced neurodegeneration may lead to a generational wave of PD cases in the near future, thus further fueling the need to understand the complex biological mechanisms underlying the disease (Baizabal-Carvallo and Alonso-Juarez, 2021).

Thus, the rationale for this study was to determine the potential cause of PD in a family of Xhosa African ancestry (a population group that has been severely understudied in PD research thus far) using previously successful whole exome sequencing (WES) techniques. This was done to determine whether novel genetic causes or susceptibility factors could be attributed to the onset of PD, or perhaps provide insight into the mechanisms underlying the disease in a family from an under-represented population.

5.1 Understudied populations in PD genetic research

African genomes, as opposed to those from European and Asian-based ancestries, tend to be comparatively diverse but severely understudied (Bentley *et al.*, 2020). Evolutionary genetic diversity, stratification of populations and the differences in genetic variation due to contrasting climates, exposure to infectious disease and even diet make individuals of African ancestry a necessary part of complex disease research (Campbell and Tishkoff, 2008). It is a hope that the inclusion of various

populations from this understudied ancestry group may lead to the identification of novel disease genes, susceptibility factors and even protective genetic factors involved in complex disease. Due to our overall limited understanding of PD genetics and its underlying mechanistic architecture, it is hoped that African populations may provide previously undiscovered genetic insight.

Thus far, a number of PD studies involving the screening of common PD genes for mutations or exonic rearrangements (including in *LRRK2*, *Parkin*, *PINK1* and *GBA*) have been performed in South Africa, however, low frequencies of pathogenic variants have been identified (Keyser *et al.*, 2010; Haylett *et al.*, 2012; du Toit *et al.*, 2019; Mahne *et al.*, 2016; Mahungu *et al.*, 2020). The use of NGS to study novel PD causes in sub-Saharan Africa (in individuals of African ancestry) is still relatively new. So far, a single study has incorporated the use of a targeted NGS panel containing 751 genes to screen for novel variants in 33 Black South African and 13 Nigerian PD-affected individuals but no known pathogenic variants were identified (Oluwole *et al.*, 2020). Thus, the probability of a novel gene being implicated within this family, after having been screened for known or probable genetic causes, is high.

This study explored the genetic basis of PD in a South African family of Xhosa ancestry (ZA 15). Notably, this study is the first known example of research into familial PD utilising NGS analysis in a family of African ancestry. These types of benchmark studies in rarely studied populations allow researchers to look at unique methods of analysing genomic data to discover novel causes and possibly understand how to diagnose or treat PD using alternate, newer methods. Furthermore, understanding the monogenic causes of PD could prove useful in the search for the more elusive causes of idiopathic PD, supporting the genetic analysis of PD families.

The Xhosa population of South Africa originate from the ancient Bantu and Khoi people who descended from North Africa to Eastern Africa (region of Africa's Great Lakes) before settling in Southern Africa (Newman, 1995). The modern Xhosa population group is descended from an admixture of the Northern Bantu and Southern San population groups (Newman, 1995). This group is now the second-largest population group behind the Zulus with most Xhosa people (~ 5.4 million people) occupying the Eastern Cape Province in South Africa (the region where ZA 15 were recruited). The study of South African Xhosa population genetics, particularly regarding neurological disorders, has been limited. Recently, a study analysed the exomes of ~1800 Xhosa individuals to determine both rare and common genetic risk factors for schizophrenia that are particular to this group, finding that schizophrenia behaves as an oligogenic disease in the Xhosa population, where a few damaging variants may be disease-causing (Gulsener *et al.*, 2020). As the largest cohort population consisting of South African Xhosa individuals, this cohort was incorporated into our study for variant screening purposes.

In another study, a cross-sectional survey was done to analyse awareness of PD among South African Xhosa individuals (25 individuals with PD, 98 control individuals and 31 traditional healers) resulted in only 18% being able to recognise the disease and almost a third believing the disease was caused by witchcraft and the affected individual should be removed from the community, indicating there is a significant lack of knowledge about the disease among black South Africans (Mokaya *et al.*, 2017). This highlights the limited number of individuals, let alone families, that would be knowledgeable or willing to receive an official PD diagnosis or partake in genetic studies. Regarding parkinsonisms, a homozygous frameshift deletion in the *PTRHDI* gene was found in a Xhosa family affected with juvenile-onset parkinsonism with intellectual disability (using WES, homozygosity mapping and linkage analysis), a gene that had produced a similar phenotype in two Iranian families (Khodadadi *et al.*, 2016). Though, to date, no WES studies have been done on a family of Xhosa ancestry with typical PD.

5.2 Main findings

This study aimed to determine the genetic cause of PD (assumed to be autosomal dominant) in a South African Xhosa family using WES. WES was conducted on four siblings (2 PD-affected and 2 non-affected individuals) and the resulting data was analysed using best-practise analysis tools. Known causes of PD were eliminated after screening for the presence of variants in both known and putative PD genes. Variants that were found to be heterozygous, non-synonymous, having minor allele frequencies (MAF) < 0.01 and a CADD score > 20 were earmarked for gene expression and pathway analysis. Twenty-four variants were found to be expressed in the brain and underwent Sanger sequencing for confirmational co-segregation analysis. Thereafter, the number of candidate variants was further reduced after screening through private (not publically available) population cohorts that were either PD- or ancestry-specific. The p.A13V variant in *MANF* was prioritised for further *in-silico* mutation analysis based on its biological role, including; promoting the survival of dopaminergic neurons in the *substantia nigra*.

The p.A13V variant was subjected to a variety of *in-silico* functional analyses to determine whether the variant could be implicated as a cause of PD in ZA 15. Phylogenetic analysis revealed that the variant occurs as a buried residue (occurring within the hydrophobic interior of a protein helix) with variable conservation. Secondary structure analysis revealed that the variant is present in the hydrophobic core of the signal peptide of the protein. Thereafter, the variant was predicted to be destabilising at the sequence level. For tertiary structural analysis, optimum theoretical models of the wildtype and mutant protein were constructed and validated. These structural protein models passed all the basic quality checks and were deemed appropriate for further structural analysis. Analysis of

the variant's effect on the wildtype structure also predicted a destabilising effect and an increase in rigidity of the signal peptide, close to the cleavage site. Lastly, short molecular dynamics (MD) simulations were performed on the wildtype and mutant models which indicated a change in the conformational stability and flexibility of the protein upon introduction of the variant, particularly at the N- and C-terminals of the protein.

5.3 Methodological approach in the present study

Performing WES analysis in PD-affected families has proven to be an efficient way of determining the presence of potentially pathogenic variants with the assistance of robust filtering workflows. A detailed look into all previously published articles outlining the use of WES in PD families (from various population groups) to ascertain potential disease variants was compiled and published (**Appendix D**). This comparative audit of each studies' methodology, bioinformatic tools, filtering criteria and variant prioritisation formed the basis of the methodology outlined for the analysis of the African Xhosa family, ZA 15. The methodology incorporated best practice tools that were used to create an optimum, reproducible pipeline. The inclusion of robust filtering techniques was employed to ensure that the most efficient and sensitive method for finding novel variants among affected siblings. The incorporation of tools such as Burrows-Wheeler Aligner (BWA) and the Genome Analysis Toolkit (GATk) and the Ensembl-Variant Effect Predictor allows the researcher to obtain a higher number of likely variants that could be disease-causing. These tools are widely-regarded as 'best-practice' due to their reproducibility and increased sensitivity during the formative steps of analysis. Furthermore, observing the prioritised variants in private cohorts, that were either PD- or ancestry specific allowed us to determine the true rarity of variants and further filter and prioritise potentially pathogenic variants. This formed a major strength of the study as many of the variants that were pathogenic and extremely rare in popular population databases, such as gNOMAD and the 1000 Genomes Project, were found to be significantly present in private population cohorts that included individuals of similar ancestry. This crucial step allowed us to eliminate variants that did not require further study. This highlights the necessity of creating databanks with genomic information obtained from individuals with African ancestry.

In-silico analyses serve as a determinative measure for predicting the possible effect of a variant on a protein and suggesting whether a resulting perturbation could be associated with the pathobiology of disease, before in-house, laboratory-based functional analysis. The methodological approach taken in this study was based on similar studies that investigated the effect of a single prioritised variant suspected to be an underlying cause of disease and incorporated both a sequence-based and structural approach (Hossain *et al.*, 2020; De Oliviera *et al.*, 2019). A secondary structural analysis of the protein under investigation using tools such as Project Hope and PredictProtein allowed for a

preliminary understanding of the region of the protein that is affected by the variant's position. These tools are useful due to the incorporation of multiple protein-based databases for detailed annotation of the protein structure, as well as the incorporation of predictive machine learning algorithms to compare the protein structure and domains to similar proteins. Secondary structure analysis of our protein indicated the presence of our variant with the signal peptide of the gene, a component that is not included in tertiary structure databases. Structure based-analysis, however, allows for analysis of the protein using its native conformation and movement. Robetta was used to build the theoretical wildtype and mutant structures using an *ab initio* approach, depicting a high-quality model of the proteins. This can be attributed to the advanced deep learning algorithms governing the builds that are continuously quality checked by Continuous Automated Model Evaluation (CAMEO) to ensure the protein models are comparable to the experimentally solved structures in the Protein Data Bank. Short molecular dynamics simulations using GROMACS were used to determine the effect of the conformational stability and flexibility of the protein models. In a study comparing the precision of MD simulations of the 5 most popular algorithms (including GROMACS, AMBER, LAMMPS, DESMOND and CHARMM), where it was found that GROMACS 5.1 showed a marked increase in performance of free energy calculations due to several new and enhanced paralleled algorithms (Sedova *et al.*, 2018; Abraham *et al.*, 2015).

5.4 p.A13V in MANF as a candidate for PD

The family ZA 15 were screened for both known and novel variants in PD genes using both laboratory and WES methods. However, no variants of significance were found, supporting the hypothesis that PD is caused by novel genetic factors in individuals with diverse genomes. Our study culminated in the nomination of a single variant (p.A13V) occurring in the *MANF* gene.

MANF is a small hormonal protein secreted by the *MANF* gene. It is a conserved neurotrophic factor protein that displays a protective role on mid-brain dopaminergic neurons and exerts a regulatory effect in response to endoplasmic reticulum (ER) stress (Yu *et al.*, 2021). Gene expression analysis revealed that *MANF* is expressed in various tissues in the body, but significantly expressed in the *substantia nigra*, the main region affected in PD. The accumulation of misfolded proteins disturbs the homeostatic environment in the ER, resulting in ER stress. ER stress then typically prompts the unfolded protein response (UPR) which involves three signalling pathways that assist in the degradation of misfolded proteins and attenuate subsequent protein synthesis. These pathways are activated by ER transmembrane proteins, namely, PERK (PRKR-like endoplasmic reticulum kinase), IRE1 (inositol requiring enzyme 1), and ATF6 (activating transcription factor 6). These proteins are inactive during cellular homeostasis and remain bound to an ER chaperone, GRP78/BiP (Glucose-regulated protein 78/Binding immunoglobulin protein). *MANF* has been found to have a

calcium-dependent interaction with the GRP78 chaperone which upregulates MANF secretion in response to ER stress (Glembotski *et al.*, 2012).

Thus far, studies on MANF as a potential therapeutic target have proposed that its induced overexpression could promote the survival of dopaminergic neurons in PD and other neurodegenerative disorders (Voutilainen *et al.*, 2009; Richman *et al.*, 2018, Zhang *et al.*, 2018). MANF has been found to improve mitochondrial function through the alleviation of oxidative stress in an MPTP/MPP⁺-induced model of PD (Liu *et al.*, 2018). Furthermore, it has also been found to have reduced progressive neuronal degradation and subsequent locomotive effects, as well as facilitated the removal of misfolded aggregates of α -synuclein in an α -synuclein *C. elegans* PD model (Zhang *et al.*, 2018). In the *Drosophila melanogaster* fly, the knockout of the MANF ortholog was seen to manifest as a deficiency in the development of DA neurons, though the deficiency was corrected upon the introduction of wildtype MANF (both fly and human orthologs) (Palgi *et al.*, 2009). MANF is known to exert protective, modulating mechanisms in response to ER stress, not unlike Parkin and leucine-rich repeat kinase 2 (*LRRK2*), which are genes that are implicated in PD (Takahashi *et al.*, 2003; Lee *et al.*, 2019). Postmortem brain tissue from PD-affected individuals has shown activation of the ER stress response, indicating a correlation with PD pathology (Conn *et al.*, 2004, Hoozemans *et al.*, 2007).

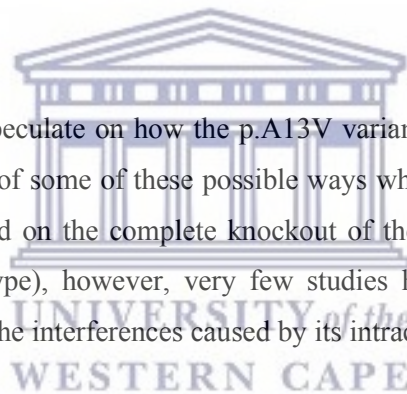
The loss of neurons in neurodegenerative disease has been previously associated with chronic ER stress, where the failure of the UPR cascade to restore cellular homeostasis can result in apoptosis. Chronic ER stress could be attributed to the prolonged production of mutant proteins thereby overwhelming the UPR cascade (Lin *et al.*, 2013). Thus, malfunctioning MANF may cause attenuation of ER stress regulation and decreased neuronal protection, thereby perpetuating common pathobiological pathways underlying PD.

The variant p.A13V in MANF is located within the hydrophobic core of the signal peptide and was found to be destabilising across all our *in silico* analyses. The AA substitution of alanine to valine may initially seem to be inconsequential as valine is also a hydrophobic AA and the difference between the two molecules includes an alkyl group. However, when an AA is substituted within a helix, steric torsion could be induced, allowing for aberrations in other functional regions of the downstream protein. More importantly, the change in hydrophobicity of the signal peptide, as well as the loss of polarity, due to the substitution may impact the processing of the immature protein.

Signal peptides play a critical role in targeting proteins in the ER, as well as dictating the translocation of the protein into the ER membrane (Liaci and Förster, 2021). The peptide contains a hydrophobic core that interacts with the signal recognition protein (SRP), which is a ribonucleoprotein found in the

cytosol that assists in translocation (Voorhees and Hegde, 2016). The resulting complex, the SRP-ribosome-nascent protein chain, enables the targeting of the protein to the ER where the SRP binds to a receptor on the ER membrane allowing the protein to travel across the membrane (Kapp *et al.*, 2009). The entire protein (including the signal peptide) is then transported to the Sec61 translocon where the mature protein is translocated into the lumen of the ER and the signal peptide is cleaved off by a signal peptidase. Thereafter, the mature protein travels through to the Golgi apparatus where it is secreted (Dudek *et al.*, 2015; Voorhees and Hegde, 2016). The signal sequence may thereby interact with different proteins during the early process of translocation, as well as after, during export to the Golgi Apparatus and subsequent secretion (Liaci and Förster, 2021). Variants present in the signal peptide of secretory proteins such as preprovasopressin and preproparathyroid hormone have been previously linked to the onset of diseases such as familial central diabetes insipidus and familial isolated hypoparathyroidism (FIH) (Ito *et al.*, 1993; Arnold *et al.*, 1990). This occurs due to a point mutation that disrupts the hydrophobic core of the signal peptide thereby impairing the processing of the hormone and resulting in a decrease in expression (Arnold *et al.*, 1990; Karaplis *et al.*, 1995). Furthermore, pathogenic variants in the hydrophobic core of the signal peptide have been found to interfere with cellular trafficking that ultimately prevents translocation of the protein (Rajpar *et al.*, 2002; Pidasheva *et al.*, 2005).

There are a number of ways to speculate on how the p.A13V variant may result in a PD phenotype. **Figure 5.1** provides an overview of some of these possible ways which are highlighted in the yellow boxes. Many studies have focused on the complete knockout of the MANF protein (which has not resulted in a typical PD phenotype), however, very few studies have focused on the effect of a heterozygous mutant protein and the interferences caused by its intracellular interactions.



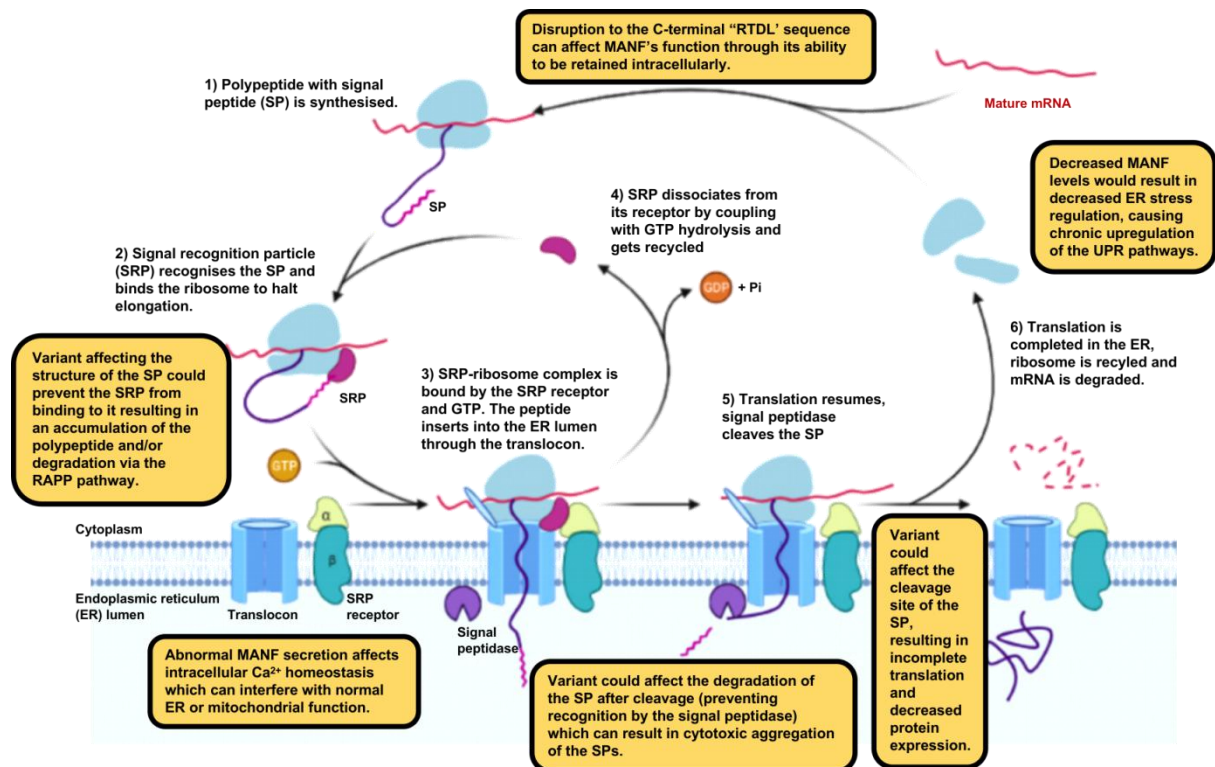


Figure 5.1: Model for the translational cycle of a secretory protein (MANF)

Inserts describing the possible effect of the variant on the protein at various stages of a secretory protein's translational cycle indicate the possibilities for how the variant could cause malfunction of the MANF protein. Created in Biorender.com.

A study was able to demonstrate that variations in the signal peptide of secretory proteins tend to cause mRNA degradation through the Regulation of Aberrant Protein Production (RAPP) pathway or downstream processing defects which can both lead to disease. Specifically, the presence of the signal peptide variant in the hydrophobic core is more likely to result in mRNA degradation via the RAPP pathway (Tikhonova *et al.*, 2019). This occurs as a result of the reduced hydrophobicity of the mutant signal peptide and its limited ability to be recognised by the SRP. Therefore, the RAPP pathway is initiated where the Argonaute-2 (AGO2) binds to the mutated signal sequence and leads to mRNA degradation of the protein (Popp and Maquat, 2014). This would cause a decrease in the levels of MANF, which could prompt a decrease in ER stress regulation and chronic upregulation of the UPR pathways. However, if the protein is ablated completely, the effects may not result in the PD phenotype as it has been shown in previous midbrain studies in mice that other neurotrophic factors are upregulated in the absence of MANF, and thus, ER stress would still be adequately regulated and dopaminergic neurons could remain protected (Pakarinen *et al.*, 2020). However, this study omitted the *MANF* gene altogether, and the effect of a defective protein may cause other aberrations in its complex role as a secretory protein.

There is also the possibility that the variant causes misfolding of the protein and interferes with the signal peptide cleavage. As illustrated, the signal peptide is cleaved off by a signal peptidase in the ER. If the signal peptide is not adequately cleaved off in the ER due to insufficient recognition of the signal peptide by the SRP, there could be an increased accumulation of the protein due to decreased disulfide bond formation and N-linked glycosylation processes (as the mature protein would not be able to pass through to the Golgi apparatus), which can cause ER stress or even cytotoxicity (Popp and Maquat, 2014). MANF would be unable to modulate the ER stress and may exacerbate it.

It has also been found that signal peptides can sometimes be released into the ER, secretory pathways or even the cytoplasm instead of being degraded by intramembrane proteases (Liaci *et al.*, 2021). This is important as the hydrophobicity of signal peptides makes them prone to potential aggregation. One study reported the concentration-dependent cytotoxicity of amyloid precursor protein signal peptides in SH-SY5Y cells that were caused by the formation of its amyloid-like aggregation *in vitro* (Gadhave *et al.*, 2021). If the signal peptide was capable of being cleaved but not degraded by the proteases within the cell, a cytotoxic aggregation of the hydrophobic peptides could occur, ultimately causing a deterioration of dopaminergic cells.

MANF also possesses a C-terminal signal sequence (RTDL) that interacts with the KDEL receptor (KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor) allowing the protein to be retained in the ER (Henderson *et al.*, 2013). This indicates that MANF has both a secretory and a retained intracellular function. If the RTDL sequence is disrupted (which may be likely based on the conformational stability analysis results produced by the MD simulations), the protein may lose its ability to be retained intracellularly, potentially interfering with processes that have not been elucidated yet.

MANF's secretion capability is also directly associated with cellular Ca²⁺ homeostasis due to the calcium-dependent binding of the protein to the GRP78/BiP complex (Eesmaa *et al.*, 2021). The fluctuation in calcium homeostasis due to the presence of a mutant MANF protein could also affect the function of the membrane potential in the ER, as well as the mitochondria. PD has been increasingly linked to mitochondrial dysfunction, particularly involving coenzyme Q deficiency and mutations occurring in the mitochondrial complex I (Eesmaa *et al.*, 2021). PD-associated toxins that are used to induce PD symptoms in animal models, including MPTP and paraquat, are also found to affect the functioning of the mitochondrial complex I. *In vivo* screening of DmMANF indicated significant interactions with genes involved in the ubiquinone synthesis pathway (including *COQ7*, *CG9249/COQ3* and *CG9613/COQ2*). This pathway is responsible for the synthesis of ubiquinone (or coenzyme Q), a protein involved in the electron transport chain whereby electrons are transferred from complexes I and II to the third complex. Furthermore, DmMANF was also found to interact with

2 homologues (NDUFS1 and FOXRED1) that are linked to human mitochondrial complex I deficiency. It was postulated that MANF may have either a direct or indirect association with mitochondrial function by affecting oxidative phosphorylation (Lindström *et al.*, 2017). Thus, the link between the mutant p.A13V MANF and its effect on mitochondrial function needs to be further examined.

Herewith, protein malfunction of MANF could inhibit the hormone of its protective effects on midbrain dopaminergic neurons, but rather cause cytotoxic degeneration of these cells. The loss of these neurons could result in a decrease in the overall quantity of dopamine produced thus indirectly directing the onset of a PD phenotype.

5.5 Limitations of the Study

The small sample size of PD-affected individuals available for research in ZA 15 is an indicator of the limitations of PD research in SSA. It is difficult to obtain a family pedigree consisting of multiple generations of affected individuals due to several reasons: 1) the lack of neurologists specialising in movement disorders and limited healthcare resources, 2) the frequency at which affected individuals contact traditional healers or hold traditional beliefs of the disease-preventing modern medical intervention and also, limited knowledge of the disease due to a lack of genetic studies within these population groups. The family analysed in this study consisted of both affected and unaffected siblings. Thus, identifying pathogenic variants within siblings can prove to be an arduous task as they share 50% of their DNA. It is also difficult to characterise co-segregation of a disease variant in a family that only possesses DNA for the siblings i.e. across one generation. Moreover, in this particular family, the presence of different ages of onset can indicate the presence of familial heterogeneity (different genetic causes of disease in the one family) or phenocopies where the environment can induce a form of the disease that is phenotypically similar to another individual in the family, who would have a genetic cause of disease. Although DNA from other members of the family (non-siblings) was obtained to observe the presence of the variant frequencies, DNA from more affected and unaffected family members would be necessary. This is particularly evident for the daughter of the affected female who possesses the *MANF* variant but is 40 years of age, and thus, may or may not end up developing PD as there is evidence of later onset PD in this family.

Although WES is the initial choice for the determination of novel disease genes in families, a number of concerns arise when solely using this method of analysis. WES limits the output of genomic data to only the exome, and even then, specific regions of the exome may be inadequately covered. This can lead to a loss of potentially disease-causing variants during analysis. Furthermore, WES is unable to detect structural variations in the genome including large genomic rearrangements such as copy

number variations and expanded repeat regions, as well as, long non-coding RNA or intronic regions. This could be detrimental when investigating a disease as multifactorial as PD, whose phenotype tends to be influenced by a large number of genetic aberrations. However, if this study were to serve as a benchmark for the analysis of PD family pedigrees like ZA 15, particularly in those sourced from understudied populations, WES is the optimum first line choice before performing more comprehensive sequencing.

5.6 Future work

For this study, a number of recommendations could be made for future work. Although the preliminary *in silico* analysis indicates a propensity towards protein destabilisation, the exact nature of the prioritised variant p.A13V and its effect on the translocation or expression of the MANF protein cannot be accurately determined. Although there are studies highlighting the effects of MANF knockout, it is important to perform functional analysis on the effect of this particular variant on the function of the protein and its specific implication regarding dopaminergic neurons in the *substantia nigra*. Thus, further laboratory-based analyses are advised.

Previous studies linking a variant in the signal peptide have shown a variety of methods to properly analyse the variant effect on the protein. To determine whether translocation of the protein is affected, MANF could be cloned using site-directed mutagenesis where transfection of the vector could be induced into neuronal cells to determine the transient transcription and translation of the protein *in vitro*. Further analysis to determine the proper cleavage of the signal peptide (using immunoblotting and proteolytic methods) as well as, the possibility for signal peptide aggregation within the cytosol (using fluorescence spectroscopy) could be later observed. Also, the link between aberrant protein function of MANF and its potential effect on mitochondrial function through homeostatic interference (through quantitative analysis of calcium concentration and mitochondrial RNA), should be explored.

Furthermore, *in vivo* analysis using an animal model could be an improved predictor of the effect of the variant. For example, the common fruit fly (*D. melanogaster*) could be an ideal model organism due to the high rate of homology (~60 %) of the gene sequence when compared to humans, and the similarity in neurological biomechanisms (Pandey and Nichols, 2011). It is also predicted that nearly 75% of human disease-associated genes have a functional homologue in *D. melanogaster*, including *DmMANF* (Mirzoyan *et al.*, 2019). Furthermore, studies have already implicated the upregulation of *MANF* in *D.melanogaster* in response to induced ER stress, where it behaves as a regulator of the unfolded protein response (UPR), highlighting the conserved role of the gene between species (Lindström *et al.*, 2016).

Although WES was found to be a suitable method of determining a potential genetic factor in the ZA 15 family, WGS could be used on this family to rule out any exonic rearrangements or copy number variations (CNVs) that would have been missed by WES. Also, third-generation sequencing or long-read sequencing are newly-developed approaches that aim to overcome the limitations of existing NGS methods. They produce long reads that are far more expansive, reducing the complexity of detecting read overlaps—thus increasing the quality of the sequencing data and improving CNV detection (Giani *et al.*, 2019). Furthermore, mutable regions in the genome may harbour pathogenic mutations, particularly compound heterozygous mutations that may only be discovered with long-read sequencing (Mantere *et al.*, 2019). Therefore, in the near future, long-read sequencing may be viewed as the more favourable sequencing alternative for disorders such as PD. As newer research methodologies develop and become available, it is important to note that new candidate genes may arise in this PD-affected family and be worth further study.

Finally, the Global Parkinson's Genetics Program (GP2; <http://gp2.org/>) is an initiative aiming to elucidate the genetic factors underlying PD by incorporating population cohorts from around the world (including those from under-represented populations) and exploring both monogenic causes, as well as, genetic risk factor meta-analysis (obtained by comparing the affected individuals with control cohorts). Our research group is in the process of submitting PD samples, including the proband sample of ZA 15, for analysis through GP2's monogenic working group in hopes of finding the novel genetic cause of disease in multiple individuals with diverse, understudied ancestry. These samples will be subjected to either whole genome or long-read sequencing dependent on a number of criteria including family history, age at onset and ethnicity (<https://gp2.org/working-groups/monogenic-network-working-group/>).

5.7 Concluding remarks

PD is a multifactorial disease that requires a faceted approach to determine the underlying pathobiology of the disease. Genetic research in PD has evolved significantly over the past two decades with varying strategies that have included linkage analysis to population-wide genome-wide association studies and more recently, NGS. Familial studies incorporating NGS have revolutionised novel disease gene discovery, however, best practice guidelines for data analysis need to be developed; considering diverse populations and ancestral origins, since it is apparent that a generic 'one-size-fits-all' approach will have significant limitations. In conclusion, determining the complex genetic architecture underlying PD, particularly in under-represented populations, is critical to providing insight into PD molecular mechanisms, detection of PD biomarkers, and elucidation of novel drug targets. Ultimately, this knowledge will change the course of future clinical diagnoses and therapeutic modalities for this currently incurable disorder.

References

- Abbas, M. M., Xu, Z., & Tan, L. C. S. (2018). Epidemiology of Parkinson's Disease—East Versus West. In *Movement Disorders Clinical Practice* (Vol. 5, Issue 1, pp. 14–28). Wiley-Blackwell. <https://doi.org/10.1002/mdc3.12568>
- Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindah, E. (2015). Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1–2, 19–25. <https://doi.org/10.1016/j.softx.2015.06.001>
- Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature methods*, 7(4), 248–249. <https://doi.org/10.1038/nmeth0410-248>
- Aftabuddin, M., & Kundu, S. (2007). Hydrophobic, hydrophilic, and charged amino acid networks within protein. *Biophysical Journal*, 93(1), 225–231. <https://doi.org/10.1529/biophysj.106.098004>
- Alcalay, R. N., Kehoe, C., Shorr, E., Battista, R., Hall, A., Simuni, T., Marder, K., Wills, A. M., Naito, A., Beck, J. C., Schwarzschild, M. A., & Nance, M. (2020). Genetic testing for Parkinson disease: current practice, knowledge, and attitudes among US and Canadian movement disorders specialists. *Genetics in Medicine*, 22(3), 574–580. <https://doi.org/10.1038/s41436-019-0684-x>
- Arnold, A., Horst, S. A., Gardella, T. J., Baba, H., Levine, M. A., & Kronenberg, H. M. (n.d.). *Mutation of the Signal Peptide-encoding Region of the Preproparathyroid Hormone Gene in Familial Isolated Hypoparathyroidism*.
- Andreu-Sánchez, S., Chen, L., Wang, D., Augustijn, H. E., Zhernakova, A., & Fu, J. (2021). A Benchmark of Genetic Variant Calling Pipelines Using Metagenomic Short-Read Sequencing. *Frontiers in genetics*, 12, 648229. <https://doi.org/10.3389/fgene.2021.648229>
- Anjos, S., Nguyen, A., Ounissi-Benkalha, H., Tessier, M. C., & Polychronakos, C. (2002). A common autoimmunity predisposing signal peptide variant of the cytotoxic T-lymphocyte antigen 4 results in inefficient glycosylation of the susceptibility allele. *Journal of Biological Chemistry*, 277(48), 46478–46486. <https://doi.org/10.1074/jbc.M206894200>
- Arnold, A., Horst, S. A., Gardella, T. J., Baba, H., Levine, M. A., & Kronenberg, H. M. (1990). Mutation of the signal peptide-encoding region of the preproparathyroid hormone gene in

- familial isolated hypoparathyroidism. *The Journal of clinical investigation*, 86(4), 1084–1087.
<https://doi.org/10.1172/JCI114811>
- Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., & Ben-Tal, N. (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic acids research*, 44(W1), W344–W350.
<https://doi.org/10.1093/nar/gkw408>
- Austin-Tse, C.A., Jobanputra, V., Perry, D.L. *et al.* (2022) Best practices for the interpretation and reporting of clinical whole genome sequencing. *npj Genom. Med.* 7, 27.
<https://doi.org/10.1038/s41525-022-00295-z>
- Baizabal-Carvalho, J., & Alonso-Juarez, M. (2021). The role of viruses in the pathogenesis of Parkinson's disease. In *Neural Regeneration Research* (Vol. 16, Issue 6, pp. 1200–1201). Wolters Kluwer Medknow Publications. <https://doi.org/10.4103/1673-5374.300437>
- Bandres-Ciga, S., Diez-Fairen, M., Kim, J. J., & Singleton, A. B. (2020). Genetics of Parkinson's disease: An introspection of its journey towards precision medicine. *Neurobiology of Disease*, 137(January), 104782. <https://doi.org/10.1016/j.nbd.2020.104782>
- Barazorda-Ccahuana, H. L., Valencia, D. E., Aguilar-Pineda, J. A., & Gómez, B. (2018). Art v 4 Protein Structure as a Representative Template for Allergen Profilins: Homology Modeling and Molecular Dynamics. *ACS Omega*, 3(12), 17254–17260.
<https://doi.org/10.1021/acsomega.8b02288>
- Barba, M., Czosnek, H., & Hadidi, A. (2013). Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses*, 6(1), 106–136.
<https://doi.org/10.3390/v6010106>
- Bartha, Á., & Gyórfy, B. (2019). Comprehensive Outline of Whole Exome Sequencing Data Analysis Tools Available in Clinical Oncology. *Cancers*, 11(11), 1725.
<https://doi.org/10.3390/cancers11111725>
- Batut, B., Hiltemann, S., Bagnacani, A., Baker, D., Bhardwaj, V., Blank, C., Bretaudeau, A., Brillet-Guéguen, L., Čech, M., Chilton, J., Clements, D., Doppelt-Azeroual, O., Erxleben, A., Freeberg, M. A., Gladman, S., Hoogstrate, Y., Hotz, H. R., Houwaart, T., Jagtap, P., ... Grüning, B. (2018). Community-Driven Data Analysis Training for Biology. *Cell Systems*, 6(6), 752-758.e1. <https://doi.org/10.1016/j.cels.2018.05.012>

- Bayrak, C., & Itan, Y. (2020). Identifying disease-causing mutations in genomes of single patients by computational approaches. *Human Genetics*, 139(6–7), 769–776. <https://doi.org/10.1007/s00439-020-02179-7>
- Begley D, Schofield PN, Sundberg JP. (2022) The Mouse Online: Open Mouse Biology and Pathology Data Resources for Biomedical Research. In: Pathology of genetically engineered mice and other mutants. Sundberg JP, Vogel P, and Ward JM, eds., Wiley. 6-21.
- Behjati, S., & Tarpey, P. S. (2013). What is next-generation sequencing? *Archives of Disease in Childhood: Education and Practice Edition*, 98(6), 236–238. <https://doi.org/10.1136/archdischild-2013-304340>
- Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A., Shang, L., Boisson, B., Casanova, J. L., & Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences of the United States of America*, 112(17), 5473–5478. <https://doi.org/10.1073/pnas.1418631112>
- Bellissent-Funel MC, Hassanali A, Havenith M, Henchman R, Pohl P, Sterpone F, van der Spoel D, Xu Y, Garcia AE. Water Determines the Structure and Dynamics of Proteins. *Chem Rev*. 2016 Jul 13;116(13):7673-97. doi: 10.1021/acs.chemrev.5b00664. Epub 2016 May 17. PMID: 27186992; PMCID: PMC7116073.
- Bentley, A. R., Callier, S. L., & Rotimi, C. N. (2020). Evaluating the promise of inclusion of African ancestry populations in genomics. *Npj Genomic Medicine*, 5(1), 1–9. <https://doi.org/10.1038/s41525-019-0111-x>
- Bentley, S. R., Guella, I., Sherman, H. E., Neuendorf, H. M., Sykes, A. M., Fowdar, J. Y., Silburn, P. A., Wood, S. A., Farrer, M. J., & Mellick, G. D. (2021). *Genome Era*.
- Bomba, L., Walter, K., & Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. In *Genome Biology* (Vol. 18, Issue 1). BioMed Central Ltd. <https://doi.org/10.1186/s13059-017-1212-4>
- Bonifati, V. (2014). Genetics of Parkinson's disease - state of the art, 2013. *Parkinsonism and Related Disorders*, 20(SUPPL.1), S23–S28. [https://doi.org/10.1016/S1353-8020\(13\)70009-9](https://doi.org/10.1016/S1353-8020(13)70009-9)

- Bough, R., & Dayan, F. E. (2022). Biochemical and structural characterization of quizalofop-resistant wheat acetyl-CoA carboxylase. *Scientific Reports*, 12(1). <https://doi.org/10.1038/s41598-021-04280-x>
- Bray, S. A., Senapathi, T., Barnett, C. B., & Grüning, B. A. (2020). Intuitive, reproducible high-throughput molecular dynamics in Galaxy: A tutorial. *Journal of Cheminformatics*, 12(1). <https://doi.org/10.1186/s13321-020-00451-6>
- Brunelli, L., Jenkins, S. M., Gudgeon, J. M., Bleyl, S. B., Miller, C. E., Tvrđik, T., Dames, S. A., Ostrander, B., Daboub, J. A. F., Zielinski, B. A., Zinkhan, E. K., Underhill, H. R., Wilson, T., Bonkowsky, J. L., Yost, C. C., Botto, L. D., Jenkins, J., Pysher, T. J., Bayrak-Toydemir, P., & Mao, R. (2019). Targeted gene panel sequencing for the rapid diagnosis of acutely ill infants. *Molecular genetics & genomic medicine*, 7(7), e00796. <https://doi.org/10.1002/mgg3.796>
- Buckner, C. A., Lafrenie, R. M., Dénoimée, J. A., Caswell, J. M., Want, D. A., Gan, G. G., Leong, Y. C., Bee, P. C., Chin, E., Teh, A. K. H., Picco, S., Villegas, L., Tonelli, F., Merlo, M., Rigau, J., Diaz, D., Masuelli, M., Korrapati, S., Kurra, P., ... Mathijssen, R. H. J. (2016). We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists TOP 1%. *Intech, 11(tourism)*, 13. <https://www.intechopen.com/books/advanced-biometric-technologies/liveness-detection-in-biometrics>
- Campbell, M. C., & Tishkoff, S. A. (2008). African genetic diversity: Implications for human demographic history, modern human origins, and complex disease mapping. *Annual Review of Genomics and Human Genetics*, 9, 403–433. <https://doi.org/10.1146/annurev.genom.9.081307.164258>
- Chen, S. M., Ma, K. Y., & Zeng, J. (2011). Pseudogene: Lessons from PCR bias, identification and resurrection. *Molecular Biology Reports*, 38(6), 3709–3715. <https://doi.org/10.1007/s11033-010-0485-4>
- Choudhury, A., Aron, S., Botigué, L. R., Sengupta, D., Botha, G., Bensellak, T., Wells, G., Kumuthini, J., Shriner, D., Fakim, Y. J., Ghoorah, A. W., Dareng, E., Odia, T., Falola, O., Adebisi, E., Hazelhurst, S., Mazandu, G., Nyangiri, O. A., Mbiyavanga, M., ... Hanchard, N. A. (2021). Author Correction: High-depth African genomes inform human migration and health (Nature, (2020), 586, 7831, (741-748), 10.1038/s41586-020-2859-7). *Nature*, 592(7856), E26. <https://doi.org/10.1038/s41586-021-03286-9>

- Choudhury, A., Aron, S., Botigué, L. R., Sengupta, D., Botha, G., Bensellak, T., Wells, G., Kumuthini, J., Shriner, D., Fakim, Y. J., Ghoorah, A. W., Dareng, E., Odia, T., Falola, O., Adebiyi, E., Hazelhurst, S., Mazandu, G., Nyangiri, O. A., Mbiyavanga, M., ... Hanchard, N. A. (2020). High-depth African genomes inform human migration and health. *Nature*, *586*(7831), 741–748. <https://doi.org/10.1038/s41586-020-2859-7>
- Coates, D. R., Chin, J. M., & Chung, S. T. L. (2011). 基因的改变 NIH Public Access. *Bone*, *23*(1), 1–7. <https://doi.org/10.1016/j.tibs.2014.02.001.Defective>
- Conn, K. J., Gao, W., McKee, A., Lan, M. S., Ullman, M. D., Eisenhauer, P. B., Fine, R. E., & Wells, J. M. (2004). Identification of the protein disulfide isomerase family member PDip in experimental Parkinson's disease and Lewy body pathology. *Brain Research*, *1022*(1–2), 164–172. <https://doi.org/10.1016/j.brainres.2004.07.026>
- Correia Guedes, L., Ferreira, J. J., Rosa, M. M., Coelho, M., Bonifati, V., and Sampaio, C. (2010). Worldwide Frequency of G2019S LRRK2 Mutation in Parkinson's Disease: a Systematic Review. *Parkinsonism Relat. Disord.* *16* (4), 237–242. <https://doi:10.1016/j.parkreldis.2009.11.004>
- Coulthurst, S. J., Dawson, A., Hunter, W. N., & Sargent, F. (2012). Conserved signal peptide recognition systems across the prokaryotic domains. *Biochemistry*, *51*(8), 1678–1686. <https://doi.org/10.1021/bi201852d>
- Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., ... Flicek, P. (2022). Ensembl 2022. *Nucleic acids research*, *50*(D1), D988–D995. <https://doi.org/10.1093/nar/gkab1049>
- Davis-Turak J, Courtney SM, Hazard ES, Glen WB Jr, da Silveira WA, Wesselman T, Harbin LP, Wolf BJ, Chung D, Hardiman G. Genomics pipelines and data integration: challenges and opportunities in the research setting. *Expert Rev Mol Diagn.* 2017 Mar;*17*(3):225-237. doi: 10.1080/14737159.2017.1282822. Epub 2017 Jan 25. PMID: 28092471; PMCID: PMC5580401.
- Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology*, *6*(12). <https://doi.org/10.1371/journal.pcbi.1001025>
- Day, J. O., and Mullin, S. (2021). The Genetics of Parkinson's Disease and Implications for Clinical Practice. *Genes* *12* (7), 1006. <https://doi:10.3390/genes12071006>

- De Oliveira, C. C. S., Pereira, G. R. C., De Alcantara, J. Y. S., Antunes, D., Caffarena, E. R., & De Mesquita, J. F. (2019). In silico analysis of the V66M variant of human BDNF in psychiatric disorders: An approach to precision medicine. *PloS one*, 14(4), e0215508. <https://doi.org/10.1371/journal.pone.0215508>
- Deans, Z. C., Costa, J. L., Cree, I., Dequeker, E., Edsjö, A., Henderson, S., Hummel, M., Ligtenberg, M. J., Loddo, M., Machado, J. C., Marchetti, A., Marquis, K., Mason, J., Normanno, N., Rouleau, E., Schuurin, E., Snelson, K. M., Thunnissen, E., Tops, B., ... Hall, J. A. (2017). Integration of next-generation sequencing in clinical diagnostic molecular pathology laboratories for analysis of solid tumours; an expert opinion on behalf of IQN Path ASBL. In *Virchows Archiv* (Vol. 470, Issue 1, pp. 5–20). Springer. <https://doi.org/10.1007/s00428-016-2025-7>
- Disratthakit, A., Toyo-Oka B, L., Thawong, P., Paiboonsiri, P., Wichukjinda, N., Ajawatanawong, P., Thipkrua, N., Suthum, K., Palittapongarnpim, P., Tokunaga, K., & Mahasirimongkol, S. (2019). *An optimized genomic VCF workflow for precise identification of Mycobacterium tuberculosis cluster from cross-platform whole genome sequencing data.*
- Dorsey, E. R., Sherer, T., Okun, M. S., & Bloem, D. B. R. (2018). The emerging evidence of the Parkinson pandemic. *Journal of Parkinson's Disease*, 8(s1), S3–S8. <https://doi.org/10.3233/JPD-181474>
- Dotchin, C., & Walker, R. (2012). The management of Parkinson's disease in sub-Saharan Africa. In *Expert Review of Neurotherapeutics* (Vol. 12, Issue 6, pp. 661–666). <https://doi.org/10.1586/ern.12.52>
- du Toit, N., van Coller, R., Anderson, D. G., Carr, J., & Bardiën, S. (2019). Frequency of the LRRK2 G2019S mutation in South African patients with Parkinson's disease. *Neurogenetics*, 20(4), 215–218. <https://doi.org/10.1007/s10048-019-00588-z>
- Dudek, J., Pfeffer, S., Lee, P. H., Jung, M., Cavalié, A., Helms, V., Förster, F., & Zimmermann, R. (2015). Protein transport into the human endoplasmic reticulum. *Journal of molecular biology*, 427(6 Pt A), 1159–1175. <https://doi.org/10.1016/j.jmb.2014.06.011>
- Duraes, C., Gomes, CP., Costa JL., Quagliata L. (2022). Demystifying the Discussion of Sequencing Panel Size in Oncology Genetic Testing. *EMJReviewsEMJ*. 2022;7[2]:68-77. DOI/10.33590/emj/22C9259. <https://doi.org/10.33590/emj/22C9259>.
- Eesmaa, A., Yu, L. Y., Göös, H., Nöges, K., Kovaleva, V., Hellman, M., Zimmermann, R., Jung, M., Permi, P., Varjosalo, M., Lindholm, P., & Saarma, M. (2021). The cytoprotective protein MANF promotes neuronal survival independently from its role as a GRP78 cofactor. *The Journal of biological chemistry*, 296, 100295. <https://doi.org/10.1016/j.jbc.2021.100295>

- El-Fishawy, P. (2013). "Common Disease-Rare Variant Hypothesis," in *Encyclopedia of Autism Spectrum Disorders* (New York: Springer), 720–722. https://doi:10.1007/978-1-4419-1698-3_1997
- Farlow, J. L., Robak, L. A., Hetrick, K., Bowling, K., Boerwinkle, E., CobanAkdemir, Z. H., *et al.* (2016). Whole-Exome Sequencing in Familial Parkinson Disease. *JAMA Neurol.* 73 (1), 68–75. <https://doi:10.1001/jamaneurol.2015.3266>
- Fatkin, D., & Johnson, R. (2020). Variants of Uncertain Significance and "Missing Pathogenicity." *Journal of the American Heart Association*, 9(3), 10–12. <https://doi.org/10.1161/JAHA.119.015588>
- Federici, G., & Soddu, S. (2020). Variants of uncertain significance in the era of high-throughput genome sequencing: A lesson from breast and ovary cancers. In *Journal of Experimental and Clinical Cancer Research* (Vol. 39, Issue 1). BioMed Central Ltd. <https://doi.org/10.1186/s13046-020-01554-6>
- Fernandez-Marmiesse, A., Gouveia, S., & Couce, M. L. (2017). NGS Technologies as a Turning Point in Rare Disease Research , Diagnosis and Treatment. *Current Medicinal Chemistry*, 25(3), 404–432. <https://doi.org/10.2174/0929867324666170718101946>
- Fingerhut, A., Reutrakul, S., Knuedeler, S. D., Moeller, L. C., Greenlee, C., Refetoff, S., & Janssen, O. E. (2004). Partial Deficiency of Thyroxine-Binding Globulin-Allentown Is Due to a Mutation in the Signal Peptide. *Journal of Clinical Endocrinology and Metabolism*, 89(5), 2477–2483. <https://doi.org/10.1210/jc.2003-031613>
- Funayama, M., Ohe, K., Amo, T., Furuya, N., Yamaguchi, J., Saiki, S., Li, Y., Ogaki, K., Ando, M., Yoshino, H., Tomiyama, H., Nishioka, K., Hasegawa, K., Saiki, H., Satake, W., Mogushi, K., Sasaki, R., Kokubo, Y., Kuzuhara, S., ... Hattori, N. (2015). CHCHD2 mutations in autosomal dominant late-onset Parkinson's disease: A genome-wide linkage and sequencing study. *The Lancet Neurology*, 14(3), 274–282. [https://doi.org/10.1016/S1474-4422\(14\)70266-2](https://doi.org/10.1016/S1474-4422(14)70266-2)
- Gadhve, K., Bhardwaj, T., Uversky, V. N., Vendruscolo, M., & Giri, R. (2021). The signal peptide of the amyloid precursor protein forms amyloid-like aggregates and enhances A β 42 aggregation. *Cell Reports Physical Science*, 2(10), 100599. <https://doi.org/10.1016/j.xcrp.2021.100599>
- Gasser, T. (2015). Usefulness of genetic testing in PD and PD trials: A balanced review. *Journal of Parkinson's Disease*, 5(2), 209–215. <https://doi.org/10.3233/JPD-140507>

- Germer, E. L., Imhoff, S., Vilariño-Güell, C., Kasten, M., Seibler, P., Brüggemann, N., Klein, C., and Trinh, J. International Parkinson's Disease Genomics Consortium (2019). The Role of Rare Coding Variants in Parkinson's Disease GWAS Loci. *Front. Neurol.* 10, 1284. doi:10.3389/fneur.2019.01284
- Giani, A. M., Gallo, G. R., Gianfranceschi, L., & Formenti, G. (2020). Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal*, 18, 9–19. <https://doi.org/10.1016/j.csbj.2019.11.002>
- Gibb, W. R., & Lees, A. J. (1988). The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease. *Journal of neurology, neurosurgery, and psychiatry*, 51(6), 745–752. <https://doi.org/10.1136/jnnp.51.6.745>
- Glembotski, C. C., Thuerauf, D. J., Huang, C., Vekich, J. A., Gottlieb, R. A., & Doroudgar, S. (2012). Mesencephalic astrocyte-derived neurotrophic factor protects the heart from ischemic damage and is selectively secreted upon sarco/endoplasmic reticulum calcium depletion. *Journal of Biological Chemistry*, 287(31), 25893–25904. <https://doi.org/10.1074/jbc.M112.356345>
- Goetz, C. G. (2011). The history of Parkinson's disease: Early clinical descriptions and neurological therapies. *Cold Spring Harbor Perspectives in Medicine*, 1(1). <https://doi.org/10.1101/cshperspect.a008862>
- Gulsuner, S., Stein, D. J., Susser, E. S., Sibeko, G., Pretorius, A., Walsh, T., Majara, L., Mndini, M. M., Mqulwana, S. G., Ntola, O. A., Casadei, S., Ngqengelele, L. L., Korchina, V., Van Der Merwe, C., Malan, M., Fader, K. M., Feng, M., Willoughby, E., Muzny, D., ... McClellan, J. M. (n.d.). *Genetics of schizophrenia in the South African Xhosa*. <https://www.science.org>
- Guo, H. B., Perminov, A., Bekele, S., Kedziora, G., Farajollahi, S., Varaljay, V., Hinkle, K., Molinero, V., Meister, K., Hung, C., Dennis, P., Kelley-Loughnane, N., & Berry, R. (2022). AlphaFold2 models indicate that protein sequence determines both structure and dynamics. *Scientific Reports*, 12(1). <https://doi.org/10.1038/s41598-022-14382-9>
- Gutierrez-Rodrigues, F., & Calado, R. T. (2018). The interpretation of rare or novel variants: damaging vs. disease-causing. *Hematology, Transfusion and Cell Therapy*, 40(1), 3–4. <https://doi.org/10.1016/j.bjhh.2017.10.003>
- Hamid, E., Ayele, B. A., Massi, D. G., Ben Sassi, S., Tibar, H., Djonga, E. E., El-Sadig, S. M., Amer El Khedoud, W., Razafimahefa, J., Kouame-Assouan, A. E., Ben-Adji, D., Lengané, Y. T. M., Musubire, A. K., Mohamed, M. H., Phiri, T. E., Nestor, N., Alwahchi, W. A., Neshuku, S. N.,

- Ocampo, C., Sakadi, F., ... Shalash, A. (2021). Availability of Therapies and Services for Parkinson's Disease in Africa: A Continent-Wide Survey. *Movement disorders : official journal of the Movement Disorder Society*, 36(10), 2393–2407. <https://doi.org/10.1002/mds.28669>
- Hancock, D. B., Martin, E. R., Stajich, J. M., Jewett, R., Stacy, M. A., Scott, B. L., Vance, J. M., & Scott, W. K. (2007). Smoking, Caffeine, and Nonsteroidal Anti-inflammatory Drugs in Families With Parkinson Disease. In *Arch Neurol* (Vol. 64). <https://jamanetwork.com/>
- Haylett, W. L., Keyser, R. J., du Plessis, M. C., van der Merwe, C., Blanckenberg, J., Lombard, D., Carr, J., & Bardien, S. (2012). Mutations in the parkin gene are a minor cause of Parkinson's disease in the South African population. *Parkinsonism and Related Disorders*, 18(1), 89–92. <https://doi.org/10.1016/j.parkreldis.2011.09.022>
- Hemminki, K., Försti, A., and Bermejo, J. L. (2008). The 'common DiseaseCommon Variant' Hypothesis and Familial Risks. *PloS one* 3 (6), e2504. <https://doi:10.1371/journal.pone.0002504>
- Henderson, M. J., Richie, C. T., Airavaara, M., Wang, Y., & Harvey, B. K. (2013). Mesencephalic astrocyte-derived neurotrophic factor (MANF) secretion and cell surface binding are modulated by KDEL receptors. *The Journal of biological chemistry*, 288(6), 4209–4225. <https://doi.org/10.1074/jbc.M112.400648>
- Hintzsche, J. D., Robinson, W. A., & Tan, A. C. (2016). *A Survey of Computational Tools to analyse and Interpret Whole Exome Sequencing Data*. <https://doi.org/10.1155/2016/7983236>
- Hommelsheim, C. M., Frantzeskakis, L., Huang, M., & Ülker, B. (2014). PCR amplification of repetitive DNA: A limitation to genome editing technologies and many other applications. *Scientific Reports*, 4, 1–13. <https://doi.org/10.1038/srep05052>
- Hoozemans, J. J. M., van Haastert, E. S., Eikelenboom, P., de Vos, R. A. I., Rozemuller, J. M., & Scheper, W. (2007). Activation of the unfolded protein response in Parkinson's disease. *Biochemical and Biophysical Research Communications*, 354(3), 707–711. <https://doi.org/10.1016/j.bbrc.2007.01.043>
- Hossain, M.S., Roy, A.S. & Islam, M.S. (2020) In silico analysis predicting effects of deleterious SNPs of human RASSF5 gene on its structure and functions. *Sci Rep* 10, 14542. <https://doi.org/10.1038/s41598-020-71457-1>
- Huber, C. D., Kim, B. Y., & Lohmueller, K. E. (2020). Population genetic models of GERP scores suggest pervasive turnover of constrained sites across mammalian evolution. *PLoS Genetics*, 16(5), 1–26. <https://doi.org/10.1371/journal.pgen.1008827>

- Iqbal, S., Erez-Palma, E. P., Jespersen, J. B., May, P., Hoksza, D., Heyne, H. O., Ahmed, S. S., Rifat, Z. T., Rahman, M. S., Lage, K., Palotie, A., Cottrell, J. R., Wagner, F. F., Daly, M. J., Campbell, A. J., & Lal, D. (n.d.). *Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants*. <https://doi.org/10.1073/pnas.2002660117/-/DCSupplemental.y>
- Ito, M., Oiso, Y., Murase, T., Kondo, K., Saito, H., Chinzei, T., Racchi, M., & Lively, M. O. (1993). Possible involvement of inefficient cleavage of preprovasopressin by signal peptidase as a cause for familial central diabetes insipidus. *The Journal of clinical investigation*, *91*(6), 2565–2571. <https://doi.org/10.1172/JCI116494>
- Jagadeesh, K. A., Wenger, A. M., Berger, M. J., Guturu, H., Stenson, P. D., Cooper, D. N., Bernstein, J. A., & Bejerano, G. (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature Genetics*, *48*(12), 1581–1586. <https://doi.org/10.1038/ng.3703>
- Jankovic, J., & Tan, E. K. (2020). Parkinson's disease: Etiopathogenesis and treatment. *Journal of Neurology, Neurosurgery and Psychiatry*, *91*(8), 795–808. <https://doi.org/10.1136/jnnp-2019-322338>
- Jarjanazi, H., Savas, S., Pabalan, N., Dennis, J.W., & Ozçelik, H. (2007). Biological implications of SNPs in signal peptide domains of human proteins. *Proteins: Structure*, *70*.
- Jiang, D., Niwa, M., Koong, A. C., & Diego, S. (2016). *Stdg.* *10*(10), 48–56. <https://doi.org/10.1038/nprot.2015.105.Genomic>
- Jumper, J., Evans, R., Pritzel, A. *et al.* (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* *596*, 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Jääntti, M., & Harvey, B. K. (2020). Trophic activities of endoplasmic reticulum proteins CDNF and MANF. *Cell and Tissue Research*, *382*(1), 83–100. <https://doi.org/10.1007/s00441-020-03263-0>
- Kalinderi, K., Bostantjopoulou, S., & Fidani, L. (2016). The genetic background of Parkinson's disease: current progress and future prospects. *Acta Neurologica Scandinavica*, *134*(5), 314–326. <https://doi.org/10.1111/ane.12563>
- Kapp K, Schrempf S, Lemberg MK, *et al* (2000 - 2013). Post-Targeting Functions of Signal Peptides. In: Madame Curie Bioscience Database [Internet]. Austin (TX): Landes Bioscience

- Karamyshev, A. L., Tikhonova, E. B., & Karamysheva, Z. N. (2020). Translational control of secretory proteins in health and disease. In *International Journal of Molecular Sciences* (Vol. 21, Issue 7). MDPI AG. <https://doi.org/10.3390/ijms21072538>
- Karaplis, A. C., Lim, S. K., Baba, H., Arnold, A., & Kronenberg, H. M. (1995). Inefficient membrane targeting, translocation, and proteolytic processing by signal peptidase of a mutant preproparathyroid hormone protein. *Journal of Biological Chemistry*, 270(4), 1629–1635. <https://doi.org/10.1074/jbc.270.4.1629>
- Keyser, R. J., Lombard, D., Veikondis, R., Carr, J., & Bardien, S. (2010). Analysis of exon dosage using MLPA in South African Parkinson's disease patients. *Neurogenetics*, 11(3), 305–312. <https://doi.org/10.1007/s10048-009-0229-6>
- Khodadadi H, Azcona LJ, Aghamollaii V, Omrani MD, Garshasbi M, Taghavi S, Tafakhori A, Shahidi GA, Jamshidi J, Darvish H, Paisán-Ruiz C. PTRHD1 (C2orf79) mutations lead to autosomal-recessive intellectual disability and parkinsonism. *Mov Disord*. 2017 Feb;32(2):287-291. doi: 10.1002/mds.26824. Epub 2016 Oct 18. PMID: 27753167; PMCID: PMC5318269.
- Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3), 310–315. <https://doi.org/10.1038/ng.2892>
- Klein, C., & Westenberger, A. (2012). Genetics of Parkinson's disease. *Cold Spring Harbor Perspectives in Medicine*, 2(1). <https://doi.org/10.1101/cshperspect.a008888>
- Kouli, A., Torsney, K. M., & Kuan, W.-L. (2018). Parkinson's Disease: Etiology, Neuropathology, and Pathogenesis Parkinson's Disease: Etiology, Neuropathology, and Pathogenesis Kouli, A. (2018). Parkinson's Disease: Etiology, Neuropathology, and Pathogenesis. 3–26. *Parkinson's Disease: Pathogenesis and Clinical Aspects*, 3–26.
- Ku, C. S., Cooper, D. N., & Patrinos, G. P. (2017). The Rise and Rise of Exome Sequencing. *Public Health Genomics*, 19(6), 315–324. <https://doi.org/10.1159/000450991>
- Kuipers, D. J. S., Carr, J., Bardien, S., Thomas, P., Sebaste, B., Breedveld, G. J., van Minkelen, R., Brouwer, R. W. W., van Ijcken, W. F. J., van Slegtenhorst, M. A., Bonifati, V., & Quadri, M. (2018). PTRHD1 Loss-of-function mutation in an african family with juvenile-onset Parkinsonism and intellectual disability. *Movement disorders : official journal of the Movement Disorder Society*, 33(11), 1814–1819. <https://doi.org/10.1002/mds.27501>

- Kulski, J. , (Ed.). (2016). Next Generation Sequencing - Advances, Applications and Challenges. IntechOpen. <https://doi.org/10.5772/60489>
- Kwong, A., Ho, C. Y. S., Shin, V. Y., Au, C. H., Chan, T. L., & Ma, E. S. K. (2022). How does re-classification of variants of unknown significance (VUS) impact the management of patients at risk for hereditary breast cancer? *BMC Medical Genomics*, 15(1). <https://doi.org/10.1186/s12920-022-01270-4>
- Landrum, M. J., Chitipiralla, S., Brown, G. R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., Lyoshin, V., Maddipatla, Z., Maiti, R., Mitchell, J., O'Leary, N., Riley, G. R., Shi, W., Zhou, G., Schneider, V., Maglott, D., ... Kattman, B. L. (2020). ClinVar: improvements to accessing data. *Nucleic acids research*, 48(D1), D835–D844. <https://doi.org/10.1093/nar/gkz972>
- Landrum, M. J., Chitipiralla, S., Brown, G. R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., Lyoshin, V., Maddipatla, Z., Maiti, R., Mitchell, J., O'Leary, N., Riley, G. R., Shi, W., Zhou, G., Schneider, V., ... Kattman, B. L. (2020). ClinVar: Improvements to accessing data. *Nucleic Acids Research*, 48(D1), D835–D844. <https://doi.org/10.1093/nar/gkz972>
- Lee, J. H., Han, J. H., Kim, H., Park, S. M., Joe, E. H., & Jou, I. (2019). Parkinson's disease-associated LRRK2-G2019S mutant acts through regulation of SERCA activity to control ER stress in astrocytes. *Acta Neuropathologica Communications*, 7(1), 68. <https://doi.org/10.1186/s40478-019-0716-4>
- Leggett, R. M., Ramirez-Gonzalez, R. H., Clavijo, B. J., Waite, D., & Davey, R. P. (2013). Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. In *Frontiers in Genetics* (Vol. 4, Issue DEC). <https://doi.org/10.3389/fgene.2013.00288>
- Lekoubou, A., Echouffo-Tcheugui, J. B., & Kengne, A. P. (2014). Epidemiology of neurodegenerative diseases in sub-Saharan Africa: A systematic review. In *BMC Public Health* (Vol. 14, Issue 1). BioMed Central Ltd. <https://doi.org/10.1186/1471-2458-14-653>
- Lesage, S., & Brice, A. (2009). Parkinson's disease: From monogenic forms to genetic susceptibility factors. In *Human Molecular Genetics* (Vol. 18, Issue R1). <https://doi.org/10.1093/hmg/ddp012>

- Lesage, S., Dürr, A., Tazir, M., Lohmann, E., Leutenegger, A.-L., Janin, S., Pollak, P., & Brice, A. (2006). LRRK2 G2019S as a Cause of Parkinson's Disease in North African Arabs . *New England Journal of Medicine*, 354(4), 422–423. <https://doi.org/10.1056/nejmc055540>
- Lesage, S., Dürr, A., Tazir, M., Lohmann, E., Leutenegger, A.-L., Janin, S., Pollak, P., & Brice, A. (2006). LRRK2 G2019S as a Cause of Parkinson's Disease in North African Arabs . *New England Journal of Medicine*, 354(4), 422–423. <https://doi.org/10.1056/nejmc055540>
- Leta, V., Urso, D., Batzu, L., Lau, Y. H., Mathew, D., Boura, I., Raeder, V., Falup-Pecurariu, C., van Wamelen, D., & Ray Chaudhuri, K. (2022). Viruses, parkinsonism and Parkinson's disease: the past, present and future. In *Journal of Neural Transmission* (Vol. 129, Issue 9, pp. 1119–1132). Springer. <https://doi.org/10.1007/s00702-022-02536-y>
- Liaci, A. M., & Förster, F. (2021). Take me home, protein roads: Structural insights into signal peptide interactions during er translocation. In *International Journal of Molecular Sciences* (Vol. 22, Issue 21). MDPI. <https://doi.org/10.3390/ijms222111871>
- Liao Y, Smyth GK, Shi W. (2017) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30;(7):923–30. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24227677>
- Lin, J. H., Walter, P., & Yen, T. S. B. (2008). Endoplasmic reticulum stress in disease pathogenesis. *Annual Review of Pathology: Mechanisms of Disease*, 3, 399–425. <https://doi.org/10.1146/annurev.pathmechdis.3.121806.151434>
- Lindström, R., Lindholm, P., Kallijärvi, J., Palgi, M., Saarna, M., & Heino, T. I. (2016). Exploring the conserved role of *MANF* in the unfolded protein response in *Drosophila melanogaster*. *PLoS ONE*, 11(3). <https://doi.org/10.1371/journal.pone.0151550>
- Lippi, A., Domingues, R., Setz, C., Outeiro, T. F., & Krisko, A. (2020). SARS-CoV-2: At the Crossroad Between Aging and Neurodegeneration. *Movement Disorders*, 35(5), 716–720. <https://doi.org/10.1002/mds.28084>
- Liu, Y., Zhang, J., Jiang, M., Cai, Q., Fang, J., & Jin, L. (2018). *MANF* improves the MPP+/MPTP-induced parkinson's disease via improvement of mitochondrial function and inhibition of oxidative stress. *American Journal of Translational Research*, 10(5), 1284–1294.
- Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J.,

- Magazine, H., Syron, J., ... Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6), 580–585. <https://doi.org/10.1038/ng.2653>
- Lunati, A., Lesage, S., & Brice, A. (2018). The genetic landscape of Parkinson's disease. *Revue Neurologique*, 174(9), 628–643. <https://doi.org/10.1016/j.neurol.2018.08.004>
- Mahmood, K., Jung, C. H., Philip, G., Georgeson, P., Chung, J., Pope, B. J., & Park, D. J. (2017). Variant effect prediction tools assessed using independent, functional assay-based datasets: Implications for discovery and diagnostics. *Human Genomics*, 11(1). <https://doi.org/10.1186/s40246-017-0104-8>
- Mahne, A. C., Carr, J. A., Bardien, S., & Schutte, C. M. (2016). Clinical findings and genetic screening for copy number variation mutations in a cohort of South African patients with Parkinson's disease. *South African Medical Journal*, 106(6), 623–625. <https://doi.org/10.7196/SAMJ.2016.v106i6.10340>
- Mahungu, A. C., Anderson, D. G., Rossouw, A. C., Coller, R. Van, Carr, J. A., Ross, O. A., Bardien, S., Sciences, H., Town, C., Africa, S., Africa, S., Sisulu, W., London, E., Africa, S., Africa, S., Sciences, H., Africa, S., & Clinic, M. (2021). *HHS Public Access*. 1–10. <https://doi.org/10.1016/j.neurobiolaging.2019.12.011.Screening>
- Mantere, T., Kersten, S., & Hoischen, A. (2019). Long-read sequencing emerging in medical genetics. In *Frontiers in Genetics* (Vol. 10, Issue MAY). Frontiers Media S.A. <https://doi.org/10.3389/fgene.2019.00426>
- McKusick, V.A. (1998). Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders. Baltimore: Johns Hopkins University Press, 12th edition).
- McNaught, K. S. P., & Jenner, P. (2001). Proteasomal function is impaired in substantia nigra in Parkinson's disease. *Neuroscience Letters*, 297(3), 191–194. [https://doi.org/10.1016/S0304-3940\(00\)01701-8](https://doi.org/10.1016/S0304-3940(00)01701-8)
- McNaught, K. S., Olanow, C. W., Halliwell, B., Isacson, O., & Jenner, P. (2001). Failure of the ubiquitin-proteasome system in Parkinson's disease. *Nature reviews. Neuroscience*, 2(8), 589–594. <https://doi.org/10.1038/35086067>
- Mercado, G., Castillo, V., Soto, P., & Sidhu, A. (2016). ER stress and Parkinson's disease: Pathological inputs that converge into the secretory pathway. *Brain research*, 1648(Pt B), 626–632. <https://doi.org/10.1016/j.brainres.2016.04.042>

- Milanowski LM, Oshinaike O, Broadway BJ, Lindemann JA, Soto-Beasley AI, Walton RL, Hanna Al-Shaikh R, Strongosky AJ, Fiesel FC, Ross OA, Springer W, Ogun SA, Wszolek ZK. Early-Onset Parkinson Disease Screening in Patients From Nigeria. *Front Neurol*. 2021 Jan 14;11:594927. doi: 10.3389/fneur.2020.594927. PMID: 33519679; PMCID: PMC7841006.
- Mirzoyan, Z., Sollazzo, M., Allocca, M., Valenza, A. M., Grifoni, D., & Bellostà, P. (2019). *Drosophila melanogaster*: A model organism to study cancer. In *Frontiers in Genetics* (Vol. 10). Frontiers Media S.A. <https://doi.org/10.3389/fgene.2019.00051>
- Mokaya, J., Gray, K., & Carr, J. (n.d.). *Beliefs, knowledge and attitudes towards Parkinson's disease among a Xhosa speaking black population in South Africa: A cross-sectional study*.
- Mulder, N. J., Adebisi, E., Adebisi, M., Adeyemi, S., Ahmed, A., Ahmed, R., *et al.* (2017). Development of Bioinformatics Infrastructure for Genomics Research. *gh* 12 (2), 91–98. <https://doi:10.1016/j.gheart.2017.01.005>
- Muzzey, D., Evans, E. A., & Lieber, C. (2015). Understanding the Basics of NGS: From Mechanism to Variant Calling. *Current Genetic Medicine Reports*, 3(4), 158–165. <https://doi.org/10.1007/s40142-015-0076-8>
- Nalls, M. A., Blauwendraat, C., Vallerga, C. L., Heilbron, K., Bandres-Ciga, S., Chang, D., *et al.* (2019). International Parkinson's Disease Genomics Consortium Identification of Novel Risk Loci, Causal Insights, and Heritable Risk for Parkinson's Disease: a Meta-Analysis of Genome-wide Association Studies. *The Lancet. Neurology* 18 (12), 1091–1102. [https://doi:10.1016/S1474-4422\(19\)30320-5](https://doi:10.1016/S1474-4422(19)30320-5)
- Newman, J. L. (1995). *The peopling of Africa: a geographic interpretation*. Yale University Press.
- Ng, P. C., & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Research*, 11(5), 863–874. <https://doi.org/10.1101/gr.176601>
- Oertel, W., & Schulz, J. B. (2016). Current and experimental treatments of Parkinson disease: A guide for neuroscientists. *Journal of Neurochemistry*, 139, 325–337. <https://doi.org/10.1111/jnc.13750>
- Okubadejo, N. U., Ojo, O. O., Wahab, K. W., Abubakar, S. A., Obiabo, O. Y., Salawu, F. K., Nwazor, E. O., Agabi, O. P., & Oshinaike, O. O. (2018). A Nationwide Survey of Parkinson's Disease Medicines Availability and Affordability in Nigeria. *Movement disorders clinical practice*, 6(1), 27–33. <https://doi.org/10.1002/mdc3.12682>
- Olgati, S., Quadri, M., Fang, M., Rood, J. P. M. A., Saute, J. A., Chien, H. F., Bouwkamp, C. G., Graafland, J., Minneboo, M., Breedveld, G. J., Zhang, J., Verheijen, F. W., Boon, A. J. W., Kievit, A. J. A., Jardim, L. B., Mandemakers, W., Barbosa, E. R., Rieder, C. R. M., Leenders, K.

- L., Wang, J., and Bonifati, V. International Parkinsonism Genetics Network (2016). D NAJC 6 Mutations Associated with Early-Onset Parkinson's Disease. *Ann. Neurol.* 79 (2), 244–256. <https://doi.org/10.1002/ana.24553>
- Oliver, G. R., Hart, S. N., & Klee, E. W. (2015). Bioinformatics for clinical next-generation sequencing. *Clinical Chemistry*, 61(1), 124–135. <https://doi.org/10.1373/clinchem.2014.224360>
- Oluwole, O. G. (2019). *Implementation of Targeted Resequencing Strategies to Identify Pathogenic Mutations in Nigerian and South African Patients with Parkinson's Disease. April.*
- Pakarinen, E., Danilova, T., Võikar, V., Chmielarz, P., Piepponen, P., Airavaara, M., Saarma, M., & Lindahl, M. (2020). *MANF* ablation causes prolonged activation of the UPR without neurodegeneration in the mouse midbrain dopamine system. *ENeuro*, 7(1). <https://doi.org/10.1523/ENEURO.0477-19.2019>
- Palgi, M., Lindström, R., Peränen, J., Piepponen, T. P., Saarma, M., & Heino, T. I. (2009). Evidence that Dm*MANF* is an invertebrate neurotrophic factor supporting dopaminergic neurons. *Proceedings of the National Academy of Sciences of the United States of America*, 106(7), 2429–2434. <https://doi.org/10.1073/pnas.0810996106>
- Pan, B., Kusko, R., Xiao, W., Zheng, Y., Liu, Z., Xiao, C., *et al.* (2019). Similarities and Differences between Variants Called with Human Reference Genome HG19 or HG38. *BMC Bioinformatics* 20, 101. <https://doi.org/10.1186/s12859-019-2620-0>
- Pandey, U. B., & Nichols, C. D. (2011). Human disease models in drosophila melanogaster and the role of the fly in therapeutic drug discovery. *Pharmacological Reviews*, 63(2), 411–436. <https://doi.org/10.1124/pr.110.003293>
- Pang, S. Y. Y., Teo, K. C., Hsu, J. S., Chang, R. S. K., Li, M., Sham, P. C., & Ho, S. L. (2017). The role of gene variants in the pathogenesis of neurodegenerative disorders as revealed by next-generation sequencing studies: A review. In *Translational Neurodegeneration* (Vol. 6, Issue 1). BioMed Central Ltd. <https://doi.org/10.1186/s40035-017-0098-0>
- Parad, R. B., & Comeau, A. M. (2005). Diagnostic dilemmas resulting from the immunoreactive trypsinogen/DNA cystic fibrosis newborn screening algorithm. *Journal of Pediatrics*, 147(3 SUPPL.), 1160–1167. <https://doi.org/10.1016/j.jpeds.2005.08.017>
- Park, H., Bradley, P., Greisen, P. Jr., Liu, Y., Mulligan, V. K., Kim, D. E., *et al.* (2016). Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J. Chem. Theory Comput.* 12, 6201–6212. doi: 10.1021/acs.jctc.6b00819

- Park, S. T., & Kim, J. (2016). Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing. *International neurourology journal*, 20(Suppl 2), S76–S83. <https://doi.org/10.5213/inj.1632742.371>
- Park, S. T., & Kim, J. (2016). Trends in next-generation sequencing and a new era for whole genome sequencing. *International Neurourology Journal*, 20, 76–83. <https://doi.org/10.5213/inj.1632742.371>
- Pearce, R., Li, Y., Omenn, G. S., & Zhang, Y. (2022). Fast and accurate Ab Initio Protein structure prediction using deep learning potentials. *PLoS computational biology*, 18(9), e1010539. <https://doi.org/10.1371/journal.pcbi.1010539>
- Pereira, R., Oliveira, J., & Sousa, M. (2020). Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics. *Journal of Clinical Medicine*, 9(1), 132. <https://doi.org/10.3390/jcm9010132>
- Petersen, B. S., Fredrich, B., Hoepfner, M. P., Ellinghaus, D., & Franke, A. (2017). Opportunities and challenges of whole-genome and -exome sequencing. *BMC Genetics*, 18(1), 1–13. <https://doi.org/10.1186/s12863-017-0479-5>
- Petrova, P. S., Raibekas, A., Pevsner, J., Vigo, N., Anafi, M., Moore, M. K., Peaire, A., Shridhar, V., Smith, D. I., Kelly, J., Durocher, Y., & Commissiong, J. W. (2004). Discovering novel phenotype-selective neurotrophic factors to treat neurodegenerative diseases. *Progress in Brain Research*, 146, 167–183. [https://doi.org/10.1016/S0079-6123\(03\)46012-3](https://doi.org/10.1016/S0079-6123(03)46012-3)
- Pidasheva, S., Canaff, L., Simonds, W. F., Marx, S. J., & Hendy, G. N. (2005). Impaired cotranslational processing of the calcium-sensing receptor due to signal peptide missense mutations in familial hypocalciuric hypercalcemia. *Human Molecular Genetics*, 14(12), 1679–1690. <https://doi.org/10.1093/hmg/ddi176>
- Pillay, N. S., Ross, O. A., Christoffels, A., & Bardien, S. (2022). Current Status of Next-Generation Sequencing Approaches for Candidate Gene Discovery in Familial Parkinson’s Disease. *Frontiers in Genetics*, 13. <https://doi.org/10.3389/fgene.2022.781816>
- Pires, D. E. V., Blundell, T. L., & Ascher, D. B. (2016). MCSM-lig: Quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Scientific Reports*, 6. <https://doi.org/10.1038/srep29575>

- Pontén, F., Jirström, K., & Uhlen, M. (2008). The Human Protein Atlas - A tool for pathology. In *Journal of Pathology* (Vol. 216, Issue 4, pp. 387–393). <https://doi.org/10.1002/path.2440>
- Popp, M. W., & Maquat, L. E. (2013). Organizing principles of mammalian nonsense-mediated mRNA decay. *Annual review of genetics*, 47, 139–165. <https://doi.org/10.1146/annurev-genet-111212-133424>
- Puschmann, A. (2017). New Genes Causing Hereditary Parkinson's Disease or Parkinsonism. *Curr. Neurol. Neurosci. Rep.* 17, 66. <https://doi:10.1007/s11910-017-0780-8>
- Quadri, M., Mandemakers, W., Grochowska, M. M., Masius, R., Geut, H., Fabrizio, E., Breedveld, G. J., Kuipers, D., Minneboo, M., Vergouw, L. J. M., Carreras Mascaro, A., Yonova-Doing, E., Simons, E., Zhao, T., Di Fonzo, A. B., Chang, H. C., Parchi, P., Melis, M., Correia Guedes, L., ... Bonifati, V. (2018). LRP10 genetic variants in familial Parkinson's disease and dementia with Lewy bodies: a genome-wide linkage and sequencing study. *The Lancet Neurology*, 17(7), 597–608. [https://doi.org/10.1016/S1474-4422\(18\)30179-0](https://doi.org/10.1016/S1474-4422(18)30179-0)
- Rajpar, M. H., Koch, M. J., Davies, R. M., Mellody, K. T., Kielty, C. M., & Dixon, M. J. (n.d.). *Mutation of the signal peptide region of the bicistronic gene DSPP affects translocation to the endoplasmic reticulum and results in defective dentine biomineralization.* <https://academic.oup.com/hmg/article/11/21/2559/551982>
- Ray Dorsey, E., Elbaz, A., Nichols, E., Abd-Allah, F., Abdelalim, A., Adsuar, J. C., Ansha, M. G., Brayne, C., Choi, J. Y. J., Collado-Mateo, D., Dahodwala, N., Do, H. P., Edessa, D., Endres, M., Fereshtehnejad, S. M., Foreman, K. J., Gankpe, F. G., Gupta, R., Hankey, G. J., ... Murray, C. J. L. (2018). Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*, 17(11), 939–953. [https://doi.org/10.1016/S1474-4422\(18\)30295-3](https://doi.org/10.1016/S1474-4422(18)30295-3)
- Reale, C., Panteghini, C., Carecchio, M., & Garavaglia, B. (2018). The relevance of gene panels in movement disorders diagnosis: A lab perspective. *European Journal of Paediatric Neurology*, 22(2), 285–291. <https://doi.org/10.1016/j.ejpn.2018.01.013>
- Reed, X., Bandrés-Ciga, S., Blauwendraat, C., & Cookson, M. R. (2019). The role of monogenic genes in idiopathic Parkinson's disease. *Neurobiology of Disease*, 124, 230–239. <https://doi.org/10.1016/j.nbd.2018.11.012>
- Reekes, T. H., Higginson, C. I., Ledbetter, C. R., Sathivadivel, N., Zweig, R. M., & Disbrow, E. A. (2020). Sex specific cognitive differences in Parkinson disease. *Npj Parkinson's Disease*, 6(1), 1–6. <https://doi.org/10.1038/s41531-020-0109-1>

- Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47(D1), D886–D894. <https://doi.org/10.1093/nar/gky1016>
- Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Research*, 39(17). <https://doi.org/10.1093/nar/gkr407>
- Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5), 405–424. <https://doi.org/10.1038/gim.2015.30>
- Richman, C., Rashid, S., Prashar, S., Mishra, R., Selvaganapathy, P. R., & Gupta, B. P. (2018). C. elegans *MANF* Homolog Is Necessary for the Protection of Dopaminergic Neurons and ER Unfolded Protein Response. *Frontiers in Neuroscience*, 12(AUG), 1–11. <https://doi.org/10.3389/fnins.2018.00544>
- Rizig, M., Ojo, O. O., Athanasiou-Fragkouli, A., Agabi, O. P., Oshinaike, O. O., Houlden, H., & Okubadejo, N. U. (2021). Negative screening for 12 rare LRRK2 pathogenic variants in a cohort of Nigerians with Parkinson's disease. *Neurobiology of aging*, 99, 101.e15–101.e19. <https://doi.org/10.1016/j.neurobiolaging.2020.09.024>
- Rogers, M. F., Shihab, H. A., Mort, M., Cooper, D. N., Gaunt, T. R., & Campbell, C. (2018). fathmm-XF: Accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*, 34(3), 511–513. <https://doi.org/10.1093/bioinformatics/btx536>
- Ruiz-Martínez, J., Azcona, L. J., Bergareche, A., Martí-Massó, J. F., and PaisánRuiz, C. (2017). Whole-exome Sequencing Associates Novel CSMD1 Gene Mutations with Familial Parkinson Disease. *Neurol. Genet.* 3 (5), e177. <https://doi:10.1212/NXG.0000000000000177>
- Ruiz-Martínez, J., Azcona, L. J., Bergareche, A., Martí-Massó, J. F., & Paisán-Ruiz, C. (2017). Whole-exome sequencing associates novel CSMD1 gene mutations with familial Parkinson disease. *Neurology: Genetics*, 3(5). <https://doi.org/10.1212/NXG.0000000000000177>
- Saier, M. H. (2019). *2019_20_NEB_Catalog_TechnicalReference.pdf*. July, 1–12.
- Salfati, E. L., Spencer, E. G., Topol, S. E., Muse, E. D., Rueda, M., Lucas, J. R., Wagner, G. N., Campman, S., Topol, E. J., & Torkamani, A. (2019). Re-analysis of whole-exome sequencing

data uncovers novel diagnostic variants and improves molecular diagnostic yields for sudden death and idiopathic diseases. *Genome Medicine*, 11(1). <https://doi.org/10.1186/s13073-019-0702-2>

Sanchez, G. (2013). Las instituciones de ciencia y tecnología en los procesos de aprendizaje de la producción agroalimentaria en Argentina. *El Sistema Argentino de Innovación: Instituciones, Empresas y Redes. El Desafío de La Creación y Apropiación de Conocimiento.*, 14(October 2006), 659–664. <https://doi.org/10.1002/prot>

Schneider, S. A., & Alcalay, R. N. (2020). Precision medicine in Parkinson's disease: emerging treatments for genetic Parkinson's disease. In *Journal of Neurology* (Vol. 267, Issue 3, pp. 860–869). Springer. <https://doi.org/10.1007/s00415-020-09705-7>

Schoonen, M., Seyffert, A. S., van der Westhuizen, F. H., & Smuts, I. (2019). A bioinformatics pipeline for rare genetic diseases in South African patients. *South African Journal of Science*, 115(3–4), 3–5. <https://doi.org/10.17159/sajs.2019/4876>

Schormair, B., Kemlink, D., Mollenhauer, B., Fiala, O., Machetanz, G., Roth, J., et al. (2018). Diagnostic Exome Sequencing in Early-Onset Parkinson's Disease confirms VPS13C as a Rare Cause of Autosomal-Recessive Parkinson's Disease. *Clin. Genet.* 93 (3), 603–612. <https://doi:10.1111/cge.13124>

Schubert, M., Lindgreen, S., & Orlando, L. (2019). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC research notes*, 9, 88. <https://doi.org/10.1186/s13104-016-1900-2>

Schutz, S., Monod-Broca, C., & Denommé-Pichon, A.-S. (2021). CutePeaks: A modern viewer for Sanger trace file. *Journal of Open Source Software*, 6(64), 3457. <https://doi.org/10.21105/joss.03457>

Sedova, A., Eblen, J. D., Budiardja, R., Tharrington, A., & Smith, J. C. (n.d.). *High-Performance Molecular Dynamics Simulation for Biological and Materials Sciences: Challenges of Performance Portability*. <http://energy.gov/>

Selvaraj, S., & Piramanayagam, S. (2019). Impact of gene mutation in the development of Parkinson's disease. *Genes and Diseases*, 6(2), 120–128. <https://doi.org/10.1016/j.gendis.2019.01.004>

Shademan, B., Biray Avci, C., Nikanfar, M., & Nourazarian, A. (2021). Application of Next-Generation Sequencing in Neurodegenerative Diseases: Opportunities and Challenges. In

NeuroMolecular Medicine (Vol. 23, Issue 2, pp. 225–235). Springer.
<https://doi.org/10.1007/s12017-020-08601-7>

Shantanam, S., & MUELLER. (2018). 乳鼠心肌提取 HHS Public Access. *Physiology & Behavior*, 176(1), 139–148. <https://doi.org/10.1038/ng.2892.A>

Shulskaya, M. V., Alieva, A. K., Vlasov, I. N., Zyrin, V. V., Fedotova, E. Y., Abramycheva, N. Y., Usenko, T. S., Yakimovsky, A. F., Emelyanov, A. K., Pchelina, S. N., Illarioshkin, S. N., Slominsky, P. A., & Shadrina, M. I. (2018). Whole-exome sequencing in searching for new variants associated with the development of Parkinson's disease. *Frontiers in Aging Neuroscience*, 10(MAY), 1–8. <https://doi.org/10.3389/fnagi.2018.00136>

Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The Missing Diversity in Human Genetic Studies. In *Cell* (Vol. 177, Issue 1, pp. 26–31). Cell Press.
<https://doi.org/10.1016/j.cell.2019.02.048>

Sosnay PR, Siklosi KR, Van Goor F, Kaniecki K, Yu H, Sharma N, Ramalho AS, Amaral MD, Dorfman R, Zielenski J, Masica DL, Karchin R, Millen L, Thomas PJ, Patrinos GP, Corey M, Lewis MH, Rommens JM, Castellani C, Penland CM, Cutting GR. Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nat Genet*. 2013 Oct;45(10):1160-7. doi: 10.1038/ng.2745. Epub 2013 Aug 25. PMID: 23974870; PMCID: PMC3874936.

Steinhaus, R., Proft, S., Schuelke, M., Schwarz, J. M., Seelow, D., & Cooper, D. N. (2021). *MutationTaster2021*. 49(April), 446–451.

Straniero, L., Guella, I., Cilia, R., Parkkinen, L., Rimoldi, V., Young, A., Asselta, R., Soldà, G., Sossi, V., Stoessl, A. J., Priori, A., Nishioka, K., Hattori, N., Follett, J., Rajput, A., Blau, N., Pezzoli, G., Farrer, M. J., Goldwurm, S., ... Duga, S. (n.d.). *DNAJC12 and dopa-responsive non-progressive Parkinsonism*.

Su, Z., Ning, B., Fang, H., Hong, H., Perkins, R., Tong, W., & Shi, L. (2011). Next-generation sequencing and its applications in molecular diagnostics. *Expert review of molecular diagnostics*, 11(3), 333–343. <https://doi.org/10.1586/erm.11.3>

Sudhaman, S., Muthane, U. B., Behari, M., Govindappa, S. T., Juyal, R. C., & Thelma, B. K. (2016). Evidence of mutations in RIC3 acetylcholine receptor chaperone as a novel cause of autosomal-dominant Parkinson's disease with non-motor phenotypes. *Journal of Medical Genetics*, 53(8), 559–566. <https://doi.org/10.1136/jmedgenet-2015-103616>

- Surmeier, D. J. (2018). Determinants of dopaminergic neuron loss in Parkinson's disease. In *FEBS Journal* (Vol. 285, Issue 19, pp. 3657–3668). Blackwell Publishing Ltd. <https://doi.org/10.1111/febs.14607>
- Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., Jensen, L. J., & von Mering, C. (2021). The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1), D605–D612. <https://doi.org/10.1093/nar/gkaa1074>
- Takahashi, R. (2004). Parkin and Endoplasmic Reticulum Stress. *Biotherapy*, 18(2), 127–133.
- Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. In *Nature Reviews Genetics* (Vol. 20, Issue 8, pp. 467–484). Nature Publishing Group. <https://doi.org/10.1038/s41576-019-0127-1>
- Teilum, K., Olsen, J. G., & Kragelund, B. B. (2011). Protein stability, flexibility and function. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1814(8), 969–976. <https://doi.org/10.1016/j.bbapap.2010.11.005>
- Tikhonova, E. B., Karamysheva, Z. N., von Heijne, G., & Karamyshev, A. L. (2019). Silencing of Aberrant Secretory Protein Expression by Disease-Associated Mutations. *Journal of Molecular Biology*, 431(14), 2567–2580. <https://doi.org/10.1016/j.jmb.2019.05.011>
- Triarhou LC. (2013). Dopamine and Parkinson's Disease. In: Madame Curie Bioscience Database [Internet]. Austin (TX): Landes Bioscience
- Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Research*, 40(15). <https://doi.org/10.1093/nar/gks596>
- van der Kamp, M. W., & Daggett, V. (2009). The consequences of pathogenic mutations to the human prion protein. *Protein engineering, design & selection : PEDS*, 22(8), 461–468. <https://doi.org/10.1093/protein/gzp039>
- van der Merwe, C., Carr, J., Glanzmann, B., & Bardien, S. (2016). Exonic rearrangements in the known Parkinson's disease-causing genes are a rare cause of the disease in South African patients. *Neuroscience letters*, 619, 168–171. <https://doi.org/10.1016/j.neulet.2016.03.028>
- van Rensburg, Z. J., Abrahams, S., Chetty, D., Step, K., Acker, D., Lombard, C. J., Elbaz, A., Carr, J., & Bardien, S. (2022). The South African Parkinson's Disease Study Collection. *Movement*

disorders : official journal of the Movement Disorder Society, 37(1), 230–232.
<https://doi.org/10.1002/mds.28828>

Voorhees, R. M., & Hegde, R. S. (2016). Toward a structural understanding of co-translational protein translocation. *Current Opinion in Cell Biology*, 41, 91–99.
<https://doi.org/10.1016/j.ceb.2016.04.009>

Voutilainen, M. H., Bäck, S., Pörsti, E., Toppinen, L., Lindgren, L., Lindholm, P., Peränen, J., Saarma, M., & Tuominen, R. K. (2009). Mesencephalic astrocyte-derived neurotrophic factor is neurorestorative in rat model of Parkinson's disease. *Journal of Neuroscience*, 29(30), 9651–9659. <https://doi.org/10.1523/JNEUROSCI.0833-09.2009>

Walusinski O. (2018). Jean-Martin Charcot and Parkinson's disease: Teaching and teaching materials. *Revue neurologique*, 174(7-8), 491–505. <https://doi.org/10.1016/j.neurol.2017.08.005>

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., De Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46(W1), W296–W303. <https://doi.org/10.1093/nar/gky427>

Wesselman, T., Harbin, L. P., Wolf, B. J., & Chung, D. (2018). *Opportunities in the Research Setting*. 17(3), 225–237. <https://doi.org/10.1080/14737159.2017.1282822>. *Genomics*

Williams, U., Bandmann, O., & Walker, R. (2018). Parkinson's Disease in Sub-Saharan Africa: A Review of Epidemiology, Genetics and Access to Care. *Journal of Movement Disorders*, 11(2), 53–64. <https://doi.org/10.14802/jmd.17028>

Worth, C.L., Blundell, T.L. (2010) On the evolutionary conservation of hydrogen bonds made by buried polar amino acids: the hidden joists, braces and trusses of protein architecture. *BMC Evol Biol* 10, 161. <https://doi.org/10.1186/1471-2148-10-161>

Yang, H., & Wang, K. (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nature protocols*, 10(10), 1556–1566. <https://doi.org/10.1038/nprot.2015.105>

Yu, Y., Liu, D. Y., Chen, X. S., Zhu, L., & Wan, L. H. (2021). *MANF* : A Novel Endoplasmic Reticulum Stress Response Protein - The Role in Neurological and Metabolic Disorders. *Oxidative Medicine and Cellular Longevity*, 2021. <https://doi.org/10.1155/2021/6467679>

Zamenhof, S. (1963). Mutations. *The American Journal of Medicine*, 34(5), 609–626.
[https://doi.org/10.1016/0002-9343\(63\)90102-5](https://doi.org/10.1016/0002-9343(63)90102-5)

Zhang, C., Shine, M., Pyle, A. M., & Zhang, Y. (n.d.). *US-align: Universal Structure Alignments of Proteins, Nucleic Acids, and Macromolecular Complexes*. <https://doi.org/10.1101/2022.04.18.488565>

Zhang, D., Jiang, H., & Xie, J. (2014). Alcohol intake and risk of Parkinson's disease: A meta-analysis of observational studies. *Movement Disorders*, 29(6), 819–822. <https://doi.org/10.1002/mds.25863>

Zhang, Z., Shen, Y., Luo, H., Zhang, F., Peng, D., Jing, L., Wu, Y., Xia, X., Song, Y., Li, W., & Jin, L. (2018). *MANF* protects dopamine neurons and locomotion defects from a human α -synuclein induced Parkinson's disease model in *C. elegans* by regulating ER stress and autophagy pathways. *Experimental neurology*, 308, 59–71. <https://doi.org/10.1016/j.expneurol.2018.06.016>



Appendices

Appendix A: Ethics approval from Stellenbosch University



09/06/2020

Project ID: 7506

Ethics Reference No: 2002C/059

Project Title: Genetic analysis of inherited Parkinson's Disease and other related movement disorders

Dear Prof Jonathan Carr

We refer to your request for an extension/annual renewal of ethics approval dated 13/04/2020 15:12.

The Health Research Ethics Committee reviewed and approved the annual progress report through an expedited review process.

The approval of this project is extended for a further year.

Approval date: 14 May 2020

Expiry date: 13 May 2021

Kindly be reminded to submit progress reports two (2) months before expiry date.

Where to submit any documentation

Kindly note that the HREC uses an electronic ethics review management system, *Infonetica*, to manage ethics applications and ethics review process. To submit any documentation to HREC, please click on the following link: <https://applyethics.sun.ac.za>.

Please remember to use your Project Id 7506 and ethics reference number 2002C/059 on any documents or correspondence with the HREC concerning your research protocol.

Yours sincerely,

Mrs. Ashleen Fortuin
Health Research Ethics Committee 2 (HREC2)

National Health Research Ethics Council (NHREC) Registration Number:
REC-130408-012 (HREC1)•REC-230208-010 (HREC2)

Federal Wide Assurance Number: 00001372
Office of Human Research Protections (OHRP) Institutional Review Board (IRB) Number:
IRB0005240 (HREC1)•IRB0005239 (HREC2)

The Health Research Ethics Committee (HREC) complies with the SA National Health Act No. 61 of 2003 as it pertains to health research. The HREC abides by the ethical norms and principles for research, established by the World Medical Association (2013), Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects; the South African Department of Health (2006), Guidelines for Good Practice in the Conduct of Clinical Trials with Human Participants in South Africa (2nd edition); as well as the Department of Health (2015), Ethics in Health Research: Principles, Processes and Structures (2nd edition).

The Health Research Ethics Committee reviews research involving human subjects conducted or supported by the Department of Health and Human Services, or other federal departments or agencies that apply the Federal Policy for the Protection of Human Subjects to such research (United States Code of Federal Regulations Title 45 Part 46); and/or clinical investigations regulated by the Food and Drug Administration (FDA) of the Department of Health and Human Services.

Appendix B: Ethics approval from the University of the Western Cape



UNIVERSITY of the
WESTERN CAPE



12 August 2021

Ms NS Pillay
SANBI
Faculty of Natural Sciences

Ethics Reference Number: BM21/4/13

Project Title: Next generation sequencing approaches for novel gene discovery in South African Parkinson's disease families.

Approval Period: 11 August 2021 – 11 August 2024

I hereby certify that the Biomedical Science Research Ethics Committee of the University of the Western Cape approved the scientific methodology and ethics of the above mentioned research project.

Any amendments, extension or other modifications to the protocol must be submitted to the Ethics Committee for approval.

Please remember to submit a progress report annually by 30 November for the duration of the project.

Permission to conduct the study must be submitted to BMREC for record-keeping.

The Committee must be informed of any serious adverse event and/or termination of the study.

Ms Patricia Josias
Research Ethics Committee Officer
University of the Western Cape

Director: Research Development
University of the Western Cape
Private Bag X 17
Bellville 7535
Republic of South Africa
Tel: +27 21 959 4111
Email: research-ethics@uwc.ac.za

NHREC Registration Number: BMREC-130416-050

FROM HOPE TO ACTION THROUGH KNOWLEDGE.

Appendix C: Data Transfer Agreement between Stellenbosch University and the University of the Western Cape

DATA TRANSFER AGREEMENT

Between

STELLENBOSCH UNIVERSITY ("Provider")

Physical Address	R.W. Wilcocks Building 2037, Victoria Street, Stellenbosch, 7600, South Africa
Postal Address	Private Bag X1, Matieland, Stellenbosch, 7602, South Africa
Telefax Number	+27 (0)21 808 4537
Telephone Number	+27 (0)21 808 9358
Contact Persons: Principal Investigator Legal matters	Prof Soraya Bardien sbardien@sun.ac.za Mark Mulder mmulder@sun.ac.za
Signature who warrants that s/he is duly authorised to sign	
Name	Prof Nico Gey van Pittius
Position	Vice-Dean Research & Internationalisation, Faculty of Medicine & Health Science
Date	Sep 23, 2020

and

UNIVERSITY OF THE WESTERN CAPE ("Recipient")

Physical Address	West Wing, 4th Floor Administration Building Robert Sobukwe Road, Bellville, 7535 South Africa
Telephone Number	+27 (0)21 959 3794
Contact Person	Mr Shervaan Rajie
Email Address	srajie@uwc.ac.za
Signature who warrants that s/he is duly authorised to sign	
Name	Professor José Frantz
Position	Deputy Vice Chancellor of Research and Innovation
Date	15 September 2020

PROF ALAN CHRISTOFFELS ("Recipient Scientist")

Facility/Laboratory Address	South African National Bioinformatics Institute (SANBI), University of the Western Cape, Robert Sobukwe Road, Bellville, 7535, South Africa
Telephone Number	+27 (0)21 959 2969
Email Address	alan@sanbi.ac.za
Signature	
Name	Prof Alan Christoffels
Position	Director & DST/NRF Research Chair in Bioinformatics & Health Genomics, SA MRC Bioinformatics Unit
Date	

SU Data Transfer UWC 2020

3

Research Project title: Next Generation Sequencing Approaches for Novel Gene Discovery in South African Parkinson's Disease Families	SU Contract number: S006840
Research Period: 2020 - 2022 (3 years)	SU Ethical Clearance number: 2002/C059

Data type:

Human tissue or blood samples []; Cell components []; Plants or organisms []; Animals []; Genetically Modified Organisms [] **Other X** [Next-generation sequencing data from human DNA samples]

Preamble

Whereas RECIPIENT and PROVIDER are working in collaboration on the project known as "Next Generation Sequencing Approaches for Novel Gene Discovery in South African Parkinson's Disease Families" ("the Research Project");

Whereas, this Agreement will govern the transfer of data from PROVIDER to RECIPIENT in furtherance of the Research Study.

Now therefore, upon execution of this Agreement, PROVIDER and RECIPIENT agree to the following:

1. The Agreement applies to the transfer of the data or portions thereof (collectively, the "DATA", clause 20) for purposes of the Research Project (clause 21). The DATA is being made available by Prof Soraya Bardien ("Investigator") and resulted from research conducted by Investigator as an employee of PROVIDER.
2. The DATA provided by the PROVIDER will be de-identified. The RECIPIENT will not be provided with any information that could be used to identify the participants from whom the DATA was collected, although the PROVIDER may retain a confidential link to the participant's identity. Neither RECIPIENT, Recipient Scientist nor the RECIPIENTS other employees or scientists shall make any attempts to determine the identity of those participants, or to contact the participants.
3. The RECIPIENT agrees that the DATA is to be used solely for teaching and academic research purposes, will not be used in human subjects, in clinical trials, or for diagnostic purposes involving human subjects or for any profit-making or commercial purposes without the explicit written consent of the PROVIDER and is to be used only at the RECIPIENT facility/laboratory and for the furtherance of the Research Project. In the case of a conflict between the provisions of this clause 3 and the provisions of clause 21, the provisions of this clause 3 shall take precedence.
4. Neither Recipient Scientist nor RECIPIENT nor any other person authorized to use the DATA under the Agreement shall make available the DATA or any portion of the DATA to any person or entity other than research personnel under the Recipients Scientist's immediate and direct control. No person authorized to use the DATA shall be allowed to take or send the DATA to any location other than the Recipient Scientist's facility/laboratory address without PROVIDERS's prior written consent.
5. Recipient Scientist and RECIPIENT will use the DATA in compliance with all applicable laws, governmental regulations and guidelines, including without limitation any regulations or guidelines pertaining to research that may be applicable to the DATA.

SU Data Transfer UWC 2020



3

6. The Recipient Scientist and RECIPIENT agrees to use the DATA in an appropriate manner and in compliance with internationally accepted scientific best practice not necessarily embedded in legislation.
7. The PROVIDER warrants that it has obtained Institutional Review Board/Ethics Committee approval required for the transfer and use of this DATA in the Research Project.
8. Legal title to the DATA shall remain with the PROVIDER and nothing in the Agreement grants RECIPIENT any rights under any patents nor any rights to use the DATA or any product(s) or process(es) derived from or with the DATA for profit-making or commercial purposes. Except as otherwise provided in clause 16 of this Agreement, RECIPIENT shall maintain the confidentiality of PROVIDER'S proprietary information relating to the DATA. RECIPIENT will hold the DATA in custody solely for the purposes of the Research Project as set forth in this Agreement.
9. The RECIPIENT acknowledges that the DATA is or may be the subject of a patent application. Except as provided in this Agreement, no express or implied licenses or other rights are provided to the RECIPIENT under any patents, patent applications, trade secrets or other proprietary rights of PROVIDER, including any altered forms of the DATA made by RECIPIENT.
10. The transfer of the DATA constitutes a non-exclusive license to use the DATA solely for teaching and academic research purposes. As required in terms of clause 3 above, RECIPIENT agrees to negotiate in good faith a license with PROVIDER prior to making any profit-making or commercial use of the DATA or any product(s) or process(es) derived from or incorporating the DATA. PROVIDER shall have no obligation to grant such a license to RECIPIENT and may grant exclusive or non-exclusive licenses to others who may be investigating uses of the DATA.
11. It is the intent of the Parties to pursue joint publications with respect to all results as consistent with the standards of the International Committee of Medical Journal Editors. The Parties agree to work together in good faith to achieve timely publication of jointly developed Research Project results. The Parties recognise that a Party ("the publishing party") may wish to publish details of academic research in scientific journals or theses. The publishing party may do so provided that the other Party ("the non-publishing Party") is provided with a copy of any such draft manuscripts of written publications or oral presentations that include any part of the Study results for review and approval approximately ten (10) business days prior to the contemplated presentation or publication (which shall be reduced to five (5) business days for abstracts). The non-publishing party shall be entitled to propose amendments to such draft manuscripts or presentation which the publishing party shall duly consider. The Parties agree to clearly mention the other Party and the other Party's contact person in any such publication.
12. Recipient Scientist and RECIPIENT shall acknowledge PROVIDER as the source of the DATA in any publication of Research Project results.
13. Except as explicitly elsewhere stated in this Agreement, the DATA is provided without warranty of merchantability or fitness for a particular purpose or any other warranty, express or implied. PROVIDER makes no representation or warranty that the use of the DATA will not infringe any patent or other proprietary right.
14. In no event shall PROVIDER be liable for any use by Recipient Scientist or RECIPIENT of the DATA or for any loss, claim, damage, or liability, of any kind or nature that may arise from or in connection with this Agreement or the use of the DATA. The RECIPIENT hereby agrees to defend, indemnify and hold harmless PROVIDER and PROVIDER'S officers, agents, and employees from any liability, claim, loss or damage, costs, or judgments of whatsoever kind or nature arising out of the use or disposition of the DATA by the Recipient Scientist or the RECIPIENT.

SU Data Transfer UWC 2020



3

15. The DATA is provided without a fee to cover the preparation and distribution of the DATA requested by RECIPIENT and the transfer of the DATA shall not be considered a sale of the DATA to RECIPIENT.
16. Unless PROVIDER specifically authorizes in writing, RECIPIENT must:
- (a) not use PROVIDER's information and DATA, except for the Research Project;
 - (b) not analyze PROVIDER's DATA to determine the composition of DATA other than required for the Research Project;
 - (c) not measure the properties of PROVIDER's DATA, except as reasonably necessary to accomplish the purpose of the Research Project;
 - (d) not make PROVIDER's information or DATA in independent form available to others (including patent offices);
 - (e) not disclose PROVIDER's information and DATA to the UWC PhD student Ms. Nikita Pillay and/or its employees requiring access to achieve the purpose of the Research Project, provided those persons are subject to obligations no less restrictive than this Agreement;
 - (f) not dispose of any of PROVIDER's DATA and RECIPIENT's copies of PROVIDER's information, in any manner other than as directed by PROVIDER; and
 - (g) not file any patent, utility model or design application disclosing any of the PROVIDER's confidential information.
 - (h) report to PROVIDER any loss or unauthorized release of the DATA in breach of this Agreement within 24 hours of becoming aware of the loss or unauthorized release.

16.1. The obligations of clause 16 are binding for a period of 10 (ten) years.

16.2. The obligations of clause 16 do not apply to any portion of PROVIDER's information that RECIPIENT can prove:

- (a) was available to the public through no fault of RECIPIENT, or
- (b) RECIPIENT already possessed prior to receipt from PROVIDER, or
- (c) RECIPIENT acquired from a third party without obligation of confidence, or
- (d) was independently developed by or for RECIPIENT.

17. Moreover, RECIPIENT may comply with a court order compelling production of PROVIDER's information/DATA, but Recipient must give PROVIDER reasonable prior notice and use reasonable efforts to obtain confidential protection for that information.

18. This Agreement will terminate on completion of the Research Period. Upon the effective date of termination, RECIPIENT will discontinue their use of the DATA and will dispose of the DATA as directed by the PROVIDER, except that RECIPIENT shall not be required to destroy any DATA which has been created pursuant to automatic archiving and back-up procedures and cannot be reasonably deleted.

19. General

19.1 This Agreement shall come into force on the date on which it is signed by both parties and shall remain in force for the duration of the Research Period unless terminated earlier.

19.2 Either Party may terminate this Agreement forthwith by thirty (30) days prior notice of termination in writing.

19.2.1 upon termination of this Agreement, RECIPIENT's rights to use the DATA will cease and RECIPIENT will discontinue all use of the DATA;

SU Data Transfer UWC 2020



3

19.2.2 all other terms hereunder will continue unaffected. For the avoidance of doubt, surviving any termination or expiration of this Agreement (unless provided otherwise by PROVIDER), is the agreement by RECIPIENT that RECIPIENT shall not use the DATA for profit-making or commercial purposes.

19.3 Neither party shall assign or transfer any interest in this Agreement without prior written approval of the other party.

19.4 No amendment, consent or waiver of terms of this Agreement shall bind either party unless in writing and signed by all parties. Any such amendment, consent, or waiver shall be effective only in the specific instance and for the specific purpose given.

19.5 This Agreement embodies the entire agreement between the parties hereto and no provision of this Agreement may be changed except by the mutual written consent of the parties hereto.

19.6 This Agreement shall be governed by the South African Law and the South African Courts shall have exclusive jurisdiction to deal with any dispute which may arise out of or in connection with this Agreement.

20. The DATA

This agreement concerns the following DATA to be provided to the Recipient:

Description of Data
Raw data from whole exome sequencing and whole genome sequencing files of Parkinson's disease patients and their family members

21. The Research Project

The manner in which and the extent to which the DATA may be used by the Recipient are as follows:

The research project aims to identify novel Parkinson's disease-causing mutations in South African patients and will involve the following:

- 1) Raw next-generation sequencing data of South African patients and their family members will be generated by collaborators of the provider scientist.
- 2) This data will be provided to the Recipient Scientist who may share the DATA with the jointly-supervised UWC PhD student (Ms Nikita Pillay), subject to the provisions of this Agreement.
- 3) This raw data will be analysed by the PhD student and the recipient scientist.
- 4) Thereafter, all parties will be involved in interpretation of the findings, and the writing of manuscripts.

~ END ~

SU Data Transfer UWC 2020



3



UNIVERSITY *of the*
WESTERN CAPE

SU Data Transfer UWC 2020

A handwritten mark or signature, possibly a stylized letter 'A' or a similar symbol.

3






22092020 DTA S006840

Final Audit Report

2020-09-22

Created:	2020-09-22
By:	Aslam Arnolds (aslam@sun.ac.za)
Status:	Signed
Transaction ID:	CBJCHBCAABAAMJbLu5_op6hxAFRzt6icNp2htWA8qZ5S

"22092020 DTA S006840" History

-  Document created by Aslam Arnolds (aslam@sun.ac.za)
2020-09-22 - 6:54:40 AM GMT- IP address: 146.232.119.1
-  Document emailed to Prof NC Gey van Pittius (ngvp@sun.ac.za) for signature
2020-09-22 - 6:55:25 AM GMT
-  Email viewed by Prof NC Gey van Pittius (ngvp@sun.ac.za)
2020-09-22 - 10:20:19 PM GMT- IP address: 197.83.213.133
-  Document e-signed by Prof NC Gey van Pittius (ngvp@sun.ac.za)
Signature Date: 2020-09-22 - 10:21:51 PM GMT - Time Source: server- IP address: 197.83.213.133
-  Agreement completed.
2020-09-22 - 10:21:51 PM GMT





Current Status of Next-Generation Sequencing Approaches for Candidate Gene Discovery in Familial Parkinson's Disease

Nikita Simone Pillay¹, Owen A. Ross^{2,3}, Alan Christoffels^{1,4} and Soraya Bardien^{5,6*}

¹South African National Bioinformatics Institute (SANBI), South African Medical Research Council Bioinformatics Unit, University of the Western Cape, Bellville, South Africa, ²Department of Neuroscience, Mayo Clinic, Jacksonville, FL, United States, ³Department of Clinical Genomics, Mayo Clinic, Jacksonville, FL, United States, ⁴Africa Centres for Disease Control and Prevention, African Union Headquarters, Addis Ababa, Ethiopia, ⁵Division of Molecular Biology and Human Genetics, Department of Biomedical Sciences, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa, ⁶South African Medical Research Council/Stellenbosch University Genomics of Brain Disorders Research Unit, Cape Town, South Africa

OPEN ACCESS

Edited by:

Amir Hossein Saeidian,
Thomas Jefferson University,
United States

Reviewed by:

Rachita Yadav,
Massachusetts General Hospital and
Harvard Medical School, United States
Thomas Gasser,
University of Tübingen, Germany

*Correspondence:

Soraya Bardien
sbardien@sun.ac.za

Specialty section:

This article was submitted to
Neurogenomics,
a section of the journal
Frontiers in Genetics

Received: 23 September 2021

Accepted: 12 January 2022

Published: 01 March 2022

Citation:

Pillay NS, Ross OA, Christoffels A and
Bardien S (2022) Current Status of
Next-Generation Sequencing
Approaches for Candidate Gene
Discovery in Familial Parkinson's
Disease.
Front. Genet. 13:781816.
doi: 10.3389/fgene.2022.781816

Parkinson's disease is a neurodegenerative disorder with a heterogeneous genetic etiology. The advent of next-generation sequencing (NGS) technologies has aided novel gene discovery in several complex diseases, including PD. This Perspective article aimed to explore the use of NGS approaches to identify novel loci in familial PD, and to consider their current relevance. A total of 17 studies, spanning various populations (including Asian, Middle Eastern and European ancestry), were identified. All the studies used whole-exome sequencing (WES), with only one study incorporating both WES and whole-genome sequencing. It is worth noting how additional genetic analyses (including linkage analysis, haplotyping and homozygosity mapping) were incorporated to enhance the efficacy of some studies. Also, the use of consanguineous families and the specific search for *de novo* mutations appeared to facilitate the finding of causal mutations. Across the studies, similarities and differences in downstream analysis methods and the types of bioinformatic tools used, were observed. Although these studies serve as a practical guide for novel gene discovery in familial PD, these approaches have not significantly resolved the "missing heritability" of PD. We speculate that what is needed is the use of third-generation sequencing technologies to identify complex genomic rearrangements and new sequence variation, missed with existing methods. Additionally, the study of ancestrally diverse populations (in particular those of Black African ancestry), with the concomitant optimization and tailoring of sequencing and analytic workflows to these populations, are critical. Only then, will this pave the way for exciting new discoveries in the field.

Keywords: Parkinson's disease, next-generation sequencing, whole-exome sequencing, familial PD, african ancestry, bioinformatic pipelines, third-generation sequencing, diverse populations

INTRODUCTION

Over the past almost 2 decades, next-generation sequencing (NGS) approaches, with their high-throughput and rapid output, have accelerated novel gene discovery for several human diseases. In this Perspective article, we summarize, analyze and highlight the studies that identified new loci for Parkinson's disease (PD) using NGS strategies.

PD is a neurodegenerative disorder, typically presenting with bradykinesia, rigidity, resting tremor, postural instability, and various non-motor symptoms (Kalinderi et al., 2016). Approximately 90% of PD cases are considered sporadic; attributed to synergistic interactions between genetic, metabolic and environmental factors (Ball et al., 2019). The remaining 5–10% of cases are accounted for by familial PD, usually displaying a Mendelian mode of inheritance (Lesage and Brice, 2012; Hernandez et al., 2016). Positional cloning approaches have been used successfully to identify disease genes within large multi-incident PD kindreds (Hildebrandt and Omran, 1999). Linked regions of the genome that co-segregated with disease were then Sanger sequenced to identify the causal variant. PD genes identified using this approach have demonstrated autosomal dominant (AD-PD) (*SNCA*, *LRKK2*), autosomal recessive (AR-PD) (*PRKN*, *PINK1*, *DJI*) and X-linked (*RAB39B*) inheritance patterns (Bras and Singleton., 2011; Gasser, 2013; Bandres-Ciga et al., 2020).

Later, development of high-throughput genotyping techniques allowed for the rapid screening of single-nucleotide variants (SNVs) - that occur with moderate to high allele frequencies - in large case/control cohorts (Shulskaya et al., 2018). This resulted in the rise of genome-wide association studies (GWAS), and adoption of the common-disease-common-variant hypothesis, which has been responsible for the discovery of many PD-susceptibility loci (Hemminki et al., 2008; Nalls et al., 2019). Yet, it has also been postulated that the gaping 'missing heritability' in complex disorders such as PD, may be attributed to larger penetrant effects of less common variants i.e., the rare-variant-common-disease hypothesis (Gasser et al., 2011; El-Fishawy, 2013; Germer et al., 2019).

Next-Generation Sequencing in PD

NGS, in the form of whole-exome sequencing (WES), captures only the coding region; while whole-genome sequencing (WGS) sequences the entire genome including all non-coding regions (Fernandez-Marmiesse et al., 2017). When considering NGS for the study of genetic disorders, WES presents as the more suitable choice as most pathogenic mutations (80–85%), found to date, are exonic (Ku et al., 2016). WES is also cheaper, and less computationally intensive than WGS (Bonnetfond et al., 2010; Chakravorty and Hegde, 2017). However, WES can result in skewed coverage due to hybridization biases and incomplete target enrichment, making detection of copy number variation (CNV) challenging (Belkadi et al., 2015). Since CNVs encompassing complete exons (in *PRKN*, *PINK1* and *DJ-1*) or spanning multiple gene copies (*SNCA*) are a significant cause of PD, this is a notable limitation of WES in PD studies. Together,

these factors indicate that WGS may be more effective for identification of novel or rare genetic variants, particularly in complex diseases like PD.

Novel Gene Discovery in PD-Affected Families Using NGS

For our search, a comprehensive search string on NCBI's PubMed Central database “((((((parkinson's disease) AND NGS) AND familial) AND novel) AND candidate) AND gene)” was done on 13 May 2021. Abstracts were read to identify studies that specifically used NGS (either WES or WGS) approaches to identify potential novel genes in familial PD or parkinsonism. We did not exclude studies with a lack of evidence of pathogenicity, and this resulted in a total of 17 relevant studies. These studies and their approaches are summarized in **Table 1** and are discussed in chronological order below.

In 2011, Vilariño-Güell and others published their WES findings on two first degree cousins from an AD PD-affected Swiss family, announcing the discovery of the p.Asp620Asn mutation in *VPS35* (Vilariño-Güell et al., 2011). In a back-to-back publication, that same mutation in *VPS35* was also identified in an Austrian family (Zimprich et al., 2011). Their study made use of haplotyping and linkage analysis in conjunction with WES, allowing for the simultaneous identification of linkage regions and the subsequent filtering of variants based on their distance to the linkage regions. Thus, postulating a time-and cost-effective approach to exome sequencing for AD-PD (Bras and Singleton, 2011; Gialluisi et al., 2020). Furthermore, the same mutation was found in six unrelated PD individuals of varying ethnicity and observed in a sporadic PD case (Zimprich et al., 2011). With these findings in several independent PD families, *VPS35* is now considered a significant gene associated with AD-PD, though with still unresolved pathology. The successes observed in these two early studies sparked hope for the discovery of rare monogenic causal factors using NGS in PD families and subsequently, several similar studies ensued.

In 2012, the discovery of *DNAJC6*, linked to AR-juvenile parkinsonism in a consanguineous Palestinian family, was published (Edvardson et al., 2012). They performed SNP genotyping and homozygosity mapping (HM) analysis in conjunction with WES (Edvardson et al., 2012; Vahidnezhad et al., 2018). This approach potentially facilitates more rapid detection of a disease gene after WES (Kim et al., 2013). HM analysis allows for the identification of large, shared regions of homozygosity (where variants associated with AR disease genes are likely to be located) between affected family members (Wakeling et al., 2019). Therefore, HM could be beneficial for the identification of pathogenic mutations in AR-PD (Bras and Singleton, 2011). The following year, the same approach on a consanguineous Iranian family affected with early-onset PD (EO-PD) led to the discovery of a homozygous mutation in *SYNJ1* (Krebs et al., 2013). Also in 2013, the finding of a heterozygous p.Ser657Asn mutation in *PLXNA4* within a large German family, was published (Schulte et al., 2013).

TABLE 1 | List of published studies that identified novel Parkinson's disease loci using next-generation sequencing approaches.

Reference	Gene	Population	Pre-NGS screening approach used	Study Participants Screened (Sequencing platform used)	QC and Read Alignment Tools	Variant Calling Tools	Variant Annotation and <i>In Silico</i> Pathogenicity Prediction Tools	Variant Inclusion/Exclusion Criteria	Mutations Identified/ (Chromosome)
Vilarinho-Güell <i>et al.</i> (2011)	VPS35 (vacuolar protein sorting 35 ortholog)	Swiss family (Family A)	None	WES on a PD-affected pair of 1st degree cousins	SOAPaligner (read alignment to the human References genome - Hg18, build 36.1)	SOAPsnp (SNP calling)	Database of Genomic Variants v6 (determination of structural variants against CNVs)	<ul style="list-style-type: none"> Variants were excluded if <ul style="list-style-type: none"> - on the X chromosome - homozygous (autosomal-dominant inheritance of disease was assumed) - non-coding - synonymous - variants present in dbSNP v.130 Variants were subsequently genotyped in a multi-ethnic case-control series (4,326 patients and 3,309 controls) Confirmation <i>via</i> Sanger sequencing 	<ul style="list-style-type: none"> Homozygous c.1858G > A - p.Asp620Asn (16q11.2)
Zimprich <i>et al.</i> (2011)	VPS35 (vacuolar protein sorting 35 ortholog)	Austrian family	Haplotyping and linkage analysis (Merlin software)	WES on two PD-affected second cousins (Genome Analyzer Ix system (Illumina))	Burrows-Wheeler Aligner (BWA version 0.6.8) (read alignment to human References genome - Hg19)	SAMtools (v 0.1.7) – (SNVs and InDel calling)	PolyPhen2, SNAP and SIFT – (pathogenicity prediction)	<ul style="list-style-type: none"> Variants were excluded if <ul style="list-style-type: none"> - present in the 72 control exomes of non-PD patients - present in dbSNP131 and 1000-Genomes Project - had an average heterozygosity of more than 0.02 Variants were included if <ul style="list-style-type: none"> - heterozygous - non-synonymous 	<ul style="list-style-type: none"> Heterozygous c.1858G > A - p.Asp620Asn (16q11.2)
Edvardson <i>et al.</i> (2012)	DNAJC6 (DnaJ Heat Shock Protein Family (Hsp40) Member C6)	Palestinian family (two patients and their unaffected brother)	Homozygosity mapping and SNP genotyping in a consanguineous family (SNP genotyping using Affymetrix GeneChip Human Mapping 250 K Nsp Array)	WES on a single index patient (GAIIx, Illumina)	Burrows-Wheeler Aligner (BWA) (sequence reads were aligned to human References genome - hg18 (GRCh36)) Picard (marking of PCR duplicates)	Genome Analysis Toolkit (GATK) (variant calling)	ANNOVAR (variant annotation) SeattleSeq Annotation (GERP score) Polyphen, SIFT and Mutation taster (pathogenicity prediction) NHLBI Exome Sequencing Project website release Version: v.0.0.9 (mutation frequency in ethnically matched controls)	<ul style="list-style-type: none"> Variants were excluded if <ul style="list-style-type: none"> - present in dbSNP132, 1000-Genomes Project and in-house databases Variants were included if <ul style="list-style-type: none"> - non-synonymous - conservation score GERP >3 Confirmation <i>via</i> Sanger sequencing 	<ul style="list-style-type: none"> Homozygous c.801-2A > G (1p31.3)

(Continued on following page)

TABLE 1 | (Continued) List of published studies that identified novel Parkinson's disease loci using next-generation sequencing approaches.

Reference	Gene	Population	Pre-NGS screening approach used	Study Participants Screened (Sequencing platform used)	QC and Read Alignment Tools	Variant Calling Tools	Variant Annotation and <i>In Silico</i> Pathogenicity Prediction Tools	Variant Inclusion/Exclusion Criteria	Mutations Identified/ (Chromosome)
Krebs <i>et al.</i> (2013)	SYNJ1 (Sac1- like inositol phosphatase domain of polyphosphoinositide phosphatase synaptojanin 1)	Iranian family (healthy parents, who were first-degree relatives, as well as two affected, and three unaffected siblings)	Genome-wide SNP genotyping and homozygosity mapping was performed on a consanguineous PD family (HumanOmniExpress beadchips and HiScanSQ system, Illumina) Genome Studio program (genotyping quality assessment) PLINK (Homozygous segment identification) Illumina genome viewer (homozygous segment visualizer)	WES on two PD-affected siblings (HiSeq 2000, Illumina)	Burrows-Wheeler Aligner (BWA) tool (alignment of raw sequence reads to the human References genome - NCBI GRCh37) Genome Analysis Toolkit (GATK v1.5-16-g58245bf) (base-quality re-calibration and local realignment)	GATK Unified Genotyper tool (SNP/SNV/InDel calling)	AnnTools (variant annotation) MutPred, SNPs&GO, Mutalyzer, HomoloGene (NCBI) and Clustalw2 (pathogenicity prediction)	Variants were excluded if - present in dbSNP137, 1,000 Genomes Project and Exome Variant Server of the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project databases Variants were included if - located in exons or splice sites Confirmation <i>via</i> Sanger sequencing	Homozygous c.773G > A - p. Arg258Gln (21q22.11)
Schulte <i>et al.</i> (2013)	PLXNA4 (plexin A4)	German Family	Genotyping of the top ten candidate variants (KORA-AGE cohort using MALDI-TOF masspectrometry on the SequenomH platform Linkage analysis on 6 family members using oligonucleotide SNP arrays (500 K Illumina) MERLIN (Linkage analysis)	WES on 2 PD-affected second cousins. (Genome Analyzer Iix system (Illumina)	Burrows-Wheeler Aligner (BWA 0.5.8) (read alignment)	SAMtools (version 0.1.7) (SNV/InDel calling)	SIFT/PROVEAN, PolyPhen-2 and MutationTaster (pathogenicity prediction)	Variants were excluded if: observed in in-house exome database, dbSNP135, 1000-Genomes Project and NHLBI-ESP (EA only) databases with a minor allele frequency >1% Variants were included if - non-synonymous - exonic/coding - missense, nonsense, stoploss, splice site or frameshift variants Confirmation <i>via</i> Sanger sequencing	Heterozygous c.1970C > T - p.Ser657Asn (7q32.3)
Vilarino-Güell <i>et al.</i> (2014)	DNAJC13 (receptor-mediated endocytosis 8/ RME-8)	Canadian (Dutch-German-Russian Mennonite) family	None	WES on three PD-affected members (Agilent SureSelect 38 Mb Human All Exon Kit, Illumina Genome Analyzer)	Bowtie 12.70 and Burrows-Wheeler Aligner (BWA 0.5.9) (read alignment to human References)	SAMtools (variant calling)	SIFT (pathogenicity prediction)	Variants were excluded if	Homozygous c.2564A > G

(Continued on following page)

TABLE 1 | (Continued) List of published studies that identified novel Parkinson's disease loci using next-generation sequencing approaches.

Reference	Gene	Population	Pre-NGS screening approach used	Study Participants Screened (Sequencing platform used)	QC and Read Alignment Tools	Variant Calling Tools	Variant Annotation and <i>In Silico</i> Pathogenicity Prediction Tools	Variant Inclusion/Exclusion Criteria	Mutations Identified/ (Chromosome)
					genome - NCBI Build 37.1) Genome Analysis Toolkit (GATk) (local realignment around insertions and deletions)			- Phred quality score <20 - frequently observed in population databases (minor allele frequency >1%) Confirmation via Sanger sequencing	- p.Asn855Ser (3q22.1)
Funayama et al. (2015)	CHCHD2 (coiled-coil-helix-coiled-coil-helix domain containing 2)	Japanese family	Genome-wide linkage analysis on 8 affected and 5 unaffected individuals of the family (Genome-Wide Human SNP Array 6.0, Affymetrix) SNPHitLink & MERLIN (linkage analysis)	WES on three patients & WGS on one patient (HiSeq 2000, Illumina)	Burrows-Wheeler Aligner (BWA-MEM version 0.5.9) (read alignment to References human genome - UCSC hg19)	SAMtools version 0.1.16 (SNV/InDel calling)	PolyPhen-2 & MutationTaster (pathogenicity prediction)	Variants were excluded if - present in the 1,000 Genomes, dbSNP138, the Human Genetic Variation database, and the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP) database Variants were included if - located in exons or splice sites - heterozygous state - non-synonymous or caused aberrant splicing - located in regions with positive log of odds greater than 1 - not noted in unaffected Japanese controls Confirmation by Sanger sequencing	Heterozygous 182C > T - p.Thr61Ile (7p11.2)
Sudhaman et al., (2016b)	RIC3 (acetylcholine receptor chaperone)	South Indian family	None	WES on a single index patient (HiSeq 2000, Illumina)	FastXToolkit (pre-alignment QC) Burrows-Wheeler Aligner (BWA) (read alignment) SAMTools (Post-alignment QC) BEDTools (assess target coverage and depth)	SAMTools and GATk (variant calling)	wANNOVAR (variant annotation) KGGSeq (variant filtering)	Variants were excluded if - present in databases (dbSNP 135, 137 and 138, 1,000 genomes and National Heart, Lung, and Blood Institute (NHLBI) 6500 exomes and ExAC) with a MAF >0.01 Variants were included if - heterozygous	Homozygous c.169C > A - p.P57T (11p15.4)

(Continued on following page)

TABLE 1 | (Continued) List of published studies that identified novel Parkinson's disease loci using next-generation sequencing approaches.

Reference	Gene	Population	Pre-NGS screening approach used	Study Participants Screened (Sequencing platform used)	QC and Read Alignment Tools	Variant Calling Tools	Variant Annotation and <i>In Silico</i> Pathogenicity Prediction Tools	Variant Inclusion/Exclusion Criteria	Mutations Identified/ (Chromosome)
Sudhaman <i>et al.</i> , 2016a	PODXL (podocalyxin-like Gene)	North Indian family	None	WES on two affected siblings (HiSeq 2000, Illumina)	FastXToolkit (pre-alignment QC) Burrows-Wheeler Aligner (BWA) (read alignment) SAMTools (Post-alignment QC) BEDTools (assess target coverage and depth)	SAMTools and GATK (variant calling)	wANNOVAR (variant annotation) KGGSeq (variant filtering)	Confirmation <i>via</i> Sanger sequencing Variants were excluded if - present in databases (dbSNP 135, 137 and 138, 1,000 genomes and National Heart, Lung, and Blood Institute (NHLBI) 6500 exomes and ExAC) with a MAF >0.01 Variants were included if - homozygous (Autosomal recessive inheritance assumed) - exonic variants - shared between the two affected individuals Confirmation <i>via</i> PCR-Sanger sequencing	Homozygous c.89_90 insGTCGCCCC - p.Gln32fs (7q32.3)
Deng <i>et al.</i> (2016)	TMEM230 (Transmembrane Protein 230)	Canadian-Mennonite (same family as DNAJC13)	None	WES on one unaffected individual and 4 distantly related affected cousins (HiSeq2500, Illumina)	Genome Analysis Tool Kit (GATK v1.1) (read alignment to human References genome - HG19)	Unified Genotyper from the Genome Analysis Tool Kit (SNV/INDEL calling and performing variant quality score (VQS) and Phred-likelihood scores)	ANNOVAR (variant annotation) PolyPhen2 (pathogenicity prediction) SpliceView, MNSplice, and ESEfinder (splicing effect prediction)	Variants were excluded if - present in multiple databases including the dbSNP (v130), HapMap and 1,000 Genome databases with a MAF >0.01 - VQSLOD < -3 - alternate Phred-scaled likelihood scores <99 Variants were included if - the average read per targeted base was >65X with the Phred quality score of ≥30 Confirmation <i>via</i> Sanger sequencing and co-segregation analysis	Heterozygous c.422G > T - p.Arg141Leu (20p13-p12.3)
Ruiz-Martinez <i>et al.</i> (2017)	CSMD1 (CUB and Sushi multiple domains 1)	Spanish Basque family	None	WES on index patient (HiSeq 2000, Illumina)	Burrows-Wheeler Aligner Tool (BWA) (read alignment to the human References genome - NCBI GRCh37.p13)	GATK Unified Genotyper tool (SNP INDEL calling)	AnnTools kit (variant annotation) PICARD (Exome statistics) MutPred, SNPs&Go, MutationTaster, and CADD (pathogenicity prediction)	Variants were excluded if - intragenic, intronic, and non-coding exonic - present in the dbSNP149 build, 1,000 Genomes	Heterozygous c.5885G > A -p.Arg1962His and c.8959G.A- p.Gly2987Arg

(Continued on following page)

TABLE 1 | (Continued) List of published studies that identified novel Parkinson's disease loci using next-generation sequencing approaches.

Reference	Gene	Population	Pre-NGS screening approach used	Study Participants Screened (Sequencing platform used)	QC and Read Alignment Tools	Variant Calling Tools	Variant Annotation and <i>In Silico</i> Pathogenicity Prediction Tools	Variant Inclusion/Exclusion Criteria	Mutations Identified/ (Chromosome)
					Genome Analysis		HomoloGene database (protein conservation across species)	Project phase 3, the Exome Variant Server of the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing and the Exome Aggregation Consortium databases with a MAF >0.05 Variants were included if	(8p23.2)
					Toolkit (GATK v1.5-16-g58245bf) (base-quality re-calibration and local realignment)		Human Gene Mutation database (HGMD) & NCBI		
							ClinVar database (genotype-phenotype correlation)	- mapping quality (q30 or higher)	
								- depth of coverage (d10 or higher)	
Straniero <i>et al.</i> (2017)	DNAJC12 (DnaJ Heat Shock Protein Family (Hsp40) Member C12)	Canadian and Italian family	Positional cloning (Ion AmpliSeq™ Exome Kit and the Ion Proton™ System, Thermo Fisher Scientific)	WES on index patient (HiSeq 2000, Illumina)	Torrent Suite Software	Torrent Variant Caller (tvc 4.2-18) (variant calling)	ANNOVAR (variant annotation)	Confirmation, segregation analysis and screening via Sanger sequencing	Homozygous c.187A > T - p.K63* (10q21.3) and c.79-2A > G - p.V27Wfs*14 (10q21.3)
Quadri <i>et al.</i> (2018)	LRP10 (Low-density lipoprotein receptor - related protein 10)	Italian family	Genome-wide SNP array genotyping and linkage analysis in ten affected Relatives (HumanCNV370 bead chip, Illumina)	WES on index PD patient (HiSeq 2000, Illumina)	Burrows-Wheeler Aligner (BWA-MEM version 0.5.9 (read alignment to human References genome - UCSC hg19)	Genome-Analysis-Tool-Kit (GATK) v3 (variant calling)	Cartagenia Bench Lab NGS v-5-0-1 (variant filtering) SpliceSiteFinder-like, MaxEntScan, NNSPLICE, GeneSplicer, and Human Splicing Finder integrated in Alamut Visual version 4.2 (splicing effect prediction)	Variants were excluded if - present in dbSNP, Exome Variant Server NHLBI GO Exome Sequencing Project (ESP), 1000 Genomes, Genome of the Netherlands (GoNL), Exome Aggregation Consortium (ExAC) and the Genome aggregation database (GnomAD) databases with a MAF >0.01 Variants were included if	Homozygous - p.Gly603Arg (14q11.2)
			Copy number analysis (Nexus Copy Number, BioDiscovery) MERLIN (linkage analysis)					- heterozygous - exonic - non-synonymous - within 5bp from a splice site - predicted to be pathogenic with ≥5 in silico tools Confirmation by Sanger sequencing	(Continued on following page)

TABLE 1 | (Continued) List of published studies that identified novel Parkinson's disease loci using next-generation sequencing approaches.

Reference	Gene	Population	Pre-NGS screening approach used	Study Participants Screened (Sequencing platform used)	QC and Read Alignment Tools	Variant Calling Tools	Variant Annotation and In Silico Pathogenicity Prediction Tools	Variant Inclusion/Exclusion Criteria	Mutations Identified/ (Chromosome)
	KCNJ15 (Potassium Inwardly Rectifying Channel Subfamily J)			the Ion Torrent (Thermo Fisher Scientific, Waltham, MA, USA) #002 (KCNJ15)	References genome)	SamTools and bedtools2 (variant calling)		ca. accessed on 4 December 2020) for Ion Torrent data Variants were included if - present in affected members of the family while taking into consideration incomplete penetrance - if were exonic or in a splicing region (RefSeq v61) - missense allele - minor allele frequency of <0.01 in the gnomAD database Confirmation via Sanger Sequencing	

Vilariño-Güell and others published their findings on identification of the p.Asn855Ser mutation in *DNAJC13* in 2014 (Vilariño-Güell et al., 2014). WES was conducted on a large PD-affected Canadian-Mennonite family of Dutch German-Russian ancestry. The same mutation and disease-associated haplotype was found in two other families of Mennonite ancestry in the greater Canadian region (Vilariño-Güell et al., 2014). Remarkably, another group, studying the original Canadian-Mennonite family, published their findings in 2016, on a different genetic causal variant, p.Arg141Leu in *TMEM230* (Deng et al., 2016). This difference in disease gene nominations in the same family may be due to differences in methodological approach, including the clinical phenotype used, genotyping approach and pathogenicity prediction scoring of mutations (Farrer, 2019). This highlights the importance of accurate clinical information, particularly in a disease like PD, where the phenotype may overlap with related neurological disorders.

Notably, in the discovery of *CHCHD2* in 2015 in AD-PD, Funayama *et al.*, performed both WES and WGS (Funayama et al., 2015). The authors state that WGS was done on one affected family member to correct for the regions that were inadequately covered during exome capture (Funayama et al., 2015). The use of WGS in combination with WES (particularly in the individual who has the variant of interest) is considered highly beneficial due to its increased coverage and enables screening for CNVs/SNVs in the regions of interest. However, WES continues to be the sequencing method of choice (and was the sole NGS approach used in 16/17 of the studies in **Table 1**), which could largely be attributed to the significant disparity in cost.

In 2016, Sudhaman and others nominated *RIC3* (Sudhaman et al., 2016a) and *PODXL* (Sudhaman et al., 2016b) in South Indian and North Indian families, respectively. For *RIC3*, microsatellite markers were used, prior to WES, to rule out linkage to known AD-PD genes including *SNCA*, *LRRK2* and *VPS35* (Sudhaman et al., 2016a). A similar approach was used to discover *PODXL*. In 2017, a study using WES on a Spanish Basque family led to the discovery of *CSMD1* as a potential disease-causing gene (Ruiz-Martínez et al., 2017). That same year, another study reported a homozygous loss-of-function mutation in *DNAJC12*, using a positional cloning approach in combination with WES (Straniero et al., 2017).

In 2018, two more novel PD genes were reported. In one study, SNP genotyping, linkage analysis, CNV analysis and WES was used in an Italian family to identify the Gly603Arg mutation in *LRP10* (Quadri et al., 2018). In PD, *de novo* mutations may potentially account for several sporadic, EO-PD cases. In the second study, WES and subsequent analysis was performed on trios of Han Chinese ancestry with EO-PD and identified potential pathogenic *de novo* mutations in *NUS1* (Guo et al., 2018). *De novo* mutations are typically rare, deleterious, and difficult to detect with traditional genotyping methods but were effectively identified using only WES in this study (Wang et al., 2019).

In 2019, the identification of *UQCRC1* (a nuclear-encoded gene associated with mitochondrial metabolism) implicated in a Taiwanese PD family with parkinsonism and polyneuropathy, was published (Chen and Lin, 2020; Courtin et al., 2021). This

study was the only one to make use of a comprehensive NGS gene panel to pre-screen ~40 PD-associated genes (including *SYNJ1*, *DNAJC13*, *DNAJC6*, *CHCHD2*, *VPS35*) before performing WES. A study published in 2021 described the discovery of a novel PD gene (*NRXN2*) in a family from South Africa (Sebate et al., 2021). They analyzed WES data from 3 affected individuals from an Afrikaner family, an ethnic group consisting of Dutch, German and French ancestry that are native to South Africa. Most recently, a study examining six families from Australia used WES to narrow down two novel potential disease-causing genes in two families - *SIPA1L1* and *KCNJ15* (Bentley et al., 2021).

It should be noted that true monogenic PD is rare and establishing a familial PD candidate gene as pathogenic can have a degree of uncertainty due to the following factors: isolated findings in familial studies, presence of disease variants in healthy controls, erroneous gene-disease associations or possession of complex phenotypes that may skew towards other, diverse parkinsonisms (Day and Mullin., 2021). Of the candidate genes outlined in this article, *VPS35*, otherwise referred to as *PARK 17*, is firmly associated with classical PD. However, *DNAJC6* (*PARK 19*), *DNAJC13* (*PARK 21*), *SYNJ1* (*PARK 20*), *VPS13C* (*PARK 23*), and *CHCHD2* (*PARK 22*) are also considered pathogenic and viewed as rare genetic contributors to PD disease (Olgiatei et al., 2016; Puschmann, 2017; Schormair et al., 2018; Correia Guedes et al., 2020; Day and Mullin., 2021; Li B et al., 2021). The remaining candidate genes require further study before being categorized as definite PD genes. “Proof of pathogenicity” of novel disease genes require that multiple mutations in the same gene co-segregate with disease in independent families, are absent in large collections of healthy controls or found to be significantly associated with sporadic PD cases (MaCarthur et al., 2014; Farrer, 2019). These criteria seem to necessitate a move away from small family studies and into population-based NGS studies for rare variant discovery - once again relying on large cohorts of individuals. This is also supported by the reasoning that many PD loci may be population-specific and therefore difficult to identify in small studies. The (International Parkinson Disease Genomics Consortium, 2020). However, confirmation of these putative mutations through functional studies or by utilizing model organisms remains a challenge due to the novelty and the large number of variants being identified through NGS.

Consequently, it is clear that there is still a need for NGS studies on PD-affected families for its ability to nominate potentially pathogenic novel genes, even if not seen in other individuals, as this may provide mechanistic insight into PD pathobiology. As seen with the discovery of *NUS1*, where knockout RNAi experiments on *Drosophila* revealed PD phenotypes, lab-based functional analysis of candidate genes is useful to uncovering disease pathogenesis (Guo et al., 2018). However, many studies omit lab-based functional analysis due to the uncertainty as to whether the gene is disease-causing (Rodenburg, 2018). Alternatively, candidate genes can be further associated with a disease of interest through phenotypic associations, determining gene or protein interaction networks or establishing functional similarity with known PD genes using computational methods

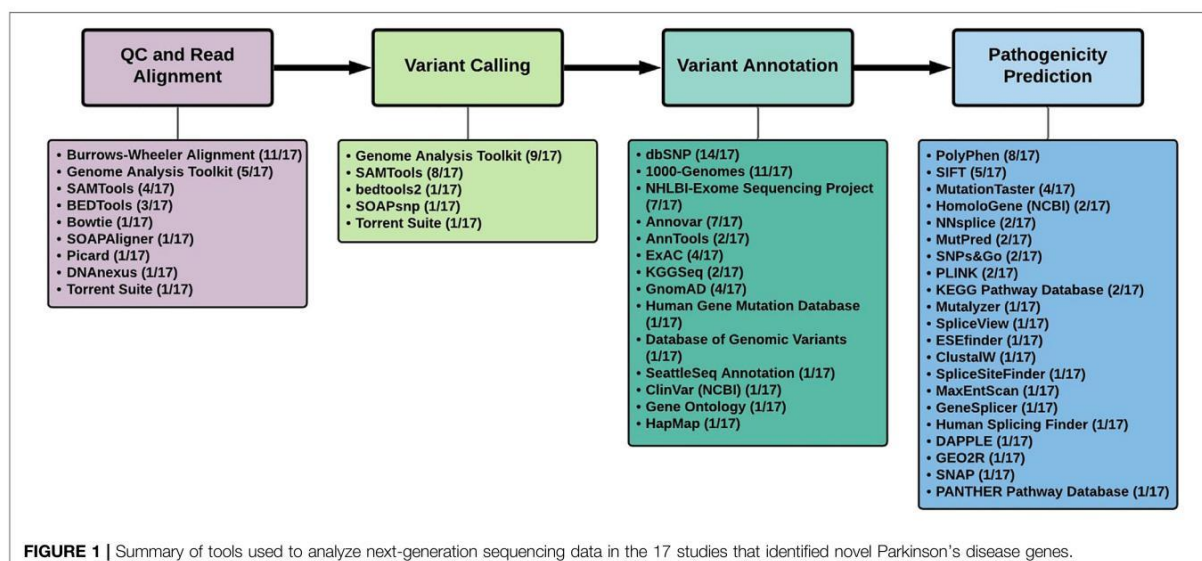
(Chen et al., 2021). Increasingly, a number of machine learning methods that incorporate information from known databases that provide functional annotations (e.g. Gene Ontology), tissue expression data (e.g., Human Protein Atlas) and metabolic/signaling pathways (e.g., Kyoto Encyclopaedia of Genes and Genomes) in order to determine protein or gene interactions between putative and established disease genes (Piro and Di Cunto, 2012). According to a recent study outlining a comprehensive PD gene database (*GENE4PD*), a functional correlation network was simulated between “high confidence” and “suggestive” PD-associated genes in PD pathways resulting in significant associations, including those seen with *RIC3* and *CHCHD2*, with the latter significantly linked to *SNCA*, *PINK1*, *LRRK2*, *PARK7*, and *VPS35* - a likely potential for expanding our knowledge on PD pathway architecture and future annotations (Li B et al., 2021). Furthermore, it is difficult to characterize a gene as being only PD-associated due to the inter-lapping of disease pathways across various parkinsonism disorders (Erratum, 2019; Li W et al., 2021).

Analysis of Bioinformatic Pipelines Used in PD Genomic Studies

Analysis of the tools used in the 17 studies, revealed several similarities and differences (Table 1; Figure 1).

Burrows-Wheeler Aligner (<http://bio-bwa.sourceforge.net/>), specifically the BWA-MEM algorithm, was the software of choice (11/17 studies) for the alignment of the NGS reads to the human reference genome [Figure 1]. The studies reviewed here made use of both the hg18/GRCh36 and hg19/GRCh37 reference genomes. According to one study, SNV detection in WGS data resulted in enhanced genome coverage and a higher number of SNV calls when using GRCh38, as opposed to GRCh37, thereby necessitating the use of the latest reference genome available for NGS analysis (Pan et al., 2019). They conclude that the selection of the aligner in NGS is not as important as the reference genome selection (Pan et al., 2019). The UnifiedGenotyper was used for variant calling in 7 of the 9 studies using the Genome Analysis Toolkit (GATK). This was until the more recent studies, including *NUS1*, *NRXN2* and *KCNJ15*, made use of GATK's HaplotypeCaller for variant calling (Guo et al., 2018). The HaplotypeCaller is now considered best practice for variant calling through GATK's Best Practices Workflows (<https://gatk.broadinstitute.org>) as it allows for SNP/inDEL detection *via de novo* haplotype assembly (Odumpatta & Mohanapriya, 2020). However, a combination of variant callers may be the most efficient method to prioritize variants (Kumaran et al., 2019; Zhao et al., 2020).

Annovar (<https://annovar.openbioinformatics.org/>) and AnnTools (<http://an-ntools.sourceforge.net/>) were the annotation tools used most frequently in 7/17 and 2/17 studies, respectively (Figure 1). These tools are capable of annotating variants using either gene-based, region-based or filtering-based approaches. A typical exome will produce ~20,000 variants with ~10% of these being novel (Belkadi et al., 2015). Thus, the variant filtering tools and exclusion/inclusion criteria must be sufficiently sensitive to identify the most likely causal factors from the ‘background noise’



(Kalinderi et al., 2016). In these PD studies, variants were searched against specific databases to determine allele frequencies. As seen in **Figure 1**, the three most frequently used databases are dbSNP (14/17), the 1000-Genomes-Project (11/17) and the NHLBI - Exome Sequencing Project (7/17), which are currently still considered the most widely used databases for NGS analysis. It was noted that GnomAD, the largest open-source population database, was only mentioned in 4/17 studies and highlights the need to prioritize the use of the larger databases (including the newly released UK Biobank database (<https://www.ukbiobank.ac.uk/>) as it may affect minor allele frequency (MAF) scores used in downstream variant filtering. Several criteria exist to prioritize possible disease-causing variants (Karczewski et al., 2020). Variants are excluded if they are synonymous as they are typically considered to be evolutionary neutral and are likely to have no functional impact on the protein. Variants are also excluded if found to appear in public databases with a MAF >0.01 indicating that the alternate allele is present in more than 1% of the population and is therefore a polymorphism. However, for inclusion, variants must possess PhRED scores >30 (indicating a base call accuracy of 99.9%), be exonic (at present, variants of interest are localized to protein-coding regions as disease-causing variants are likely to impact protein function), have either heterozygous or homozygous genotypes specific to the Mendelian inheritance pattern observed in the family, and also be validated through Sanger sequencing (Vilariño-Güell et al., 2011).

Notably, several caveats need to be considered in the case of PD. Homozygous variants may be disease-causing but may commonly appear in databases such as dbSNP and the 1,000 Genomes Project in heterozygous form, and therefore may be filtered out before variant prioritization (Bras & Singleton., 2011). Furthermore, there are instances in which not all PD affected family members carry the same pathogenic mutation and present as phenocopies (whereby two affected PD individuals with matching phenotypes in a family have different genotypes possibly due to an environmental risk

factor). This phenomenon can easily be confused with intrafamilial heterogeneity (where one affected individual has a different mutation to the family mutation but where this difference may be due to *de novo* mutations, epigenetic changes, or pleiotropy or, in another instance, where multiple rare variants contribute to individual disease risk as seen in oligogenic inheritance (Klein et al., 2011; Farlow et al., 2016; Bentley et al., 2021). True phenocopies in a family may also lead to incorrect conclusions regarding the inheritance pattern within the family (Klein et al., 2011). These confounding factors are relevant in PD, thus requiring adaptation of inclusion criteria in bioinformatic tools going forward.

Popular tools used in these studies to predict the pathogenicity of variants included SIFT (<https://sift.bii.a-star.edu.sg/>) (5/17) and PolyPhen-2 (<http://genetics.bwh.harvard.edu/pph-2/>) (8/15) (Planagan et al., 2010). SIFT determines the effect of amino acid substitution on the protein function whereas PolyPhen-2 predicts the structural and functional impact non-synonymous SNPs have on the protein based on phylogenetic analysis (Odumputta and Mohanapriya., 2020). Furthermore, many of the other pathogenicity prediction tools in **Figure 1** were aimed at identifying variants with splice site effects. Subsequent performance assessment of pathogenicity assessment tools identified other options that outperform PolyPhen-2 and SIFT (Niroula and Vihinen, 2019). Recently, it has been noted that deep neural network models, in conjunction with general pathogenicity predictors such as CADD, are capable of improved variant prioritization as opposed to using the tool alone (Rentzsch et al., 2021). This may open the door to novel machine learning approaches, tailored to the disease of interest, in identifying or confirming disease-causing genes. Many of these newer tools, including RENOVO (Favalli et al., 2021) and DeepPVP (Boudellioua et al., 2019), typically make use of phenotypes to identify gene-disease associations by employing the use of publically available databases including ClinVar.

Also, there is a push to validate the functionality of these novel genes with wet-laboratory-based methods. However, the development of bioinformatic tools to aid the functional analysis of candidate variants may be useful in the interim. VS-CNV (Fortier et al., 2018), dudeML (Hill and Unckless, 2019), CNV-MEANN (Huang et al., 2021) are examples of newer computational software developed to detect and call CNVs in NGS data (including both exome and gene panel data) with CNVnator (Abyzov et al., 2011), Control-FREEC (Boeva et al., 2012) and LUMPY (Layer et al., 2014) still widely used to replace standard multiplex ligation-dependent probe amplification (MLPA), fluorescence *in situ* hybridization (FISH) or microarray CNV detection (Zhang et al., 2019). In the discovery of *NRXN2*, computational protein modelling was performed using the Swissmodel webserver to simulate the potentially disruptive effect of the mutation on protein structure (Sebate et al., 2021).

NGS Approaches to Study PD Genetics in Sub-Saharan Africa

As observed for LRRK2 p.G2019S, some PD-causing mutations may be population-specific (Correia Guedes et al., 2010). Therefore, given the significant differences in ancestral origins, it is likely that the genetic etiology of sub-Saharan populations may be different to that of European and Asian populations (Bope et al., 2019). Mutation screening of Sub-Saharan African individuals with PD has revealed a low frequency in the known PD-causing genes, thus fueling this hypothesis (Williams et al., 2018; Dekker et al., 2020). Additionally, a recent study, using commercial MLPA kits to detect CNVs in individuals with PD from South Africa and Nigeria, observed false-positive deletions due to the presence of SNPs, highlighting the need for data from diverse populations when designing genomic assays for detecting PD mutations (Müller-Nedebock et al., 2021).

The current human reference genome build (GRCh38) is derived from a small sample size, with ~70% of the build derived from a single donor of European ancestry, thereby lacking genetic diversity and therefore inadequate in the context of genetic research in Africa (Wong et al., 2020). Attempts to bridge this fundamental gap in African genomics are currently underway. An example is the South African Human Genome Project initiative to develop a local reference genome based on 24 African ancestry individuals (<https://sahgp.sanbi.ac.za/>). Another initiative is the H3Africa Consortium which aims to develop a pan-African bioinformatics network (H3ABionet) and infrastructure to enhance African genomics research on the continent (Mulder et al., 2017). Additionally, South African researchers have developed a secondary data analysis pipeline to overcome the lack of African allele frequency data in population databases (Schoonen et al., 2019). Their software incorporates Ensembl Variant Effect Predictor (<https://www.ensembl.org/info/docs/tools/vep>) to annotate variants and GENome MINing (GEMINI v0.20) (<https://gemini.readthedocs.io/>) to effectively filter variants according to African allele frequencies, resulting in higher quality output (Schoonen et al., 2019). Furthermore, international efforts in PD are underway to bring underrepresented populations to the fore, through standardized NGS data storage and analysis, as seen with the Global Parkinson's Genetics Program

(Global Parkinson's Genetics Program, 2021) that aims to sequence and analyze PD-affected, at-risk and control individuals from diverse populations to bridge the gap in the 'missing heritability' witnessed in PD.

Recently, the exponential increase in large genomic datasets has necessitated the use of cloud-based systems for the ease of storage, analysis and data-sharing (Navale and Bourne, 2018). However, cloud-based systems can be expensive and require careful consideration of the data use policies to adhere to security in the cloud. Another glaring issue in computational biology is inconsistencies regarding the reproducibility of genomic data analysis and reuseability of open-source analytic software (Russell et al., 2018). A review examining the state of Github repositories of popular bioinformatic tools found that nearly half (46%) of all public repositories had no opensource license and nearly 12% had no version control (Russell et al., 2018). They suggested that software need to be vetted for consistent maintenance by a developer team. Thus, it is important to check the credibility of analysis software before use in a research or clinical setting, and a need for journals to insist on providing datasets and code to reproduce analyses.

Future Directions and Conclusions

The initial studies that discovered *VPS35*, created excitement about the subsequent elucidation of the genetic etiology of PD. However, that initial hope has not been realized with most of the genes identified through NGS, only being found in a single family. This may be due to the complexity of PD etiology, with either, each family having its own rare genetic cause, or that the more common genetic causes underlying PD have not yet been identified. This leads us to question the future direction of NGS approaches in PD.

Third-Generation Sequencing or long-read sequencing are newly-developed approaches that aim to overcome the limitations of existing NGS methods. They produce long-reads that are far more expansive, reducing the complexity of detecting read overlaps—thus increasing the quality of the sequencing data and improving CNV detection (Giani et al., 2019). Approximately 15% of the genome is assumed to be inaccessible due to atypical GC content and repeat elements including trinucleotide repeat expansions which are disease-causing in several neurological disorders, including PD (Keogh and Chinnery, 2013). These mutable regions may harbor pathogenic mutations, particularly compound heterozygous mutations that may only be discovered with long-read sequencing (Mantere et al., 2019). Another limitation of short-read lengths produced by traditional NGS, is potential misalignment of *GBA* (a common genetic risk factor for PD) to its pseudogene which is ~96% identical, resulting in false-positive mutations (Bras and Singleton, 2011). Furthermore, a study that explored the use of targeted-capture long-read sequencing of *SNCA* transcripts, detected previously-undiscovered isoforms capable of translating novel proteins (Tseng et al., 2019). Therefore, in the near future, long-read sequencing may be viewed as the more favorable sequencing alternative for disorders such as PD.

In conclusion, determining the complex genetic architecture underlying PD, particularly in underrepresented populations, is critical to provide insight into PD molecular mechanisms, detection of PD biomarkers, and elucidation of novel drug targets. Thus, this

knowledge will change the course of future clinical diagnoses and therapeutic modalities for this currently incurable disorder. The aim of this article was to explore the use of NGS approaches to identify novel candidate genes in familial PD to consider not only their current relevance in research, but also their future potential in unraveling PD genetics. From our analysis, we recommend the use of third-generation sequencing technologies to identify complex genomic rearrangements and new sequence variation, in combination with current sequencing techniques, to propel future PD genetics research. Furthermore, we recommend that NGS researchers optimize and adjust their sequencing and analytic workflows according to the genetic background of their study participants with PD, and the constant evolution of bioinformatic tools. NGS approaches have revolutionized novel disease gene discovery, however, best practice guidelines need to be developed; taking into account diverse populations and ancestral origins, since it is apparent that a “one-size-fits-all” approach will have significant limitations.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

NSP searched the literature, compiled the table and figure, and wrote the first draft of the manuscript. OAR, AC, and SB wrote

sections and edited the manuscript. All authors approved the final version.

FUNDING

NSP is supported by South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation of South Africa (award number UID 64751); OAR is supported by grants from NIH/NINDS (U54-NS100693, UG3-NS104095, U54-NS110435), Department of Defense (DOD) (W81XWH-17-1-0249), The Michael J. Fox Foundation and American Parkinson disease Association Center for Advanced Research; AC is supported by the South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation of South Africa (award number UID 64751); SB is supported by grants from the South African Medical Research Council (Self-Initiated Research Grant) and the National Research Foundation of South Africa (Grant Numbers: 106052 and 129249).

ACKNOWLEDGMENTS

We acknowledge the support of the DST-NRF Centre of Excellence for Biomedical Tuberculosis Research, South African Medical Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa.

REFERENCES

- Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: An Approach to Discover, Genotype, and Characterize Typical and Atypical CNVs from Family and Population Genome Sequencing. *Genome Res.* 21, 974–984. PMID:21324876. doi:10.1101/gr.114876.110
- Ball, N., Teo, W.-P., Chandra, S., and Chapman, J. (2019). Parkinson's Disease and the Environment. *Front. Neurol.* 10, 218. doi:10.3389/fneur.2019.00218
- Bandres-Ciga, S., Diez-Fairen, M., Kim, J. J., and Singleton, A. B. (2020). Genetics of Parkinson's Disease: An Introspection of its Journey towards Precision Medicine. *Neurobiol. Dis.* 137 (January), 104782. doi:10.1016/j.nbd.2020.104782
- Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A., et al. (2015). Whole-genome Sequencing Is More Powerful Than Whole-Exome Sequencing for Detecting Exome Variants. *Proc. Natl. Acad. Sci. USA* 112 (17), 5473–5478. doi:10.1073/pnas.1418631112
- Bentley, S. R., Guella, L., Sherman, H. E., Neuendorf, H. M., Sykes, A. M., Fowdar, J. Y., et al. (2021). Hunting for Familial Parkinson's Disease Mutations in the Post Genome Era. *Genes* 12 (3), 430. doi:10.3390/genes12030430
- Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., et al. (2012). Control-FREEC: a Tool for Assessing Copy Number and Allelic Content Using Next-Generation Sequencing Data. *Bioinformatics (Oxford, England)* 28 (3), 423–425. doi:10.1093/bioinformatics/btr670
- Bonnefond, A., Durand, E., Sand, O., De Graeve, F., Gallina, S., Busiah, K., et al. (2010). Molecular Diagnosis of Neonatal Diabetes Mellitus Using Next-Generation Sequencing of the Whole Exome. *PLOS ONE* 5 (10), e13630. doi:10.1371/journal.pone.0013630
- Bope, C. D., Chimusa, E. R., Nembaware, V., Mazandu, G. K., de Vries, J., and Wonkam, A. (2019). Dissecting In Silico Mutation Prediction of Variants in African Genomes: Challenges and Perspectives. *Front. Genet.* 10, 601. doi:10.3389/fgene.2019.00601
- Boudelloua, I., Kulmanov, M., Schofield, P. N., Gkoutos, G. V., and Hoehndorf, R. (2019). DeepPVP: Phenotype-Based Prioritization of Causative Variants Using Deep Learning. *BMC Bioinformatics* 20, 65. doi:10.1186/s12859-019-2633-8
- Bras, J., and Singleton, A. (2011). Exome Sequencing in Parkinson's Disease. *Clin. Genet.* 80 (2), 104–109. doi:10.1111/j.1399-0004.2011.01722.x
- Chakravorty, S., and Hegde, M. (2017). Gene and Variant Annotation for Mendelian Disorders in the Era of Advanced Sequencing Technologies. *Annu. Rev. Genom. Hum. Genet.* 18, 229–256. doi:10.1146/annurev-genom-083115-022545
- Chen, J., Althagafi, A., and Hoehndorf, R. (2021). Predicting Candidate Genes from Phenotypes, Functions and Anatomical Site of Expression. *Bioinformatics (Oxford, England)* 37 (6), 853–860. doi:10.1093/bioinformatics/btaa879
- Chen, S., and Lin, X. (2020). Analysis in Case-Control Sequencing Association Studies with Different Sequencing Depths. *Biostatistics (Oxford, England)* 21 (3), 577–593. doi:10.1093/biostatistics/kxy073
- Correia Guedes, L., Ferreira, J. J., Rosa, M. M., Coelho, M., Bonifati, V., and Sampaio, C. (2010). Worldwide Frequency of G2019S LRRK2 Mutation in Parkinson's Disease: a Systematic Review. *Parkinsonism Relat. Disord.* 16 (4), 237–242. doi:10.1016/j.parkrel.2009.11.004
- Correia Guedes, L., Mestre, T., Outeiro, T. F., and Ferreira, J. J. (2020). Are Genetic and Idiopathic Forms of Parkinson's Disease the Same Disease. *J. Neurochem.* 152, 515–522. doi:10.1111/jnc.14902
- Courtin, T., Tesson, C., Corvol, J.-C., Lesage, S., and Brice, A. (2021). Lack of Evidence for Association of UQCRC1 with Autosomal Dominant Parkinson's Disease in Caucasian Families. *Neurogenetics* 22, 365–366. doi:10.1007/s10048-021-00647-4
- Day, J. O., and Mullin, S. (2021). The Genetics of Parkinson's Disease and Implications for Clinical Practice. *Genes* 12 (7), 1006. doi:10.3390/genes12071006

- Dekker, M. C. J., Coulibaly, T., Bardien, S., Ross, O. A., Carr, J., and Komolafe, M. (2020). Parkinson's Disease Research on the African Continent: Obstacles and Opportunities. *Front. Neurol.* 11 (June). doi:10.3389/fneur.2020.00512
- Deng, H.-X., Shi, Y., Yang, Y., Ahmeti, K. B., Miller, N., Huang, C., et al. (2016). Identification of TMEM230 Mutations in Familial Parkinson's Disease. *Nat. Genet.* 48 (7), 733–739. doi:10.1038/ng.3589
- Edvardson, S., Cinnamon, Y., Ta-Shma, A., Shaag, A., Yim, Y.-I., Zenvirt, S., et al. (2012). A Deleterious Mutation in DNAJC6 Encoding the Neuronal-specific Clathrin-Uncoating Co-chaperone Auxilin, Is Associated with Juvenile Parkinsonism. *PLoS ONE* 7 (5), e36458–8. doi:10.1371/journal.pone.0036458
- El-Fishawy, P. (2013). "Common Disease-Rare Variant Hypothesis," in *Encyclopedia of Autism Spectrum Disorders* (New York: Springer), 720–722. doi:10.1007/978-1-4419-1698-3_1997
- Erratum (2019). Genetic Risk of Parkinson Disease and Progression: An Analysis of 13 Longitudinal Cohorts. *Neurol. Genet.* 5 (4), e354. doi:10.1212/NXG.0000000000000354
- Farlow, J. L., Robak, L. A., Hetrick, K., Bowling, K., Boerwinkle, E., Coban-Akdemir, Z. H., et al. (2016). Whole-Exome Sequencing in Familial Parkinson Disease. *JAMA Neurol.* 73 (1), 68–75. doi:10.1001/jamaneurol.2015.3266
- Farrer, M. J. (2019). Doubts about TMEM230 as a Gene for Parkinsonism. *Nat. Genet.* 51 (3), 367–368. doi:10.1038/s41588-019-0354-6
- Favalli, V., Tini, G., Bonetti, E., Voza, G., Guida, A., Gandini, S., et al. (2021). Machine Learning-Based Reclassification of Germline Variants of Unknown Significance: The RENOVO Algorithm. *Am. J. Hum. Genet.* 108 (4), 682–695. doi:10.1016/j.ajhg.2021.03.010
- Fernandez-Marmiesse, A., Gouveia, S., and Couce, M. L. (2018). NGS Technologies as a Turning Point in Rare Disease Research, Diagnosis and Treatment. *Cmc* 25 (3), 404–432. doi:10.2174/0929867324666170718101946
- Flanagan, S. E., Patch, A.-M., and Ellard, S. (2010). Using SIFT and PolyPhen to Predict Loss-Of-Function and Gain-Of-Function Mutations. *Genet. Test. Mol. Biomarkers* 14 (4), 533–537. doi:10.1089/gtmb.2010.0036
- Fortier, N., Rudy, G., and Scherer, A. (2018). Detection of CNVs in NGS Data Using VS-CNV. *Methods Mol. Biol. (Clifton, N.J.)* 1833, 115–127. doi:10.1007/978-1-4939-8666-8_9
- Funayama, M., Ohe, K., Amo, T., Furuya, N., Yamaguchi-Saiki, J., Saiki, S., et al. (2015). CHCHD2 Mutations in Autosomal Dominant Late-Onset Parkinson's Disease: a Genome-wide Linkage and Sequencing Study. *Lancet Neurol.* 14 (3), 274–282. doi:10.1016/S1474-4422(14)70266-2
- Gasser, T., Hardy, J., and Mizuno, Y. (2011). Milestones in PD Genetics. *Mov. Disord.* 26 (6), 1042–1048. doi:10.1002/mds.23637
- Gasser, T. (2015). Usefulness of Genetic Testing in PD and PD Trials: A Balanced Review. *Jpd* 5 (2), 209–215. doi:10.3233/JPD-140507
- Germer, E. L., Imhoff, S., Vilariño-Güell, C., Kasten, M., Seibler, P., Brüggemann, N., Klein, C., and Trinh, J. International Parkinson's Disease Genomics Consortium (2019). The Role of Rare Coding Variants in Parkinson's Disease GWAS Loci. *Front. Neurol.* 10, 1284. doi:10.3389/fneur.2019.01284
- Gialluisi, A., Reccia, M. G., Tirozzi, A., Nutile, T., Lombardi, A., De Sanctis, C., et al. (2020). Whole Exome Sequencing Study of Parkinson Disease and Related Endophenotypes in the Italian Population. *Front. Neurol.* 10, 1362. doi:10.3389/fneur.2019.01362
- Giani, A. M., Gallo, G. R., Gianfranceschi, L., and Formenti, G. (2019). Long Walk to Genomics: History and Current Approaches to Genome Sequencing and Assembly. *Comput. Struct. Biotechnol. J.* doi:10.1016/j.csbj.2019.11.002
- Global Parkinson's Genetics Program (2021). GP2: The Global Parkinson's Genetics Program. *Move. Disord. : official J. Move. Disord. Soc.* 36 (4), 842–851. doi:10.1002/mds.28494
- Guo, J.-f., Zhang, L., Li, K., Mei, J.-p., Xue, J., Chen, J., et al. (2018). Coding Mutations in NUS1 contribute to Parkinson's Disease. *Proc. Natl. Acad. Sci. USA* 115 (45), 11567–11572. doi:10.1073/pnas.1809969115
- Hemminki, K., Försti, A., and Bermejo, J. L. (2008). The 'common Disease-Common Variant' Hypothesis and Familial Risks. *PLoS one* 3 (6), e2504. doi:10.1371/journal.pone.0002504
- Hernandez, D. G., Reed, X., and Singleton, A. B. (2016)., 139. Blackwell Publishing Ltd, 59–74. doi:10.1111/jnc.13593 Genetics in Parkinson Disease: Mendelian versus Non-mendelian Inheritance. *Neurochem.*
- Hildebrandt, F., and Omran, H. (1999). "Positional Cloning and Linkage Analysis," in *Techniques in Molecular Medicine* (Springer Berlin Heidelberg), 352–363. doi:10.1007/978-3-642-59811-1_23
- Hill, T., and Unckless, R. L. (2019). A Deep Learning Approach for Detecting Copy Number Variation in Next-Generation Sequencing Data. *G3 (Bethesda, Md.)* 9 (11), 3575–3582. doi:10.1534/g3.119.400596
- Huang, T., Li, J., Jia, B., and Sang, H. (2021). CNV-MEANN: A Neural Network and Mind Evolutionary Algorithm-Based Detection of Copy Number Variations from Next-Generation Sequencing Data. *Front. Genet.* 12, 700874. doi:10.3389/fgene.2021.700874
- International Parkinson Disease Genomics Consortium (IPDGC) (2020). Ten Years of the International Parkinson Disease Genomics Consortium: Progress and Next Steps. *J. Parkinsons Dis.* 10 (1), 19–30. doi:10.3233/JPD-191854
- Kalinderi, K., Bostantjopoulou, S., and Fidani, L. (2016). The Genetic Background of Parkinson's Disease: Current Progress and Future Prospects. *Acta Neurol. Scand.* 134 (5), 314–326. doi:10.1111/ane.12563
- Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., et al. (2020). The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans. *Nature* 581 (7809), 434–443. doi:10.1038/s41586-020-2308-7
- Keogh, M. J., and Chinnery, P. F. (2013). Next Generation Sequencing for Neurological Diseases: New hope or New Hype. *Clin. Neurol. Neurosurg.* 115 (7), 948–953. doi:10.1016/j.clineuro.2012.09.030
- Kim, H.-J., Won, H.-H., Park, K.-J., Hong, S. H., Ki, C.-S., Cho, S. S., et al. (2013). SNP Linkage Analysis and Whole Exome Sequencing Identify a Novel POU4F3 Mutation in Autosomal Dominant Late-Onset Nonsyndromic Hearing Loss (DFNA15). *PLoS one* 8 (11), e79063. doi:10.1371/journal.pone.0079063
- Klein, C., Chuang, R., Marras, C., and Lang, A. E. (2011). The Curious Case of Phenocopies in Families with Genetic Parkinson's Disease. *Mov. Disord.* 26 (10), 1793–1802. doi:10.1002/mds.23853
- Krebs, C. E., Karkheiran, S., Powell, J. C., Cao, M., Makarov, V., Darvish, H., et al. (2013). The Sac1 Domain of SYNJ 1 Identified Mutated in a Family with Early-Onset Progressive P Arkinsonism with Generalized Seizures. *Hum. Mutat.* 34 (9), 1200–1207. doi:10.1002/humu.22372
- Ku, C.-S., Cooper, D. N., and Patrinos, G. P. (2016). The Rise and Rise of Exome Sequencing. *Public health genomics* 19 (6), 315–324. doi:10.1159/000450991
- Kumaran, M., Subramanian, U., and Devarajan, B. (2019). Performance Assessment of Variant Calling Pipelines Using Human Whole Exome Sequencing and Simulated Data. *BMC Bioinformatics* 20, 342. doi:10.1186/s12859-019-2928-9
- Layer, R. M., Chiang, C., Quinlan, A. R., and Hall, I. M. (2014). LUMPY: a Probabilistic Framework for Structural Variant Discovery. *Genome Biol.* 15 (6), R84. doi:10.1186/gb-2014-15-6-r84
- Lesage, S., and Brice, A. (2012). Role of Mendelian Genes in "sporadic" Parkinson's Disease. *Parkinsonism Relat. Disord.* 18 (Suppl. 1), S66–S70. doi:10.1016/s1353-8020(11)70022-0
- Li, B., Zhao, G., Zhou, Q., Xie, Y., Wang, Z., Fang, Z., et al. (2021). Gene4PD: A Comprehensive Genetic Database of Parkinson's Disease. *Front. Neurosci.* 15, 679568. doi:10.3389/fnins.2021.679568
- Li, W., Fu, Y., Halliday, G. M., and Sue, C. M. (2021). PARK Genes Link Mitochondrial Dysfunction and Alpha-Synuclein Pathology in Sporadic Parkinson's Disease. *Front. Cel Dev. Biol.* 9, 612476. doi:10.3389/fcell.2021.612476
- MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., et al. (2014). Guidelines for Investigating Causality of Sequence Variants in Human Disease. *Nature* 508 (7497), 469–476. doi:10.1038/nature13127
- Mantere, T., Kersten, S., and Hoischen, A. (2019). Long-read Sequencing Emerging in Medical Genetics. *Front. Genet.* 10 (MAY), 1–14. doi:10.3389/fgene.2019.00426
- Mulder, N. J., Adebisi, E., Adebisi, M., Adeyemi, S., Ahmed, A., Ahmed, R., et al. (2017). Development of Bioinformatics Infrastructure for Genomics Research. *gh* 12 (2), 91–98. doi:10.1016/j.ghcart.2017.01.005
- Müller-Nedebock, A. C., Komolafe, M. A., Fawale, M. B., Carr, J. A., Westhuizen, F. H., Ross, O. A., et al. (2021). Copy Number Variation in Parkinson's Disease: An Update from Sub-Saharan Africa. *Mov Disord.* 36, 2442–2444. doi:10.1002/mds.28710
- Nalls, M. A., Blauwendraat, C., Vallerga, C. L., Heilbron, K., Bandres-Ciga, S., Chang, D., et al. (2019). International Parkinson's Disease Genomics Consortium Identification

- of Novel Risk Loci, Causal Insights, and Heritable Risk for Parkinson's Disease: a Meta-Analysis of Genome-wide Association Studies. *The Lancet. Neurology* 18 (12), 1091–1102. doi:10.1016/S1474-4422(19)30320-5
- Navale, V., and Bourne, P. E. (2018). Cloud Computing Applications for Biomedical Science: A Perspective. *Plos Comput. Biol.* 14 (6), e1006144. doi:10.1371/journal.pcbi.1006144
- Niroula, A., and Vihinen, M. (2019). How Good Are Pathogenicity Predictors in Detecting Benign Variants. *Plos Comput. Biol.* 15 (2), e1006481. doi:10.1371/journal.pcbi.1006481
- Odumpatta, R., and Mohanapriya, A. (2020). Next Generation Sequencing Exome Data Analysis Aids in the Discovery of SNP and INDEL Patterns in Parkinson's Disease. *Genomics* 112 (5), 3722–3728. doi:10.1016/j.ygeno.2020.04.025
- Olgati, S., Quadri, M., Fang, M., Rood, J. P. M. A., Saute, J. A., Chien, H. F., Bouwkamp, C. G., Graafland, J., Minneboon, M., Breedveld, G. J., Zhang, J., Verheijen, F. W., Boon, A. J. W., Kievit, A. J. A., Jardim, L. B., Mandemakers, W., Barbosa, E. R., Rieder, C. R. M., Leenders, K. L., Wang, J., and Bonifati, V. International Parkinsonism Genetics Network (2016). D NAJC 6 Mutations Associated with Early-Onset Parkinson's Disease. *Ann. Neurol.* 79 (2), 244–256. doi:10.1002/ana.24553
- Pan, B., Kusko, R., Xiao, W., Zheng, Y., Liu, Z., Xiao, C., et al. (2019). Similarities and Differences between Variants Called with Human Reference Genome HG19 or HG38. *BMC Bioinformatics* 20, 101. doi:10.1186/s12859-019-2620-0
- Piro, R. M., and Di Cunzio, F. (2012). Computational Approaches to Disease-Gene Prediction: Rationale, Classification and Successes. *FEBS J.* 279 (5), 678–696. doi:10.1111/j.1742-4658.2012.08471.x
- Puschmann, A. (2017). New Genes Causing Hereditary Parkinson's Disease or Parkinsonism. *Curr. Neurol. Neurosci. Rep.* 17, 66. doi:10.1007/s11910-017-0780-8
- Quadri, M., Mandemakers, W., Grochowska, M. M., Masius, R., Geut, H., Fabrizio, E., et al. (2018). LRP10 Genetic Variants in Familial Parkinson's Disease and Dementia with Lewy Bodies: a Genome-wide Linkage and Sequencing Study. *Lancet Neurol.* 17 (7), 597–608. doi:10.1016/S1474-4422(18)30179-0
- Rentsch, P., Schubach, M., Shendure, J., and Kircher, M. (2021). CADD-Splice-improving Genome-wide Variant Effect Prediction Using Deep Learning-Derived Splice Scores. *Genome Med.* 13 (1), 1–12. doi:10.1186/s13073-021-00835-9
- Rodenburg, R. J. (2018). The Functional Genomics Laboratory: Functional Validation of Genetic Variants. *J. Inherit. Metab. Dis.* 41 (3), 297–307. doi:10.1007/s10545-018-0146-7
- Ruiz-Martinez, J., Azcona, L. J., Bergareche, A., Martí-Massó, J. F., and Paisán-Ruiz, C. (2017). Whole-exome Sequencing Associates Novel CSMD1 Gene Mutations with Familial Parkinson Disease. *Neurol. Genet.* 3 (5), e177. doi:10.1212/NXG.0000000000000177
- Russell, P. H., Johnson, R. L., Ananthan, S., Harnke, B., and Carlson, N. E. (2018). A Large-Scale Analysis of Bioinformatics Code on GitHub. *PLoS one* 13 (10), e0205898. doi:10.1371/journal.pone.0205898
- Schoonen, M., Seyffert, A. S., van der Westhuizen, F. H., and Smuts, I. (2019). A Bioinformatics Pipeline for Rare Genetic Diseases in South African Patients. *S. Afr. J. Sci.* 115 (3–4), 1–3. doi:10.17159/sajs.2019/4876
- Schormair, B., Kemlink, D., Mollenhauer, B., Fiala, O., Machetzanz, G., Roth, J., et al. (2018). Diagnostic Exome Sequencing in Early-Onset Parkinson's Disease confirms VPS13C as a Rare Cause of Autosomal-Recessive Parkinson's Disease. *Clin. Genet.* 93 (3), 603–612. doi:10.1111/cge.13124
- Schulte, E. C., Stahl, I., Czamara, D., Ellwanger, D. C., Eck, S., Graf, E., et al. (2013). Rare Variants in PLXNA4 and Parkinson's Disease. *PLoS one* 8 (11), e79145. doi:10.1371/journal.pone.0079145
- Sebate, B., Cuttler, K., Cloete, R., Britz, M., Christoffels, A., Williams, M., et al. (2021). Prioritization of Candidate Genes for a South African Family with Parkinson's Disease Using In-Silico Tools. *PLoS one* 16 (3), e0249324. doi:10.1371/journal.pone.0249324
- Shulska, M. V., Alieva, A. K., Vlasov, I. N., Zyrin, V. V., Fedotova, E. Y., Abramychova, N. Y., et al. (2018). Whole-Exome Sequencing in Searching for New Variants Associated with the Development of Parkinson's Disease. *Front. Aging Neurosci.* 10 (MAY), 1–8. doi:10.3389/fnagi.2018.00136
- Straniero, L., Guella, I., Cilia, R., Parkkinen, L., Rimoldi, V., Young, A., et al. (2017). DNAJC12 and Dopa-Responsive Nonprogressive Parkinsonism. *Ann. Neurol.* 82 (4), 640–646. doi:10.1002/ana.25048
- Sudhaman, S., Muthane, U. B., Behari, M., Govindappa, S. T., Juyal, R. C., and Thelma, B. K. (2016b). Evidence of Mutations in RIC3acetylcholine Receptor Chaperone as a Novel Cause of Autosomal-Dominant Parkinson's Disease with Non-motor Phenotypes. *J. Med. Genet.* 53 (8), 559–566. doi:10.1136/jmedgenet-2015-103616
- Sudhaman, S., Prasad, K., Behari, M., Muthane, U. B., Juyal, R. C., and Thelma, B. (2016a). Discovery of a Frameshift Mutation in Podocalyxin-like (PODXL) Gene, Coding for a Neural Adhesion Molecule, as Causal for Autosomal-Recessive Juvenile Parkinsonism. *J. Med. Genet.* 53 (7), 450–456. doi:10.1136/jmedgenet-2015-103459
- Tseng, E., Rowell, W. J., Glenn, O.-C., Hon, T., Barrera, J., Kujawa, S., et al. (2019). The Landscape of SNCA Transcripts across Synucleinopathies: New Insights from Long Reads Sequencing Analysis. *Front. Genet.* 10 (JUL), doi:10.3389/fgene.2019.00584
- Vahidnezhad, H., Youssefian, L., Jazayeri, A., and Uitto, J. (2018). Research Techniques Made Simple: Genome-wide Homozygosity/Autozygosity Mapping Is a Powerful Tool for Identifying Candidate Genes in Autosomal Recessive Genetic Diseases. *J. Invest. Dermatol. B.V.* 138 (Issue 9), 1893–1900. doi:10.1016/j.jid.2018.06.170
- Vilariño-Güell, C., Rajput, A., Milnerwood, A. J., Shah, B., Szu-Tu, C., Trinh, J., et al. (2014). DNAJC13 Mutations in Parkinson Disease. *Hum. Mol. Genet.* 23 (7), 1794–1801. doi:10.1093/hmg/ddt570
- Vilariño-Güell, C., Wider, C., Ross, O. A., Dachsel, J. C., Kachergus, J. M., Lincoln, S. J., et al. (2011). VPS35 Mutations in Parkinson Disease. *Am. J. Hum. Genet.* 89 (1), 162–167. doi:10.1016/j.ajhg.2011.06.001
- Wakeling, M. N., Laver, T. W., Wright, C. F., De Franco, E., Stals, K. L., Patch, A.-M., et al. (2019). Homozygosity Mapping Provides Supporting Evidence of Pathogenicity in Recessive Mendelian Disease. *Genet. Med.* 21 (4), 982–986. doi:10.1038/s41436-018-0281-4
- Wang, W., Corominas, R., and Lin, G. N. (2019). De Novo Mutations from Whole Exome Sequencing in Neurodevelopmental and Psychiatric Disorders: From Discovery to Application. *Front. Genet.* 10 (APR), doi:10.3389/fgene.2019.00258
- Williams, U., Bandmann, O., and Walker, R. (2018). Parkinson's Disease in Sub-Saharan Africa: A Review of Epidemiology, Genetics and Access to Care. *Jmd* 11 (2), 53–64. doi:10.14802/jmd.17028
- Wong, K. H. Y., Ma, W., Wei, C.-Y., Yeh, E.-C., Lin, W.-J., Wang, E. H. F., et al. (2020). Towards a Reference Genome that Captures Global Genetic Diversity. *Nat. Commun.* 11, 5482. doi:10.1038/s41467-020-19311-w
- Zhang, L., Bai, W., Yuan, N., and Du, Z. (2019). Correction: Comprehensively Benchmarking Applications for Detecting Copy Number Variation. *Plos Comput. Biol.* 15 (9), e1007367. doi:10.1371/journal.pcbi.1007367
- Zhao, S., Agafonov, O., Azab, A., Stokowy, T., and Hovig, E. (2020). Accuracy and Efficiency of Germline Variant Calling Pipelines for Human Genome Data. *Sci. Rep.* 10 (1), 20222. doi:10.1038/s41598-020-77218-4
- Zimprich, A., Benet-Pagès, A., Struhal, W., Graf, E., Eck, S. H., Offman, M. N., et al. (2011). A Mutation in VPS35, Encoding a Subunit of the Retromer Complex, Causes Late-Onset Parkinson Disease. *Am. J. Hum. Genet.* 89 (1), 168–175. doi:10.1016/j.ajhg.2011.06.008

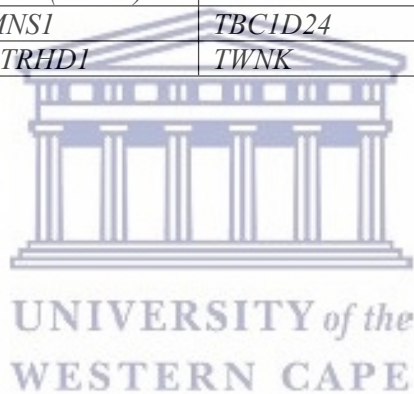
Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Pillay, Ross, Christoffels and Bardien. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Appendix E: List of known and putative PD genes obtained from literature

Known Genes	PD	Putative PD Associated Genes				
<i>SNCA</i>		<i>MAPT</i>	<i>PZP</i>	<i>SLC18A2</i>	<i>CAPS2</i>	<i>NUS1</i>
<i>PRKN</i>		<i>TH</i>	<i>UQCRC1</i>	<i>COL6A5</i>	<i>CEL</i>	<i>OR8B3</i>
<i>UCHL1</i>		<i>ASNA1</i>	<i>DCTN1</i>	<i>ATP1A3</i>	<i>SVOPL</i>	<i>PCDHA9</i>
<i>PARK7</i>		<i>RAB39B</i>	<i>ABCG2</i>	<i>CD36</i>	<i>ATG4C</i>	<i>PRB3</i>
<i>LRRK2</i>		<i>GCH1</i>	<i>ATOH1</i>	<i>CP</i>	<i>CABIN1</i>	<i>PRMT3</i>
<i>PINK1</i>		<i>TNR (TNRC6A)</i>	<i>CCSER1</i>	<i>GRN</i>	<i>COL15A1</i>	<i>PRSS48</i>
<i>POLG2</i>		<i>PODXL</i>	<i>FAM13A</i>	<i>PSEN1</i>	<i>DARS</i>	<i>PTCHD3</i>
<i>HTRA2</i>		<i>CSMD1</i>	<i>FAM13A-AS1</i>	<i>SMPD1</i>	<i>DNAH8</i>	<i>RFLP2</i>
<i>ATP13A2</i>		<i>PTEN</i>	<i>GRID2</i>	<i>FAM83</i>	<i>ELOA2</i>	<i>SCARF2</i>
<i>FBXO7</i>		<i>GPRIN3</i>	<i>HERC3</i>	<i>KIF21A</i>	<i>FAM71A</i>	<i>SPPL2C</i>
<i>GIGYF2</i>		<i>PPMIK</i>	<i>HERC5</i>	<i>PTPRH</i>	<i>FAM90A1</i>	<i>TMEM134</i>
<i>GBA</i>		<i>TARDBP</i>	<i>HERC6</i>	<i>COMT</i>	<i>FER1L6</i>	<i>UHRF1BP1L</i>
<i>PLA2G6</i>		<i>SLC6A3</i>	<i>PIGY</i>	<i>SPG7</i>	<i>GH2</i>	<i>USP20</i>
<i>EIF4G1</i>		<i>ATP10B</i>	<i>PKD2</i>	<i>MCCC1</i>	<i>GPATCH2L</i>	<i>ZNF516</i>
<i>VPS35</i>		<i>MMRN1</i>	<i>PYURF</i>	<i>PLIN4</i>	<i>GRAMD1C</i>	<i>ZNF543</i>
<i>DNAJC6</i>		<i>VAAPB</i>	<i>SMARCARD1</i>	<i>TNK2</i>	<i>IFI35</i>	
<i>SYNJ1</i>		<i>L2HGDH</i>	<i>TIGD2</i>	<i>APOE</i>	<i>KALRN</i>	
<i>DNAJC13</i>		<i>SPP1</i>	<i>ANKRD30A</i>	<i>OGN</i>	<i>KCNK16</i>	
<i>TMEM230</i>		<i>SCN3A</i>	<i>DIS3 (DIS3L)</i>	<i>WDR45</i>	<i>LIPI</i>	
<i>VPS13C</i>		<i>NAPIL5</i>	<i>MNS1</i>	<i>TBC1D24</i>	<i>LPA</i>	
<i>LRP10</i>		<i>ITPR1</i>	<i>PTRHD1</i>	<i>TWNK</i>	<i>MAP3K6</i>	



Appendix F: WES analysis steps, tools and commands

Steps	Tool	Function
Adaptor Identification	bbmap	bbmerge.sh in1=6899-1.FCH22VFBBXY_L2_R1_IAGCAGGAA.FASTQ.gz in2=6899-1.FCH22VFBBXY_L2_R2_IAGCAGGAA.FASTQ.gz outa=adapters.fa ./
Adaptor Trimming	bbmap	bbduk.sh in1=read1.fq in2=read2.fq out1=clean1.fq out2=clean2.fq ref=adapters.fa ktrim=r k=23 mink=11 hdist=1 tpe tbo qtrim=rl trimq=20
Create Reference Genome Index	BWA	bwa index resources_broad_hg38_v0_Homo_sapiens_assembly38.fasta
Align FASTQ files to Reference Genome (hg38)	BWA MEM	bwa mem -M -R resources_broad_hg38_v0_Homo_sapiens_assembly38.fasta clean1_R1.FASTQ.gz clean2_R2.fq.gz > ALN/clean_1.sam
SAM to BAM	Samtools	samtools view -Sb ZA 15_6899-1.sam > ZA 15_6899-1.bam
Namesort BAMs	Picard	SortSam I=ZA 15_6899-1.bam O=ZA 15_6899-1_namesorted.bam SO=queryname ./
BAM File Summary Statistics	Samtools	samtools stat ZA 15_6899-1.bam
Variant Calling(gVCF)	GATk	gatk HaplotypeCaller -R GRCh38_latest_genomic.fna -I ZA 15BAM/ZA 15_6899-1.bam -O ZA 15BAM/VCF/6899-1.g.vcf -ERC GVCF
Combine gVCFs	GATk	gatk CombineGVCFs -R ZA 15MY/resources_broad_hg38_v0_Homo_sapiens_assembly38.fasta -V ZA 15BAM/VCF/6899-1.g.vcf -V ZA 15BAM/VCF/6899-2.g.vcf -V ZA 15BAM/VCF/6899-18.g.vcf -V ZA 15BAM/VCF/6899-19.g.vcf -O ZA 15BAM/VCF/CombinedZA 15.g.vcf
Convert gVCF format to VCF format	GATk	gatk GenotypeGVCFs -R ZA 15MY/resources_broad_hg38_v0_Homo_sapiens_assembly38.fasta -V ZA 15BAM/VCF/CombinedZA 15.g.vcf -O CombinedZA 15.vcf
Isolate SNPs from VCF	GATk	gatk SelectVariants -R ZA 15MY/resources_broad_hg38_v0_Homo_sapiens_assembly38.fasta -V CombinedZA 15.vcf --select-type SNP -o ZA 15BAM/VCF/ZA 15.snps.vcf
Filter Isolated SNPs	GATk	gatk VariantFiltration -R ZA 15MY/resources_broad_hg38_v0_Homo_sapiens_assembly38.fasta -V ZA 15BAM/VCF/ZA 15_snps.vcf -filter "QD < 2.0" --filter-name "QD2" -filter "QUAL < 30.0" --filter-name "QUAL30" -filter "SOR > 3.0" --filter-name "SOR3" -filter "FS > 60.0" --filter-name "FS60" -filter "MQ < 40.0" --filter-name "MQ40" -filter "MQRankSum < -12.5" --filter-name "MQRankSum-12.5" -filter "ReadPosRankSum < -8.0" --filter-name "ReadPosRankSum-8" -O ZA 15BAM/VCF/ZA 15_filtered_snps.vcf
Isolate INDELS from VCF	GATk	gatk SelectVariants -R ZA 15MY/resources_broad_hg38_v0_Homo_sapiens_assembly38.fasta -V CombinedZA 15.vcf --select-type INDEL -o ZA 15BAM/VCF/ZA 15.indels.vcf
Filter Isolated INDELS	GATk	gatk VariantFiltration -R ZA 15MY/resources_broad_hg38_v0_Homo_sapiens_assembly38.fasta -V ZA 15BAM/VCF/ZA 15_indels.vcf -filter "QD < 2.0" --filter-name "QD2" -filter "QUAL < 30.0" --filter-name "QUAL30" -filter "FS >

		200.0" --filter-name "FS200" -filter "ReadPosRankSum < -20.0" --filter-name "ReadPosRankSum-20" -O ZA 15BAM/VCF/ZA 15_filtered_indels.vcf
Merge filtered SNPs and INDELS back into one VCF	GATk	gatk MergeVcfs -I ZA 15_filtered_snps.vcf -I ZA 15_filtered_indels.vcf -O ZA 15_Merged_Filtered.vcf
Include only SNPs and INDELS that passed filtering step in merged VCF	GATk	gatk SelectVariants -R ZA 15MY/resources_broad_hg38_v0_Homo_sapiens_assembly38.fasta -V ZA 15BAM/VCF/ZA 15_Merged_Filtered.vcf -O ZA 15BAM/VCF/ZA 15_Merged_Filtered_PASS.vcf --exclude-filtered
Variant Annotation	Ensembl VEP	vep -I ZA 15BAM/VCF/ZA 15_Merged_Filtered_PASS.vcf --cache -all --fasta genome.fa.gz -O ZA 15_Anno.vcf
Heterozygous SNP Filtering	BCFTools and all2VCF	bcftools isec -i 'GT="het"' -n=3 ~ ZA 15_Anno.vcf "6899-1" "6899-2" - O sites.txt python all2vcf isec --sites sites.txt --vcf hetsaffect.vcf



Appendix G: Pathway and gene expression analysis of the prioritised 24 variants with a CADD score > 20 and expressed in neuro-specific tissue

Variant Identifiers			Tissue Expression			Pathway Analysis		Disease Associations		Gene Interactions	
Gene Symbol	rsID	Protein	GTex	Human Protein Atlas	mGI	KEGG	Reactome	ClinVar	OMIM	STRING	GO
DNAH5	rs113742238	p.R3077Q	Fallopian tube, lung, pituitary gland, cerebellum	Nasopharynx, fallopian tube, bronchus, endometrium	Postnatal lethality, hydrocephalus, respiratory infections, situs inversus and ciliary immotility.	Amyotrophic lateral sclerosis - hsa05014, Huntington disease - hsa05016, Pathways of neurodegeneration - multiple diseases - hsa05022	N/A	Primary ciliary dyskinesia 3	Primary ciliary dyskinesia 3	DNAI2, DNAI1, CCDC114, DNAL1, PAFAH1B1, DCTN1, CLIP1, DCTN2, DNAH6, DNAH3	cilium movement determination of left/right symmetry sperm motility outer dynein arm assembly cilium morphogenesis
NPHP3	rs111727307	p.R1167H	Cervix (uterine), fallopian tube, ovary, endometrium, urinary bladder	Cerebral cortex, cerebellum, testis, ovary, heart muscle	Kidney cysts, enlarged kidneys, increased blood urea nitrogen, kidney inflammation and associated fibrosis, and premature death	N/A	ARL13B-mediated ciliary trafficking of INPP5E, cargo trafficking to the periciliary membrane, trafficking of myristoylated proteins to the cilium, VxPx cargo-targeting to cilium BBSome-mediated cargo-targeting to cilium	Meckel Syndrome, Type 7, Nephronophthisis 3, Renal-Hepatic-Pancreatic Dysplasia 1	Nephronophthisis 3	NEK8, NPHP1, INVS, TMEM67, NEK9, NPHP3, ARL3, NPHP4, CYS1, UNC119B, UNC119	Kidney development heart looping atrial septum development determination of left/right symmetry lung development
DNAH10	rs186639935	p.E698G	Fallopian tube, testis, cervix (uterine), lung, cerebellum	Testis, choroid plexus, bronchus, ovary, fallopian tube	N/A	Amyotrophic lateral sclerosis - hsa05014, Huntington disease - hsa05016, Pathways of neurodegeneration - multiple diseases - hsa05022	N/A	Lipodystrophy, Familial Partial, Type 1, Primary Ciliary Dyskinesia, Charcot-Marie-Tooth Disease	Spermatogenic failure 56	DCTN1, DCTN3, DCTN2, ACTR1A, ACTR10, DCTN4, ACTR, DCTN5, DCTN6,	N/A

										HAP1		
FRMD4B	rs144459338	p.R360W	Thyroid gland	Cerebellum, Adrenal gland, Nasopharynx	N/A	N/A	N/A	N/A	N/A	N/A	CYTH1, CYTH2, CYTH3, SPON1, TOX2, CPNE4, TMTC4, HSPB7, MIPOL1	N/A
DAAM2	rs375083979	p.R209G	Spinal cord, midbrain, hippocampal formation, amygdala, basal ganglia, hypothalamus	Cerebral cortex, hippocampus, caudate, parathyroid gland, adrenal gland	Abnormal ventricular morphology and pressure	Wnt signaling pathway - hsa04310	N/A	Idiopathic Steroid-Resistant Nephrotic Syndrome, Molybdenum Cofactor Deficiency, Cerebral Amyloid Angiopathy, Itm2b-Related, 2	Nephrotic syndrome, type 24		DAAM1, CDC42, PSTPIP1, PSTPIP2, DVL3, DVL1, DVL2, RHOC, RHOA, GBF1	N/A
CLSTN2	rs147617850	p.D289N	Ovary, breast, cerebral cortex	Cerebellum, breast, adipose tissue, kidney	Deficiency in spatial learning and memory	N/A	N/A	Astigmatism, Achondroplasia, Severe, With Developmental Delay And Acanthosis Nigricans	N/A		SLC12A9, KLC2, MTMR2, ICA1L, FAM134C, EPC2, GRIP1, NMNAT3, SLC18A1, CDH17	N/A

STAC	rs111403865	p.P103L	Parathyroid gland, prostate, lung	Cerebellum, epididymis, seminal vesicle, fallopian tube, endometrium, breast, cerebral cortex	N/A	N/A	N/A	Myopathy, Congenital, Bailey-Bloch, Exudative Vitreoretinopathy 6	N/A	KIR2DL1, KIR3DL2, KIR3DL1, KIR3DL3, HLA-G, LILRB2, KIR2DL4, SPG20, TRIM4S, FCHSD2	N/A
CLSTN2	rs140202819	p.E910K	Ovary, breast, cerebral cortex	Cerebellum, breast, adipose tissue, kidney	Deficiency in spatial learning and memory	N/A	N/A	Astigmatism, Achondroplasia, Severe, With Developmental Delay And Acanthosis Nigricans	N/A	SLC12A9, KLC2, MTMR2, ICA1L, FAM134C, EPC2, GRIP1, NMNAT3, SLC18A1, CDH17	N/A
MRE11		p.E406K	Cervix, Endometrium, Spleen	High expression in all tissue	Reduced fertility in female	Homologous recombination - hsa03440, Non-homologous end-joining - hsa03450, Cellular senescence - hsa04218	N/A	Ataxia-Telangiectasia-Like Disorder 1, Ataxia, Early-Onset, With Oculomotor Apraxia And Hypoalbuminemia, Nijmegen Breakage Syndrome-Like Disorder	Ataxia-telangiectasia-like disorder 1	RBBP8, RAD50, XRCC5, BRCA1, ATM, MDC1, XRCC, EXO1, NBN, TP53BP1, RBBP8, TP53BP1	N/A
CD47	rs761086667	p.A252S	Cerebellum, lung, salivary gland	Cerebral cortex, cerebellum, urinary bladder, prostate, fallopian tube	Reduced CD3+ fraction of peripheral lymphocytes and inability to clear infection by E.coli	N/A	Cell surface interactions at the vascular wall, cell-cell communication, extracellular matrix organisation, hemostasis, immune system, innate immune system, integrin cell surface interactions, neutrophil	Hereditary Spherocytosis, Hereditary Elliptocytosis	N/A	ALG1L, TRIM61, POTEI, POTEJ, CCDC183, PPP1R37, NWD1, PRDM7, PQLC1	Positive regulation of cell proliferation positive regulation of cell-cell adhesion positive regulation of T cell activation

							degranulation, signal regulatory protein (SIRP) family interactions				
ZDHC11	rs528116435	p.R276P	Cerebellum, testis, pituitary gland	N/A	Reduced circulating IL6 levels in response to LPS and D-galactosamine or HSV-1	N/A	N/A	N/A	N/A	ZNF30, ZNF181, ZDHC23, TPPP, CLPTMIL, SLCA18, ZDHC13	N/A
KNTC1	rs141767241	p.A1083T	Testis, spleen, small intestine	Cerebral cortex, parathyroid gland, adrenal gland, lung, stomach	N/A	N/A	Cell cycle, cell cycle, mitotic, M Phase, mitotic anaphase, mitotic metaphase and anaphase, mitotic prometaphase, resolution of sister chromatid cohesion, RHO GTPase effectors, RHO GTPases activate formins, separation of sister chromatids, signal transduction, signaling by Rho GTPases	N/A	N/A	BUB1B, BUB1, CDK1, ZWILCH, KIF11, ASPM, CDCA8, CENPF, ZW10, MAD2L1	Mitotic cell cycle checkpoint
MANF	rs545661735	p.A13V	Thyroid gland, pituitary gland, spleen	Cerebellum, hippocampus, caudate, thyroid gland	KO affects de novo protein synthesis in differentiating neuronal stem cells and pancreatic beta cells, disrupting the migration and neurite growth of developing cortical neurons and causing severe growth	N/A	Hemostasis, platelet activation, signaling and aggregation, platelet degranulation, response to elevated platelet cytosolic Ca ²⁺	N/A	N/A	HSPA5, PDIA4, SOD1, HABP4, ANXA5, HSP90B1, HYOU1, CRELD2, DNAJB11	dopaminergic neuron differentiation

					retardation and hyperglycemia						
AHNAK2	rs776830611	p.D1540H	Skin, colon, cervix, uterine	Cerebral cortex, oral mucosa, esophagus, testis, vagina	N/A	Salmonella infection - hsa05132	N/A	Episodic Ataxia	Episodic Ataxia	AHNAK, MYOF, ANXA2, S100A10, CAV1, CACNA1S, ACTN1, PDAP1, ZNF280B, FER1L6	N/A
EIF2A	rs561839835	p.A143V	Cerebellum, cervix (uterine)	Testis, endometrium, skin	No visible phenotypes	N/A	N/A	N/A	N/A	EIF2S1, EIF2S2, EIF2AK2, STAT1, TGFB1, TGFB3, EIF3B, EIF2S3L, EIF2S3	Regulation of translation ribosome assembly
KLHL35		p.R179C	Testis, basal ganglia, hypothalamus	Testis	N/A	N/A	N/A	N/A	N/A	ATP5G2, CORO6, QPCT, ZSCAN18, CCDC8, SCUBE3, SPDYA, CNN2, ALAS1, DLEC	N/A
FAM149B1	rs377021877	p.I149M	Spinal cord, midbrain, hippocampal formation, amygdala, basal ganglia, hypothalamus	Parathyroid gland, adrenal gland, bronchus	N/A	N/A	N/A	Orofaciodigital Syndrome Vi, Spinocerebellar Ataxia 29	Joubert Syndrome 36	DNAJC9, CFAP20, TBC1D32, DHX29, SCAF11, PROM2, OTUD5, SSH2, FFRS1L, C21ORF62	N/A

			mus								
CX3CR1	rs137947370	p.A313V	Spinal cord, midbrain, hypothalamus, spleen, amygdala, basal ganglia	Caudate, adrenal gland, lung	Impaired monocyte recruitment after vascular injury, kidney ischemia and reperfusion	N/A	Chemokine receptors bind chemokines, Class A/1 (Rhodopsin-like receptors), GPCR ligand binding, peptide ligand-binding receptors, signal transduction, signaling by GPCR	Macular Degeneration, Age-Related, Human Immunodeficiency Virus Type 1	Macular Degeneration, Age-Related	CCL22, CX3CL1, CXCL12, CXCL1, CCL5, CCL21, CCL26, CCL2, CCL3, ITGAM	N/A
SALL3	rs150707152	p.R1012Q	Vagina, spinal cord, prostate, basal ganglia, amygdala, hypothalamus, cerebral cortex	N/A	Neonatal lethality with an impaired suckling ability, truncated soft palate, small epiglottis, and abnormal cranial nerve morphology	N/A	N/A	Ureteral Benign Neoplasm, Chromosome 18q Deletion Syndrome	N/A	NANOG, SOX2, POU5F1, DNMT3A, COQ2, CD38, RBFA, HHLA, NOX3, ABCD4	N/A
MZF1		p.S721R	Thyroid gland, cerebellum, fallopian tube	Cerebral cortex, hippocampus, caudate, parathyroid gland, adrenal gland	Late-onset (>2 yr) neoplasias characterised by infiltration, enlargement and disruption of the liver by monomorphic cells	N/A	N/A	N/A	Inflammatory Bowel Disease 27	HELT, MGARP, TERT, PKD2L2, CCBE1, PLVAAP, FOXL1, GGH, GMFG	Negative regulation of transcription from RNA polymerase II promoter regulation of transcription, DNA-templated positive regulation of transcription from RNA polymerase II promoter

AHNAK2	rs11852016	p.P1711L	Skin, colon, cervix, uterine	Cerebral cortex, oral mucosa, esophagus, testis, vagina	N/A	Salmonella infection - hsa05132	N/A	Episodic Ataxia	Episodic Ataxia	AHNAK, MYOF, ANXA2, S100A10, CAV1, CACNA1S, ACTN1, PDAP1, ZNF280B, FER1L6	N/A
ZNF418	rs201309448	p.R667G	Parathyroid gland, skin, thyroid gland, cerebral cortex	Adrenal gland, nasopharynx, bronchus, salivary gland	Aortic banding-induced cardiac hypertrophy and fibrosis	N/A	Gene expression, generic transcription pathway	Ovarian Squamous Cell Carcinoma	N/A	CCDC183, POTEJ, PPP1R37, POTEI, NWD1, ALG1L, TRIM61, KIAA1549L, PQLC1, PRDM7	N/A
IL3RA	rs776812933	p.S91C	Lung, adipose tissue, breast	Cerebral cortex, fallopian tube	N/A	N/A	N/A	Hairy Leukemia Cell	N/A	IL3, STAT5A, STAT5B, PTPN11, JAK2, CSF2RB, GRB2, CSF2, INPP5D	N/A
NPHP3	rs113364886	p.F1324S	Cervix (uterine), fallopian tube, ovary, endometrium, urinary bladder	Cerebral cortex, Cerebellum, Testis, Ovary, Heart Muscle	Kidney cysts, enlarged kidneys, increased blood urea nitrogen, kidney inflammation and associated fibrosis, and premature death	N/A	ARL13B-mediated ciliary trafficking of INPP5E, Cargo trafficking to the periciliary membrane, Trafficking of myristoylated proteins to the cilium, VxPx cargo-targeting to cilium	Meckel Syndrome, Type 7, Nephronophthisis 3, Renal-Hepatic-Pancreatic Dysplasia 1	Nephronophthisis 3	NEK8, NPHP1, INVS, TMEM67, NEK9, NPHP3, ARL3, NPHP4, CYS1, UNC119B, UNC119	Kidney development heart looping atrial septum development determination of left/right symmetry lung development

Appendix H: Primer sequences designed for Sanger sequencing

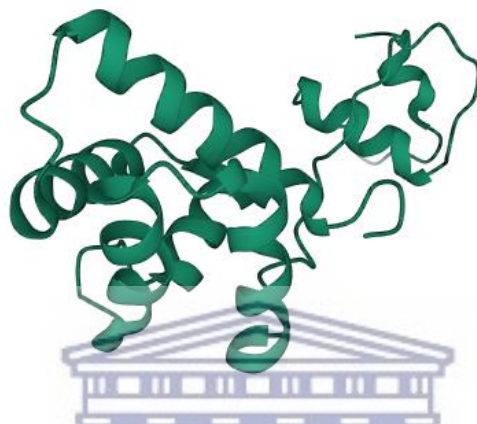
Gene Symbol	Primer pair	Template sequence (5' → 3')	Template strand	Length (bp)	Start (genomic location)	Stop (genomic location)	T _m (°C)	GC (%)	Self-complementarity	Self - 3' complementarity	Product length (bp)
<i>DNAH5</i>	Forward Primer:	AGGGCAGGGAACCTCAAAGC	Plus	20	13776498	13776517	60.54	55.00	5.00	3.00	169
	Reverse Primer:	ATAGCGACCTGGCATCAGTC	Minus	20	13776666	13776647	59.61	55.00	3.00	2.00	
<i>NPHP3</i>	Forward Primer:	ACTAACCTGTCCCTCATAAAGAC	Plus	23	132684456	132684478	57.32	43.48	3.00	2.00	288
	Reverse Primer:	AGGACCAGATCACCTGACT	Minus	20	132684743	132684724	59.58	55.00	4.00	3.00	
<i>DNAH10</i>	Forward Primer:	TCAACTTTTCACCGGCTCTC	Plus	20	-	-	60.40	50.00	0.00	N/a	209
	Reverse Primer:	CTTTTGGAGGGGTCCTCAAT	Minus	20	-	-	60.30	50.00	3.00	N/a	
<i>FRMD4B</i>	Forward Primer:	GCAAAGGGGAATCTGTCCCTA	Plus	22	69196855	69196876	60.03	50.00	4.00	2.00	157
	Reverse Primer:	GCAAAGTGGCTTGTGTGTC	Minus	20	69197011	69196992	59.35	50.00	4.00	2.00	
<i>DAAM2</i>	Forward Primer:	GATGAACAACCTCCAGGGGC	Plus	20	39867633	39867652	60.68	60.00	5.00	2.00	129
	Reverse Primer:	CACACAGCACCCAGGATCTC	Minus	20	39867761	39867742	60.39	60.00	4.00	2.00	
<i>CLSTN2</i>	Forward Primer:	ACTGACTGCACCTGATTCAC	Plus	20	140448473	140448492	57.82	50.00	5.00	3.00	209
	Reverse Primer:	TCAGAGTAGGTCTCCCGGTC	Minus	20	140448681	140448662	59.74	60.00	4.00	1.00	
<i>STAC</i>	Forward Primer:	CAGCGAACCAACAGCGAAGA	Plus	20	36443436	36443455	61.22	55.00	2.00	0.00	260
	Reverse Primer:	CTCACTCAGCTTCCGGGG	Minus	19	36443695	36443677	60.08	63.16	4.00	2.00	
<i>CLSTN2</i>	Forward Primer:	CAGCCCCTGATGAGCATTTG	Plus	20	140566012	140566031	59.26	55.00	4.00	2.00	233
	Reverse Primer:	GAGGGTGGAGTCATCCCACT	Minus	20	140566244	140566225	60.62	60.00	5.00	3.00	
<i>MRE11</i>	Forward Primer:	TTCCCACTGTCAATTTGTTTAAAGA	Plus	24	-	-	59.90	33.30	5.0	3.0	211
	Reverse Primer:	AAATTTGTGGATCGGGTAGC	Minus	20	-	-	58.90	45.00	6.0	2.0	
<i>CD47</i>	Forward Primer:	AGAAAGATGACTCTTACCCGCA	Plus	22	108058320	108058341	59.17	45.45	5.00	0.00	101
	Reverse Primer:	TAACCTCCTTCGTCATTGCCA	Minus	21	108058420	108058400	59.37	47.62	3.00	3.00	
<i>ZDHC11</i>	Forward Primer:	GGAGCAGAGAGACAGGTGTA	Plus	21	837343	837363	60.62	57.14	2.00	2.00	183

	Reverse Primer:	CCGCAGGGCTGGTATCTTGT	Minus	20	837525	837506	62.26	60.00	4.00	0.00	
<i>KNTC1</i>	Forward Primer:	AACAAGAGCTGGAGGCAGAG	Plus	20	122582910	122582929	59.68	55.00	4.00	0.00	106
	Reverse Primer:	GCTATGCAATTCAGGGATCTGG	Minus	22	122583015	122582994	59.18	50.00	4.00	0.00	
<i>MANF</i>	Forward Primer:	AGGAGGAGGAGGATGAGGAG	Plus	20	-	-	59.80	60.00	2.00	0.00	188
	Reverse Primer:	TGGTGATGTTGTGGGGTTC	Minus	19	-	-	60.20	52.60	2.00	0.00	
<i>AHNAK2</i>	Forward Primer:	CCTCTGGGAGTTTCACGTCC	Plus	20	104950761	104950780	60.04	60.00	4.00	3.00	154
	Reverse Primer:	GAGGCCTCAGTGGATGTGTC	Minus	20	104950914	104950895	60.11	60.00	8.00	1.00	
<i>EIF2A</i>	Forward Primer:	AGGTGTCCATCCTGGTCAGAG	Plus	20	-	-	60.50	55.00	5.00	3.00	180
	Reverse Primer:	CAGTATTTAGGAAAACCCTATGATGTC	Minus	27		-	59.60	37.00	4.00	2.00	
<i>KLHL35</i>	Forward Primer:	TTCAAACACGGCCTCCTCG	Plus	19	75429976	75429994	60.30	57.89	4.00	2.00	232
	Reverse Primer:	CTGCGTGCCTTTCTCG	Minus	17	75430207	75430191	59.55	64.71	4.00	2.00	
<i>FAM149B1</i>	Forward Primer:	TTGTATGTATGCCAGTGAAGGTA	Plus	25	73193422	73193446	59.81	40.00	4.00	2.00	128
	Reverse Primer:	GGAAACAGCAGAAGGGGATCT	Minus	21	73193549	73193529	59.72	52.38	4.00	2.00	
<i>CX3CR1</i>	Forward Primer:	AGAACACTTCCATGCCTGCT	Plus	20	39265497	39265516	59.60	50.00	4.00	0.00	186
	Reverse Primer:	TGTGACTGAGACGGTTGCAT	Minus	20	39265682	39265663	59.61	50.00	4.00	2.00	
<i>SALL3</i>	Forward Primer:	GAAATCCACTACCGCAGCCA	Plus	20	78994968	78994987	60.39	55.00	3.00	0.00	123
	Reverse Primer:	AGGCAGCTCTTCAATCTGTGT	Minus	22	78995090	78995069	60.22	45.45	4.00	0.00	
<i>MZF1</i>	Forward Primer:	GGAGCTACTCGGCGCTGT	Plus	18	58562068	58562085	61.83	66.67	4.00	2.00	108
	Reverse Primer:	GCACCCACCGACGAGAGAAAG	Minus	20	58562175	58562156	62.82	65.00	3.00	0.00	
<i>AHNAK2</i>	Forward Primer:	AAACTGGGCATCTGCACCTT	Plus	20	104950205	104950224	60.18	50.00	4.00	2.00	186
	Reverse Primer:	GGTGGAAGCTGATGTGAGCC	Minus	20	104950390	104950371	61.03	60.00	4.00	2.00	
<i>ZNF418</i>	Forward Primer:	TCCCACGTTTGTCACACTCA	Plus	20	57926052	57926071	59.46	50.00	5.00	1.00	288
	Reverse Primer:	TCGAGGAAAGCCTTACGAGT	Minus	20	57926339	57926320	58.46	50.00	5.00	2.00	
<i>IL3RA</i>	Forward Primer:	CAACTACACCGTCCGAGTGG	Plus	20	1348484	1348503	60.39	60.00	4.00	3.00	105
	Reverse Primer:	GGGAGGGAATAGAGAATAAACAAAC	Minus	25	1348588	1348564	57.43	40.00	2.00	0.00	
<i>NPHP3</i>	Forward Primer:	TGCAAAGAATTCTAACTGCTGCT	Plus	23	132681886	132681908	59.18	40.13	8.00	2.00	114
	Reverse Primer:	GCTCCTTCACGCCATTCATC	Minus	20	132681999	132681980	59.34	55.00	2.00	0.00	

Appendix I: MANF analysis figures

A) Dataset identifiers

Gene Symbol	dbSNP ID	AA change	Protein Sequence
<i>MANF</i>	rs545661735	p.A13V	MRRMWATQGLAVALALSVLPGSRALRPGDCEV CISYLGRFYQDLKDRDVTFSPATIENELIKFCREA RGKENRLCYYGATDDAATKIINEVSKPLAHHIP VEKICEKLLKKKDSQICELKYDKQIDLSTVDLKKL RVKELKKILDDWGETCKGCAEKSDYIRKINELM PKYAPKAASARTDL

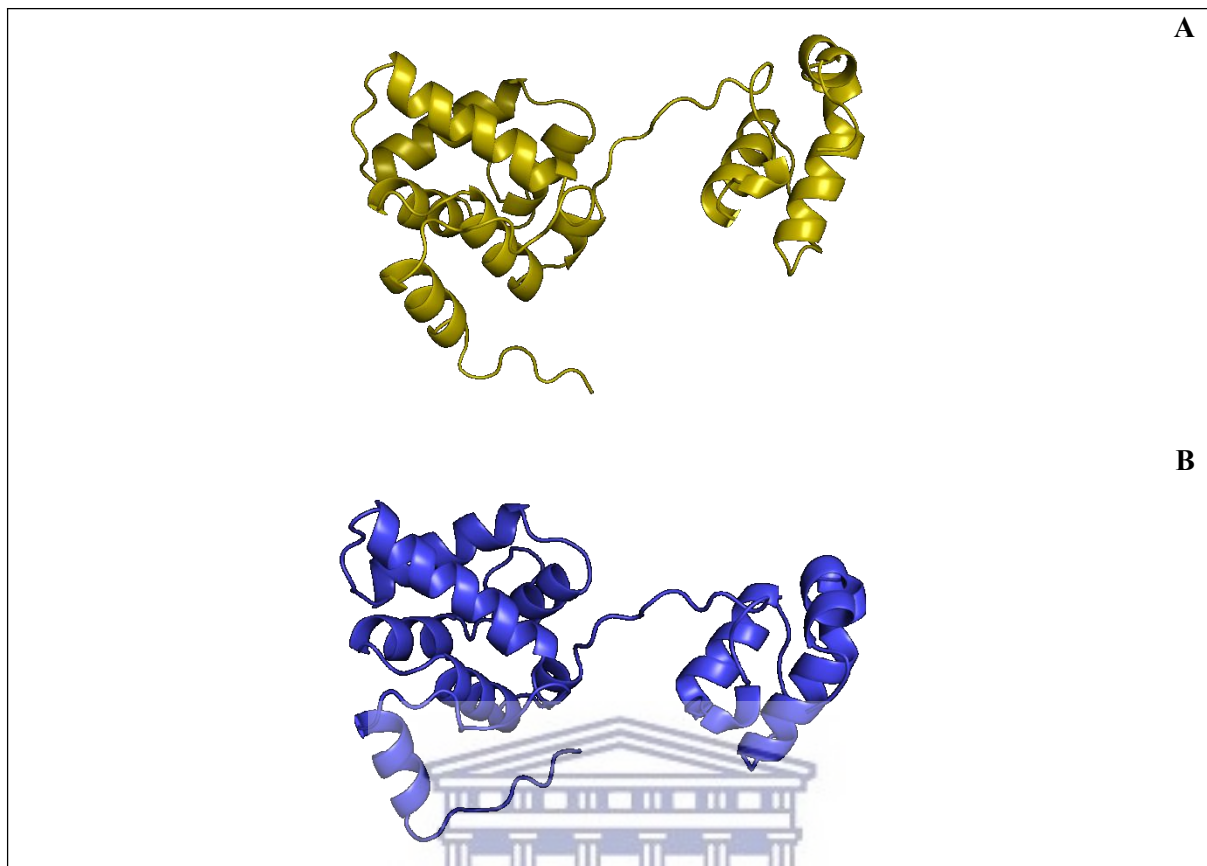


B) Multiple sequence alignment of the MANF protein obtained from ConSurf

001 Input_protein_seq	MRRMWATQGLAVALALS-VLPGSRALRPGDCEV	CISYLGRFYQDLKDR
002 UniRef90_A0AB7AR7_50_230	-RPMWAMAKK-LAVALALS-VLPGSRALRPGDCEV	CISFLGRFYQDLKDR
003 UniRef90_UPI00064AE4F_81_261	-RPMWATQGLAVALALS-VLPGSRALRPGDCEV	CISYLGRFYQDLKDR
004 UniRef90_F7A6B3_2_183	-RPMWVVRG-LAVALAVA-VLPGGTRALRPGDCEV	CISFLGRFYQDLKDR
005 UniRef90_A0A1A6H9P9_1_166	---MRTAKK-LAVALALS-VLPGNSRALRPGDCEV	CISYLGRFYQDLKDR
006 UniRef90_A0A3M0JIA1_2_180	---LAAAKK-LCAVALALL-LLPAGGRRALRDGDCEV	CVTFLLGRFYQSLKDN
007 UniRef90_TOMC43_653_809	---LAAAKK-LCAVALALL-LLPAGGRRALRDGDCEV	CVTFLLGRFYQDLKDR
008 UniRef90_A0A6P7YJCI_1_181	---MVRTKSLVVAVALA-LVPGDSGCTLRREGDCEV	CVSFLGRFYQSLKDD
009 UniRef90_A0A8C3GPD8_96_275	---MVAALSLGKVALLL-LLPAGGRRALRDGDCEV	CVTFLLGRFYQSLKDN
010 UniRef90_H3BAJ2_17_193	---YLLAKK-LVAVALTF-FLHGSQALREGDCEV	CMSFLGRFYQSLKDK
011 UniRef90_A0A1V4K8T3_1_170	---MVAAKK-LWVALALL-LLPAGSRALRDGDCEV	CVTFLLGRFYQSLKDN
012 UniRef90_UPI001CE1966B_4_180	---LSSK-LSVALALA-LLPGSSEALKEGCEV	CVTFLLGKFFQSLKDN
013 UniRef90_A0A226NTT9_2_193	---LAAAKK-LWVALALL-LLPAGSRALRDGDCEV	CVTFLLGRFYQSLKDN
014 UniRef90_UPI000F5PDBFD_106_285	---MDCLSK-LSVALALA-LVPGFGAALKDGDCEV	CVSFLGKFFQTLQDN
015 UniRef90_A0A5C6N1K4_83_262	---MSSVSK-LSVALALV-LLPASAEALKEGCEV	CVIAFLGRFYQSLKDN
016 UniRef90_A0A671W811_3_180	---SLSSK-LSVALV-LVPGPTEALKEGCEV	CLSFLGKFFQSLKEN
017 UniRef90_A0A6J2PSB5_36_210	---LAAAKK-LVAVALT-LVPGSAEALKEGCEV	CVITFLGRFYQSLQDN
018 UniRef90_B7ZTI5_11_180	---LAAAKK-LVAVALT-LVPGSAEALKEGCEV	CVITFLGRFYQSLKER
019 UniRef90_A0A3Q3AQF8_4_180	---LSSK-LSVALV-LVPGAVEALKEGCEV	CVAVFLGRFYQSLKDS
020 UniRef90_UPI001C42C2AC_4_180	---LSSK-LSVALALA-LVLGAAGALKDGDCEV	CVTFLLGKFFQSLSDN
021 UniRef90_g1MTM6_24_180	---LSSK-LSVALALA-LVPGSAEALKEGCEV	CVSFLGKFFQSLKDD
022 UniRef90_UPI00148D1F2A_97_276	MVAALSK-LSVALALA-LLPAPADALKEGDCEV	CLGFLGKFFQSLRDN
023 UniRef90_A0A0P7UQ56_4_178	---LSSK-LSVALALA-LVPGSSEALKEGCEV	CVSFLSRFYQSLKGG
024 UniRef90_A0A096MC24_3_178	---SLSSK-LSVALALA-LVPGFVEGLKEGCEV	CVSFLGKFFQSLKDS
025 UniRef90_A0A8C5DR98_1_179	MSSVSK-LSVVLG-LVPGFPCAGLKAGECEV	CVTFLLSKFYQTLKDN
026 UniRef90_A0A401SWA7_5_179	---LAAAKK-LTGAVALV-FFVVTSSKALKEGCEV	CLSFLERFYNSLKEQ
027 UniRef90_A0A8C5PLZ6_16_176	---LAAAKK-LTGAVALV-FFVVTSSKALKEGCEV	CLSFLERFYNSLKEQ
028 UniRef90_V9L857_7_184	---LVAAK-VAVVVAV-MAVMPVEALKEGDCEV	CVSFLERLYKILNDD
029 UniRef90_S4RGZ2_13_182	---LAAAKK-LVAVALS-RIK-YLVSPNALLTQGDWTV	CVITFLERFYNSLKER
030 UniRef90_UPI000C8287B0_1_121	MWATQGLAVALALS-VLQ-GSRALRAGDCEV	CLSYLGRFYQDLRDR
031 UniRef90_A0A061IBD4_827_960	---LAAAKK-LVAVALS-VLQ-GSRALRAGDCEV	CLSYLGRFYQDLKDR
032 UniRef90_UPI001885BD50_36_194	---LAAAKK-LVAVALS-VLQ-GSRALRAGDCEV	CLSYLGRFYQDLKDR
033 UniRef90_A0A6P3FK51_23_182	---LAAAKK-LVAVALS-VLQ-GSRALRAGDCEV	CLSYLGRFYQDLKDR
034 UniRef90_Q49AH0_30_180	---LAAAKK-LVAVALS-VLQ-GSRALRAGDCEV	CLSYLGRFYQDLKDR
035 UniRef90_H3APH4_10_179	---LAAAKK-LVAVALS-VLQ-GSRALRAGDCEV	CLSYLGRFYQDLKDR

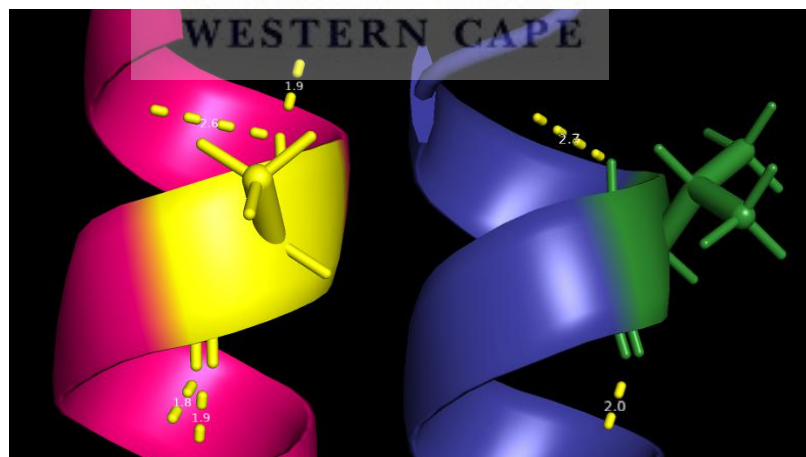
The red bar outlines the 13th residue (Alanine in *H.sapiens*) across the first 35 homologous sequences from different species.

C] Theoretical structures of the wildtype and mutant protein produced by Robetta



[A]- wildtype protein and [B]- variant protein.

D] Polar contacts of the wildtype and mutant protein



The pink helix is the wildtype protein (residue 13; Alanine) and the blue helix is the variant protein (residue 13: Valine).

Appendix J: ACMG classification of p.A13V in MANF

Chromosome ▲	Position ▲	Ref ▲	Alt ▲	Gene (refGene) ▲	InterVar-Adjusted
3	51422811	C	T	MANF	Likely pathogenic

- PVS1: null variant (nonsense, frameshift, canonical +/- 2 splice sites, initiation codon, single or multiexon deletion) in a gene where LOF is a known mechanism of disease
- Strong** PS1: Same amino acid change as a previously established pathogenic variant regardless of nucleotide change
- Strong** PS2: De novo (both maternity and paternity confirmed) in a patient with the disease and no family history
- Strong** PS3: Well-established in vitro or in vivo functional studies supportive of a damaging effect on the gene or gene product
- Strong** PS4: The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls
- Strong** PS5: The user has additional | 1 strong pathogenic evidence
- Moderate** PM1: Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation
- Moderate** PM2: Absent from controls (or at extremely low frequency if recessive) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium
- Moderate** PM3: For recessive disorders, detected in trans with a pathogenic variant
- Moderate** PM4: Protein length changes as a result of in-frame deletions/insertions in a nonrepeat region or stop-loss variants
- Moderate** PM5: Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before
- Moderate** PM6: Assumed de novo, but without confirmation of paternity and maternity
- Moderate** PM7: The user has additional | 1 moderate pathogenic evidence
- Supporting** PP1: Cosegregation with disease in multiple affected family members in a gene definitively known to cause the disease
- Supporting** PP2: Missense variant in a gene that has a low rate of benign missense variation and in which missense variants are a common mechanism of disease
- Supporting** PP3: Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.)
- Supporting** PP4: Patient's phenotype or family history is highly specific for a disease with a single genetic etiology
- Supporting** PP5: Reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory to perform an independent evaluation
- Supporting** PP6: The user has additional | 1 supporting pathogenic evidence
- BA1: Allele frequency is >5% in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium
- Strong** BS1: Allele frequency is greater than expected for disorder
- Strong** BS2: Observed in a healthy adult individual for a recessive (homozygous), dominant (heterozygous), or X-linked (hemizygous) disorder, with full penetrance expected at an early age
- Strong** BS3: Well-established in vitro or in vivo functional studies show no damaging effect on protein function or splicing
- Strong** BS4: Lack of segregation in affected members of a family
- Strong** BS5: The user has additional | 1 strong benign evidence
- Supporting** BP1: Missense variant in a gene for which primarily truncating variants are known to cause disease
- Supporting** BP2: Observed in trans with a pathogenic variant for a fully penetrant dominant gene/disorder or observed in cis with a pathogenic variant in any inheritance pattern
- Supporting** BP3: In-frame deletions/insertions in a repetitive region without a known function
- Supporting** BP4: Multiple lines of computational evidence suggest no impact on gene or gene product (conservation, evolutionary, splicing impact, etc.)
- Supporting** BP5: Variant found in a case with an alternate molecular basis for disease
- Supporting** BP6: Reputable source recently reports variant as benign, but the evidence is not available to the laboratory to perform an independent evaluation
- Supporting** BP7: A synonymous (silent) variant for which splicing prediction algorithms predict no impact to the splice consensus sequence nor the creation of a new splice site AND the nucleotide is not highly conserved
- Supporting** BP8: The user has additional | 1 supporting benign evidence