Identification of novel Parkinson's disease genes in the South African population using a whole exome sequencing approach

Brigitte Glanzmann

Dissertation presented for the degree Doctor of Philosophy (Human Genetics) in the Faculty of Medicine and Health Sciences at Stellenbosch University

Supervisor: Professor Soraya Bardien
Faculty of Medicine and Health Sciences
Department of Biomedical Sciences

Co-supervisor: Dr. Junaid Gamieldien
South African National Bioinformatics Institute
University of the Western Cape

March 2016

DECLARATION

By submitting this dissertation, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature: ……..

Date: ……..

ABSTRACT

Parkinson's disease (PD) is a progressive and severely debilitating neurodegenerative disorder that is characterised by a range of motor symptoms and the selective loss of dopaminergic neurons in the substantia nigra. While the aetiology of PD remains poorly understood, it is hypothesised to involve a combination of various environmental, genetic and cellular factors that independently or collectively contribute to neurodegeneration and ultimately disease. To date, a number of genes including *Parkin, PINK1, LRRK2, SNCA, DJ-1, ATP13A2* and *VPS35* that have been directly associated with disease and investigations of their functions have provided significant insights into the pathobiology of PD. However, these genes do not play a significant role in the South African PD cohort and for this reason, novel genes and pathogenic mutations must be investigated and identified. This will aid in early diagnosis of patients and also ultimately for the design of more effective therapeutic strategies to treat this debilitating and poorly understood chronic systemic disorder.

The present study aimed to identify novel PD-causing mutations in the South African Afrikaner population using a genealogical and whole exome sequencing (WES) approach.. The Afrikaner are unique to South Africa and are known to have undergone a bottleneck in the 1800s which has led to genetic founder effects for a number of different disorders in this particular group. Additionally, we further aimed to determine whether the identified putative disease-causing mutation(s) could be attributed to the development of PD in other South African ethnic groups. A total of 458 patients were recruited, of which 148 were self-identified as Afrikaner. From these, a total of 48 Afrikaner probands were subjected to extensive genealogical analyses and 40 of them could be traced back to a single common couple. For this reason, it was hypothesised that the disorder in these patients may be due to a genetic founder effect.

The use of a whole genome SNP array confirmed the relatedness of the individuals to varying degrees (8 to 12 generations back) and subsequently three of the probands and one affected sibling were selected for WES. The selected individuals were sequenced using the Illumina Genome Hiseq 2000™ and approximately 78 000 variants were identified for each individual. Numerous bioinformatics tools were used to scrutinize the variants but none were able to produce a candidate list of plausible disease-causing variants. All variants identified were either present at high frequency, did not co-segregate with the disorder or were

artefacts. In order to facilitate and expedite the variant prioritisation process, a novel method for the filtration of WES data was designed in-house. This strategy named TAPER™ (Tool for Automated selection and Prioritisation for Efficient Retrieval of sequence variants) implements a set of logical steps by which to prioritise candidate variants that could be pathogenic. It is primarily aimed at the support of resource-constrained scientific environments with limited bioinformatics capacity. As a proof of concept various independent WES datasets for PD, severe intellectual disability and microcephaly as well as ataxia and myoclonic epilepsy were used, and TAPER™ was able to successfully prioritise and identify the causal variants in each case.

Through the use of TAPER™, two putative candidate variants in *SYNJ1* and *USP17* were identified. The homozygous V1405I variant in *SYNJ1* was found only in the affected sibling pair and in none of the 458 patients and 690 control individuals that had been screened. This variant is predicted to be deleterious across multiple platforms and has a CADD score of 29.40 and may alter synaptic vesicle recycling. The homozygous C357S variant in *USP17* was found in 18/458 probands (12 Afrikaner, two white and four mixed ancestry) but was identified in 0.14% of the controls (1/184 Afrikaner, 0/160 white, 0/180 mixed ancestry and 0/160 black). This variant is also anticipated to be deleterious across multiple platforms and has a CADD score of 34.89. In summary, the results of the present study reveal that PD in the 40 South African Afrikaner patients studied is not due to a founder effect, but highlights two variants of interest for future studies. Further work is necessary to analyse both of these variants and to assess their possible effect on protein structure and function.

The discovery of novel PD-causing genes is important as this allows for the generation of disease-linked protein networks, thereby facilitating identification of additional disease genes and subsequently providing insights into the underlying pathobiology. Moreover, this knowledge is critical for the development of improved treatment strategies and drug interventions that will ultimately prevent or halt neuronal cell loss in susceptible individuals. Although the present study did not conclusively identify a novel PD-causing gene, it does provide a solid foundation for future work in our laboratory in the challenging and rapidly evolving research area of WES and bioinformatics, and its application to studies on PD.

OPSOMMING

Parkinson se siekte (PS) is ʼn erg aftakelende neuro-degeneratiewe siekte wat gekenmerk word deur 'n verskeidenheid van simptome en uiteindelik die inkorting van beweging veroorsaak. Hierdie toestand is die gevolg van selektiewe degenerasie van die *dopaminergiese* neurone substantia nigra pars compacta in die midbrein. Dit lei tot patologiese simptome naamlik bradikinese, rus tremore, posturale onstabiliteit en rigiditeit. Aanvanklik was die hipotese dat persone wat PS ontwikkel blootgestel was aan omgewingsverwante snellers wat die aanvang van die siekte veroorsaak. Maar onlangse bewyse dui daarop beide omgewing- en genetiese faktore speel ʼn rol in die patogenese van die siekte. Tans is daar sewe gene (*Parkin, PINK1, LRRK2, SNCA, DJ-1, ATP13A2* en *VPS35*) wat direk betrokke is by PD.

Die doel van die huidige studie is om ʼn 'n PS oorsaak-mutasies in die Suid-Afrikaanse Afrikaner bevolking te identifiseer met behulp van 'n genealogiese en die heel eksoom volgorde-benadering (WES). Die Afrikaner is uniek aan Sui Afrika en het in die 1800s ń genetiese knelpunt ondervind wat tot genetiese stigterseffek gelei het. Daarbenewens het ons verder ten doel om te bepaal of die geïdentifiseerde vermeende siekte-veroorsakende mutasie(s) toegeskryf kan word aan die ontwikkeling van PS in ander Suid-Afrikaanse etniese groepe. ʼn Totaal van 458 pasiënte is vir die studie gewerf, waarvan 148 self-geïdentifiseerde Afrikaners is. ʼn Totaal van 48 Afrikaner probandi was onderworpe aan genealogiese analise en 40 van hulle kon teruggevoer word na 'n enkele gemeenskaplike voorouer. Dit word dus veronderstel dat die individue aan mekaar verwant is en dat PS weens ń stigterseffek is.

Die gebruik van 'n hele genoom SNP verskeidenheid bevestig die verwantskap van die individue in verskillende grade (tussen 8 en 12 generasies) en daarvolgens is drie van die probandi en een geaffekteerde bloedverwant gekies vir WES. Die gekose eksooms is georden volgens die Illumina Genome Hiseq 2000™ en ongeveer 78 000 variante is geïdentifiseer vir elke individu. Verskeie bio-informatika instrumente is gebruik om die variante wat deur WES verkry is te bestudeer maar geen een was in staat om ʼn beweerde lys van geloofwaardige siekte-veroorsakende variante te identifiseer nie. Ten einde die variante identifikasie proses te ondersteun, is ʼn nuwe metode vir filtrasie van WES-data ontwikkel, naamlik TAPER™ (Tool for Automated selection and Prioritization for Efficient Retrieval of sequence variants). TAPER™ implementeer ʼn stel logiese stappe waardeur kandidaat variante gekies word wat

met die siekte geassosieer word; dit het ten doel om ondersteuning te bied aan wetenskaplike omgewings met beperkte bioinformatika kapasiteit. Verder is die sukses van TAPER™ geëvalueer op reeds bestaande data-stelsels wat die konsep bewys.

Met behulp van TAPER™ is twee waarskynlike kandidaat variante in *SYNJ1* en *USP17* geïdentifiseer. Die V1405I variant in *SYNJ1* is slegs in ’n geaffekteerde bloedverwant paar gevind en in geen van die 458 pasiënte of 690 gekeurde kontrole groep individue. Dit word voorspel dat hierdie variant skadelik is en het ń CADD telling van 29.40. Die C357S variant is homosigoties in *USP17* in 18/458 probandi (12 Afrikaner, twee wit en vier gemengde afkoms) gevind is. Maar dit is ook geïdentifiseer in 0.14% van die kontrole individu (1/184 Afrikaner, 0/160 wit, 0/180 gemengde afkoms en 0/160 swart) wat verkry is van die Westelike Provinsie Bloedoortappingsdienste. Dit word voorspel dat hierdie variant skadelik is en het ń CADD telling van 34.89. Die resultate van die huidige studie toon dat PD in die Suid-Afrikaanse Afrikaner nie die oorsprong het by 'n stigterslid nie, maar beklemtoon twee variante van belang. Verdere werk is nodig om elkeen van die variante te analiseer en hul moontlike patogenese te ondersoek.

Die ontdekking van nuwe PS veroorsakende gene is belangrik omdat dit help met die ontwikkeling van siekte-verwante proteïen netwerke, en om sodoende addisionele gene te identifiseer in sleutel siekte prosesse en gevolglik kern biologiese insig in onderliggende prosesse te verskaf. Alhoewel die huidige studie nie ń nuwe PS-veroorsakende geen geïdentifiseer het nie, dit bied wel ń ferm platform vir toekomstige navorsing in die uitdagende en versnellende veranderende velde van WES en bioinformatika en die toepassing daarvan op PS studies.

ACKNOWLEDGEMENTS

I would like to express my sincerest gratitude to the following people and institutions, without whose help this PhD would not have been possible:

Prof. Soraya Bardien, my supervisor and mentor; thank you for your advice, your patience, your positivity and for working as hard as I did on this PhD. I can only hope to follow the example that you have set for me and give back in the way that you have given me.

Dr. Junaid Gamieldien for your support, advice and assistance with the bioinformatics. Thank you for giving me a platform from which to learn and grow.

Drs. Craig Kinnear and Marlo Möller – thank you both for the encouragement with the bioinformatics and helping me wherever and with whatever you could.

Dr. Sihaam Boolay – thank you for being my sounding board, reigning in my bad temper when no one else would and giving me a space on the end of your desk from which I could vent.

Annika Neethling and Juanelle du Plessis for the endless coffee breaks, the pick-me-ups, the time wasting, laughs, blond moments and troubleshooting advice when I needed it.

Celia van der Merwe –over the course of 5 years we have grown together and you are one of the best scientists that I know. I hope to meet up with you soon and compare notes.

The National Research Foundation (NRF), Stellenbosch University and Prof. Paul van Helden for providing financial support.

Hennie and Estelle for giving me a home away from home. Thank you both for taking me in, dusting me off when I needed it and giving me a place to go to when my own company was not good enough.

My parents Konrad and Joan, for teaching me the profound values of perseverance, diligence and inquisitiveness. I would not have been able to complete this dissertation without your unwavering support and words of encouragement even through the difficult times. Papa, danke für ein wichtige Lektion *"Anfangen ist leicht, beharren eine Kunst"*.

Hendri, for your unfailing support, endless patience and most of all... for loving me when I wasn't very lovable.

RESEARCH OUTPUTS

Geldenhuys, G., B. Glanzmann, D. Lombard., et al. 2014. "Identification of a Common Founder Couple for 40 South African Afrikaner Families with Parkinson's Disease." *South African Medical Journal* 104 (6): 413-417.

Glanzmann, B., H. Herbst, C. Kinnear., et al 2015. "A new tool for prioritization of sequence variants from whole exome sequencing data" – **Manuscript submitted.**

PATENT

B. Glanzmann, H. Herbst, C. Kinnear, M. Möller, J. Gamieldien, S. Bardien - Method and System for Filtering Whole Exome Sequence Variants. Patent pending (Provisional patent number: 2015/05726).

TABLE OF CONTENTS

INDEX	PAGE
List of abbreviations	x
List of figures	xv
List of tables	xvii
Outline of dissertation	xviii
Chapter 1: Introduction	1
Chapter 2: Materials and methods	38
Chapter 3: Results	63
Chapter 4: Discussion	106
References	126
Appendix I	140
Appendix II	141
Appendix III	146
Appendix IV	164
Appendix V	167
Appendix VI	168
Appendix VII	174
Appendix VIII	176
Appendix IX	177
Appendix X	178

LIST OF ABBREVIATIONS

1KGP	1000 Genomes Project
AAO	Age at onset
AD	Autosomal dominant
ALS	Amyotrophic lateral sclerosis
ANK	Ankyrin repeat domain
AP/MS	Affinity Purification or Mass Spectrometry
AR	Autosomal recessive
ARM	Armadillo domain
ATP13A2	ATPase type 13 A2
BAM	Binary Alignment/Map
BLAST	Basic Local Alignment Search Tool
CADD	Combined Annotation Dependent Depletion
CAF	Central Analytical Facility
CDC27	Cell division cycle protein 27
CHR	Chromosome
CK	Casein Kinase
CMA	Chaperone-mediated Autophagy
CNS	Central nervous system
CNV	Copy number variations
COR	Carboxy terminal of ROC
CSV	Comma Separated Values
Ct	Cycle threshold
DBS	Deep brain stimulation
ddNTP	Di-deoxyribonucleotide triphosphate
DEIC	Dutch East India Company
DHODH	dihydroorotate dehydrogenase
DJ-1	Daisuke-Junko-1
DNAJC13	DNAJ- Homolog Subfamily C Member 13
dNTP	Deoxyribonucleotide triphosphate
dsDNA	Double stranded DNA
DUB	Deubiquitinating enzyme

EIF4G	Eukaryotic translation initiation factor 4 gamma
ELM	Eukaryotic linear motifs
ESP6500	Exome Sequencing Project 6500
ExAC	Exome Aggregation Consortium
ExoI	Exonuclease I
FATHMM	Functional Analysis Through Hidden Markov Models
FBOX7	F-box only protein 7
FH	Familial hypocholestrolemia
FID	Family indentification
FRET	Fluorescent resonance energy transfer
GATK	Genome Analysis Toolkit
GBA	Glucocerebrosidase
GBD	Global Burden of Disease
GEOPD	Genetic Epidemiology of Parkinson's disease
Grb2	Growth factor receptor-bound protein
GSK	Glycogen Synthase Kinase
GO	Gene Ontology
GTP	Guanosine triphosphate
GWAS	Genome wide association studies
HD	Huntington's disease
HEK23	Human embryonic kidney
HGP	Human Genome Project
HMM	Hidden Markov Model
HP	Human Phenotype
HRM	High Resolution Melt
IBD	Identity by Descent
IBR	In-between RING
IDT	Integrated DNA Technologies
IID	Individual identification
IL	Interleukin
ISFET	In-sensitive field-effect transistor
IVA	Ingenuity Variant Analysis

IVS	Intervening Sequence
LB	Lewy body
LD	Linkage disequilibrium
LRR	Leucine rich repeat domain
LRRK2	Leucine rich repeat kinase 2
MAF	Minor Allele Frequency
MAO	Monoamine oxidase
MAP	Microtubule associated protein
MAPKKK	Mitogen–activated protein kinase kinase kinase
MAPT	Microtubule-associated protein tau
MLPA	Multiplex ligation-dependent probe amplification
MNS	Mental, neurological and substance abuse
MP	Mammalian Phenotype
MPTP	1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine
MRI	Magnetic resonance imaging
MTS	Mitochondrial targeting domain
NAC	Non-amyloid-B component
NEDD4	Neural precursor cell expressed developmentally down-regulated protein 4
NGS	Next generation sequencing
NHGRI	National Human Genome Research Institute
NHLS	National Health Laboratory Services
NQF	Non-fluorescent Quencher
OFS	Orange Free State
OMIM	Online Mendelian Inheritance in Man
PARK2	Parkin
PCA	Principal Component Analysis
PCR	Polymerase chain reaction
PD	Parkinson's disease
PEP	Postencephalitic Parkinsonism
PET	Positron emission tomography
PIKK	PI3 Kinase –related Kinase

PINK1	PTEN-induced kinase 1
PRD	Protein rich domain
PW	Pathway
PXE	Pseudoxanthoma elasticum
QC	Quality Control
qPCR	Quantitative polymerase chain reaction
RNF40	Ring finger protein 40
ROC	Ras of complex proteins
ROS	Reactive oxygen species
RRM	RN recognition motif
SAM	Sequence Alignment/Map
SANBI	South African National Bioinformatics Institute
SAP	Shrimp alkaline phosphatase
SCA	Spinocerebellar ataxia
scaRNA	small Cajal body-specific RNA
SD	Standard deviation
SDS	Sequence detection system
SIFT	Sorting Intolerant From Tolerant
SMTL	SWISS-MODEL template library
SNCA	α-synuclein
snoRNA	small nucleolar RNA
SNP	Single nucleotide polymorphism
SNpc	Substantia nigra pars compacta
SNV	Single nucleotide variant
SPECT	Single photon emission computerized tomography
SSA	sub-Saharan Africa
ssDNA	Single stranded DNA
Ta	Annealing temperature
TAPER	Tool for Automated selection and Prioritisation for Efficient Retrieval of sequence variants
TBCC	Tubulin folding cofactor C
TERT	Telomerase reverse transcriptase

Tm	Melting temperature
TM	Transmembrane region
TNF	Tumor necrosis factor
TRAP	TNF receptor associated protein
Ub	Ubiquitin
UBC	University of British Columbia
UBL	Ubiquitin-like domain
UCHL1	Ubiquitin carboxyterminal hydrolase
UPD	Unique Parkin domain
UPS	Ubiquitin proteasome system
USD	Ubiquitin specific domain
USP	Ubiquitin specific processing protease
VIF	Variance Inflation Factor
VPS35	Vacuolar protein sorting-associated protein 35
WES	Whole exome sequencing
WHO	World Health Organization
WPBTS	Western Province Blood Transfusion Services
WT	Wild type
ZAR	South African Republic
αSYN	α-synuclein protein

LIST OF FIGURES

		PAGE
Figure 1.1	Regions of the brain affected by Parkinson's disease (PD).	3
Figure 1.2	Substantia nigra pars compacta (SNpc) and Parkinson's disease (PD).	4
Figure 1.3	Immunohistochemical stain showing Lewy bodies (LBs) in a Parkinson's disease (PD) patient.	5
Figure 1.4	Key molecular processes implicated in Parkinsonism through genetic findings and exploratory models of disease.	9
Figure 1.5	The Ubiquitin Proteasome System.	20
Figure 1.6	The Autophagy Lysosomal Pathway.	21
Figure 1.7	Sample pipeline for whole exome sequencing result filtration.	28
Figure 1.8	Graphic representation of the South African PD patient group according to ethnicity and disease inheritance pattern.	33
Figure 1.9	Pedigree of the sic Afrikaner PD probands shown to be distantly related through genealogical studies.	36
Figure 2.1	Basic workflow for the generation of a variant called file for further analysis.	46
Figure 2.2	Flow diagram of the two approaches used for variant prioritisation.	48
Figure 2.3	A diagrammatical representation of the approach used for the hypothesis-free approach to novel variant discovery and the backbone for TAPER™.	54
Figure 2.4	Overview of TaqMan® allelic discrimination technology.	57
Figure 2.5	Illustration of the principle underlying high resolution melt (HRM).	59
Figure 2.6	Example of a HRM normalised graph.	60
Figure 2.7	Example of a HRM difference graph.	60
Figure 3.1	Pedigree of the 40 individuals affected with Parkinson's disease shown to be linked to a common founder couple.	65
Figure 3.2	Relatedness inferences from IBD estimates.	73
Figure 3.3	Relatedness inferences from IBD estimates including the control individuals.	77
Figure 3.4	Pedigree of family ZA92.	80
Figure 3.5	Pedigree of family ZA106.	81
Figure 3.6	Pedigree of family ZA111.	81
Figure 3.7	Diagrammatic representation of the amino acid change inducing the G23E variant in *TIMM23*.	87

Figure 3.8	TaqMan® SNP genotyping assay result obtained from IKMB.	89
Figure 3.9	Sequence alignments of ten controls as well as the probands and affected sibling.	90
Figure 3.10	Diagrammatic representation of the amino acid change inducing the P1150S variant in *EFCAB6*.	91
Figure 3.11	Sequence alignments of ZA92 family as well as an unrelated, unaffected control for the P1150S variant in *EFCAB6*.	93
Figure 3.12	HRM normalised graph indicating the heterozygous P1150S variant in the sequence confirmed positive controls.	93
Figure 3.13	HRM difference graph indicating the heterozygous P1150S variant in the sequence confirmed positive controls.	94
Figure 3.14	Sequencing results from the proband with the P1150S variant and additional family members.	95
Figure 3.15	Diagrammatic representation of the amino acid change inducing the V1366I variation in *SYNJ1*.	101
Figure 3.16	HRM difference graph for the V1405I variant in *SYNJ1*.	102
Figure 3.17	Sequence alignments of selected samples for the V1405I variant in *SYNJ1*.	102
Figure 3.18	Diagrammatic representation of the amino acid change inducing the C357S variant in *USP17*.	103
Figure 3.19	HRM difference graph for the C357S variant in *USP17*.	104
Figure 3.20	Sequence alignments of selected samples for the C357S variant in *USP17*.	105
Figure 4.1	Functional and interaction domains of isoforms A and B of SYNJ1.	114
Figure 4.2	Synaptic recycling and PD genes.	116

LIST OF TABLES

		PAGE
Table 1.1	List of genes involved in Parkinson's disease and how they were first identified.	10
Table 1.2	Ethnic breakdowns of 458 South African Parkinson's disease patients recruited for genetic studies.	32
Table 2.1	Afrikaner Parkinson's disease patients selected for whole exome sequencing.	44
Table 3.1	Identity by descent (IBD) scores shared between the siblings of family ZA92.	67
Table 3.2	IBD shared between the original six probands and affected sibling traced back to a common founder couple.	68
Table 3.3	Highest percentage of the chromosomes shared across the six original probands.	70
Table 3.4	Degrees of relatedness between the 40 Afrikaner probands.	71
Table 3.5	The number of shared segments across the 40 probands (chromosomally).	74
Table 3.6	IBD shared between the original six probands and four randomly selected, unaffected Afrikaner controls.	75
Table 3.7	Sequence variants found in *Parkin* in 22 Afrikaner patients.	78
Table 3.8	Summary of WES results across three probands and one affected sibling.	82
Table 3.9	Variants detected in the known PD genes in the three PD probands ZA92, ZA106 and ZA111 as well as the affected sibling.	82
Table 3.10	Overlapping prioritised SNPs across four individuals affected with PD.	86
Table 3.11	Global frequency data of P1150S in *EFCAB6.*	92
Table 3.12	Summary of the total number of variants obtained through each filtration step.	97
Table 3.13	Shortlist of candidate genes prioritised for further analysis	97
Table 3.14	Summary of the genotyping results obtained for the six variants shortlisted for further analysis.	99
Table 4.1	Global population frequencies of V1405I in *SYNJ1* and *USP17* as compared to other PD causing genes.	117

OUTLINE OF THE DISSERTATION

This dissertation involves a next generation sequencing, more specifically whole exome sequencing (WES) investigation of Parkinson's disease (PD) in the South African cohort, identifies numerous Afrikaner PD patients that are related to one another through genealogical tracking and a whole genome SNP array and makes use of WES as a means for the discovery of putative disease-causing candidates. Moreover, this dissertation also provides a detailed description of a novel bioinformatics filtration pipeline.

This dissertation is divided into four chapters:

Chapter One provides a comprehensive background and overview of what is currently known about PD, with specific focus on the genetics and pathobiology of the disease. In addition to this, previous findings of studies conducted on the South African PD patients as well as the overall aims and objectives of the present study will be outlined.

Chapter Two provides a detailed overview of the methodological approaches used throughout the course of this study. Moreover, it describes in detail the design and implementation of a novel bioinformatics pipeline, TAPER™ (Tool for Automated selection and Prioritisation for Efficient Retrieval of sequence variants) that was utilised during the course of the study so as to identify novel, putative disease-causing variants in the South African PD cohort.

Chapter Three is a detailed description of the results obtained throughout the course of the present study. This includes results from the whole genome SNP array showing the relatedness of the 40 Afrikaner PD probands as well as results obtained using conventional WES filtration processes and those obtained through the use of TAPER™.

Chapter Four provides a detailed discussion of the important findings of the dissertation, highlight the possible relevance of the findings and the relevance that they may have to the understanding of PD. In addition to this, it advises on possible future work that may expand the current knowledge of PD in South African patients.

CHAPTER 1: INTRODUCTION

INDEX **PAGE**

1.1 Symptoms and diagnosis of PD 3
1.2 Prevalence and incidence of PD 5
1.3 Genetic aetiology of PD 7
1.3.1 Genes directly implicated in PD 12
1.4 Pathways implicated in PD 17
1.4.1 Mitochondrial dysfunction and oxidative stress 18
1.4.2 The Ubiquitin-Proteasome System 19
1.4.3 The Autophagy-Lysosomal Pathway 20
1.5 Next generation sequencing and whole exome sequencing 22
1.6 Whole exome sequencing platforms and bioinformatics 24
1.6.1 Commonly used WES platforms 25
1.6.2 Data analysis strategies 27
1.6.3 Proof of concept: the use of WES to identify PD-causing genes 29
1.7 Parkinson's disease research in South Africa 30
1.7.1 The South African Afrikaner population 33
1.8 The present study 35
1.8.1 Hypothesis 36
1.8.2 Aims and objectives 37

CHAPTER 1: Introduction

"Involuntary tremulous motion, with lessened muscular power, in parts not in action and even when supported; with a propensity to bend the trunk forwards, and to pass from a walking to a running pace: the senses and intellects being uninjured." Dr James Parkinson – 1817

Although first medically described in detail as a neurological syndrome by Dr James Parkinson in an essay entitled *"An essay on the shaking palsy"*, Parkinson's disease (PD) is a condition that had been identified long before its first official medical documentations (Parkinson 1817); (Raudino 2011). It was first described in the ancient Indian medical system and was (and in some places still is) known as *"Kampa Vata"*. In Western medicine and medical literature, PD was described in 175AD as the 'shaking palsy' by a medical physician known as Galen (Pearce 1989). However, it was only 1642 years later that it was established as a recognised medical condition. Much has been learned about the disease but concomitantly, much of it remains a mystery.

PD (OMIM # 168600) is a severely debilitating neurodegenerative disorder that is characterized by a range of motor symptoms, all of which significantly compromise the movement abilities of an individual (Goetz 2011; Caviness 2014). This disorder is currently without a cure and the root cause for the disease has been pinpointed to the substantia nigra pars compacta (SNpc) in the midbrain (Figure 1.1) (Caviness 2014). Here, the pathological degeneration of the dopaminergic neurons results in an overall decrease in the production of the neurotransmitter dopamine, specifically at the nerve terminals, thus leading to motor circuit dysregulation (Cookson and Bandmann 2010).

The pathology of PD is well understood, but the aetiology remains unclear. For this reason, there are numerous hypotheses that have been constructed in various attempts to solve the conundrum that is PD. Initially, it was suggested that PD is an environmental disease and subsequently caused by environmental factors (Dawson and Dawson 2003), but more recent developments have suggested that it is more likely to be a combination of genetic susceptibility as well as environmental contributors that will result in disease development (Goetz 2011; Caviness 2014).

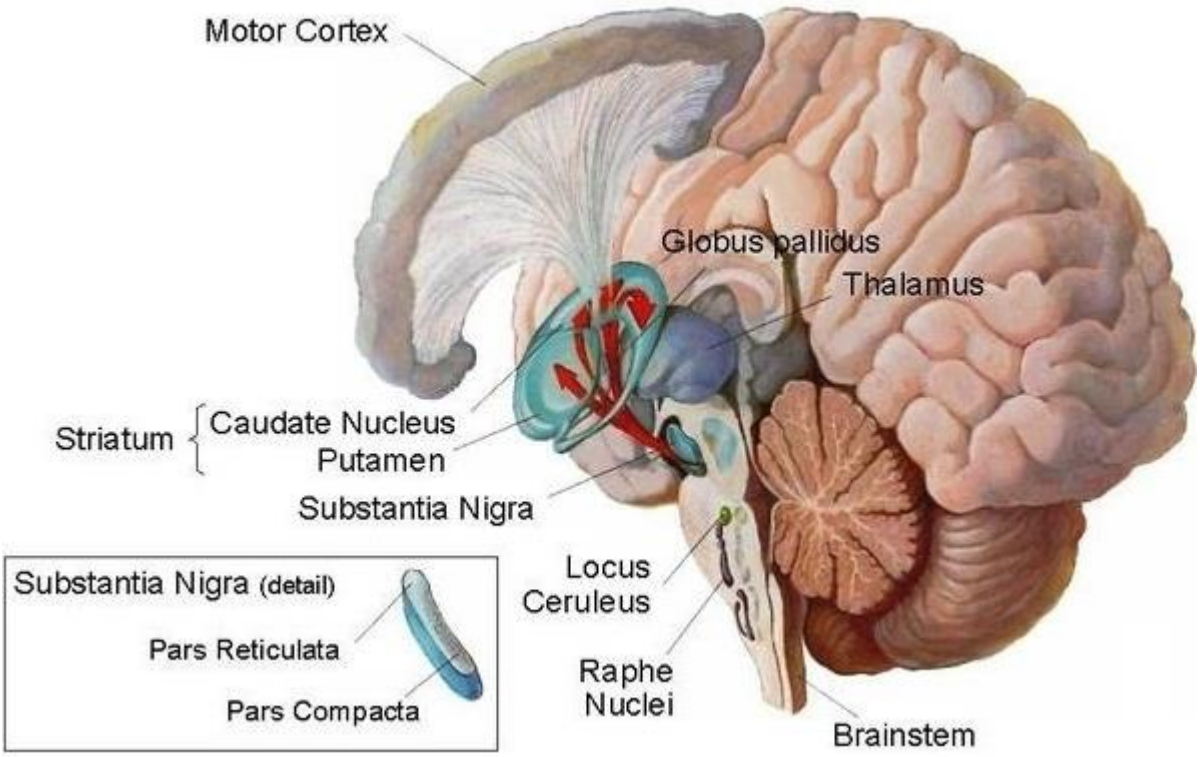


Figure 1.1 Regions of the brain affected by Parkinson's disease (PD). The regions affected by PD are specifically identified in the diagram. Voluntary movements are established in the motor cortex and the output is regulated to the brain stem. The output signal is managed by the subcortical targets that include the thalamus and putamen (taken from http://ayurvedayogashram.com/parkinson-disease.asp).

1.1 Symptoms and diagnosis of PD

The clinical diagnosis of PD is predominantly based on the motor symptoms that include bradykinesia (the inability of a patient to start and continue movements, as well as the inability to adjust the body's position), resting tremor, postural instability and rigidity. In order for the patient to be diagnosed with PD, at least three of the above-mentioned four symptoms must be present. However, bradykinesia is a hallmark characteristic of PD and for this reason, is always required as one of the symptoms for diagnosis (Gibb and Lees 1988).

Due to the complexity of the disease, the United Kingdom (UK) Parkinson's disease Brain Bank has established a standardized method of diagnosis. These criteria are divided into three major steps, each with specific subsections so as to ensure a diagnosis that it is as accurate as possible. These steps for proper disease diagnosis are highlighted in Appendix I.

It should be noted that motor symptoms will only arise in PD patients when approximately 80% of the striatal dopamine and 50% of the nigral neurons have been lost (Bezard and Fernagut 2014). In addition to these prominent motor symptoms, a range of non-motor symptoms that occur prior to the first motor signs characterize PD. Throughout this so-called "premotor period" patients may present with an array of non-motor symptoms with the most common being olfactory disturbances/dysfunction which is characterized by hyposmia (lessened sensitivity to odours) or anosmia (loss of smell – may be total or partial) (Savica, Rocca, and Ahlskog 2010; Bezard and Fernagut 2014). Moreover, patients may also suffer from depression and anxiety, anaemia, rapid eye movement, sleep disturbances as well as gastrointestinal disturbances (Savica, Rocca, and Ahlskog 2010; Bezard and Fernagut 2014).

Regardless of the advances in imaging, clinical diagnostic approaches and tools that are available, pathological confirmation through the use of autopsy is still considered to be the gold standard for PD diagnosis (Poulopoulos, Levy, and Alcalay 2012). Anatomically, the loss of dopaminergic neurons in the SNpc (Figure 1.2) is considered to be the main pathological feature, whereas the accompaniment of intraneuronal accumulation of Lewy bodies (LBs) and Lewy neurites (LNs) both of which are responsible for the dopamine deficiency supports a PD diagnosis (Figure 1.3).

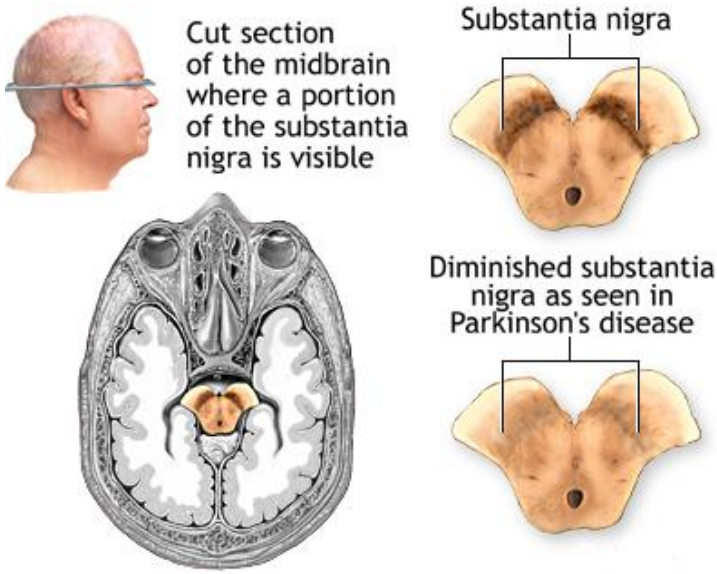


Figure 1.2 Substantia nigra pars compacta (SNpc) and Parkinson's disease. The SNpc is almost intangible in an individual that is affected with PD (taken from http://health.kernan.org/imagepages/19515.htm).

LBs are intra-cytoplasmic inclusions that have a tremendously dense eosiniophilic core and are highly proteinaceous (Gasser 2001; Poulopoulos, Levy, and Alcalay 2012). Interestingly, the major fibrillar component of LBs is α-synuclein, a protein predominantly expressed in the thalamus, SNpc, neocortex and cerebellum. It is hypothesised that amino acid changes or whole gene duplications and triplications of α-synuclein may lead to an increase in aberrant proteins, ultimately leading to neuronal dysfunction and death (Gasser 2001; Dawson and Dawson 2003; Poulopoulos, Levy, and Alcalay 2012).

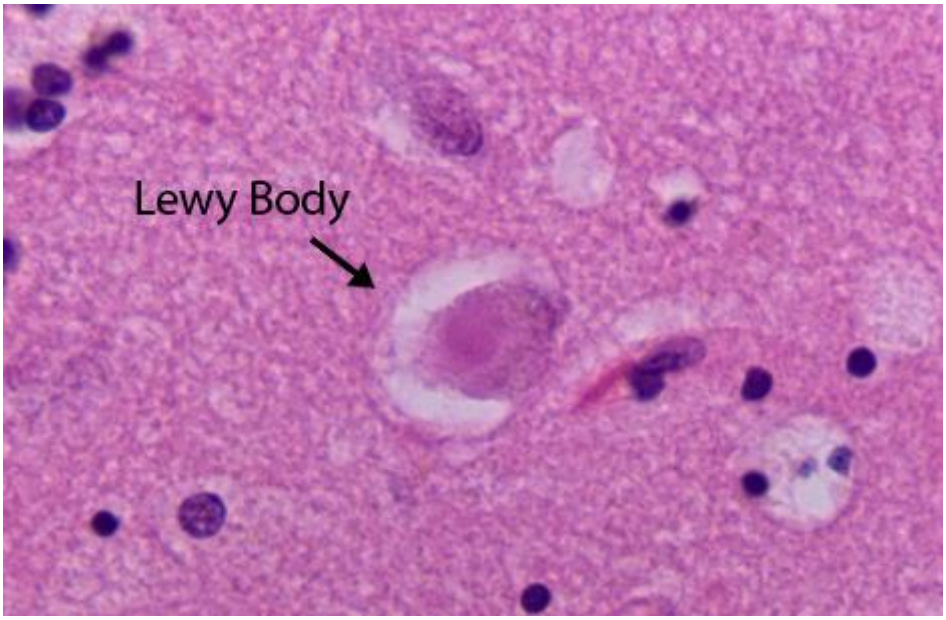


Figure 1.3 Immunohistochemical stain showing Lewy bodies (LBs) in a Parkinson's disease (PD) patient. Lewy bodies are intra-cytoplasmic inclusions that can be identified in patients following autopsy (taken from http://www.medicinenet.com/image-collection/lewy_body_dementia_picture/picture.htm).

1.2 Prevalence and incidence of PD

There is currently no diagnostic test or marker that can be used to identify PD in patients without performing an autopsy post mortem. Although sophisticated equipment such as single photon emission computerized tomography (SPECT) scans and positron emission tomography (PET) scans have been developed, these have yet to be applied to large, population based epidemiological studies. For this reason, clinical criteria established is the most effective means for the diagnosis of PD (de Lau and Breteler 2006; Jankovic 2012). PD is a global disease that affects numerous individuals across various ethnic groups, however the individual and prevalence estimates may vary based on methodological applications as well as geographic locations – both of these factors significantly complicate comparisons of individual studies (de Lau and Breteler 2006).

The prevalence of PD in developed countries is estimated at approximately 0.3% of the entire population and in individuals that are over the age of 60 years of age, this figure increases to around 1% (de Lau and Breteler 2006). It is estimated that the prevalence rate of PD in European countries lies between 108 and 257 per 100 000 individuals but it should be noted that this figure differs from country to country. Interestingly the prevalence of PD in Asian countries is significantly lower, with figures varying from 51.3 to 176.9 per 100 000 individuals across all age groups (Muangpaisan, Hori, and Brayne 2009). Globally, investigations have shown that the prevalence of PD among populations is rising with age; in individuals between 40 – 49 years of age, the prevalence is estimated at 41 per 100 000, between 50 – 59 years of age, prevalence lies at 107 per 100 000. This estimated figure increases four fold between 60 – 69 years of age as the prevalence lies at 428 per 100 000 and nearly triples at 1087 per 100 000 in the 70 - 79 years age categories. This figure increases to 1903 per 100 000 in individuals that age beyond 80 years (Pringsheim et al. 2014). Interestingly, this figure is significantly lower in developing countries and in Africa this figure is strikingly lower, with reported prevalence rates falling between 7 and 43 per 100 000 (Melcon et al. 1997; Okubadejo et al. 2006; de Lau and Breteler 2006).

There are currently standardized incidence rates for PD. The reported incidence rates of PD in developed countries lie between 8 – 18 per 100 000 person years, with a 1.5% lifetime risk of developing the disease (de Lau and Breteler 2006). Moreover, the incidence rates for PD across all age groups has been reported to range between 1.5 – 22 per 100 000 person years but studies that focus exclusively on older populations (where individuals are older than 60 years of age) report PD incidence rates of 529 per 100 000 per year, with an estimated 59 000 new cases per year being reported in the United States alone (Kaplin et al. 2007). The incidence rates of PD in developing countries such as those in Africa is estimated to be around the 4.5 per 100 000 person years mark (Okubadejo et al. 2006; de Lau and Breteler 2006).

The use of stricter diagnostic criteria yields significantly lower estimates of incidence and prevalence and concurrently, these estimates are also directly influenced by so-called case-finding strategies (de Lau and Breteler 2006; Jankovic 2012). Additionally, the estimates surrounding incidence and prevalence rates in developing countries such as sub-Saharan Africa (SSA) are likely to be a gross underestimation due to the methodological problems experienced with some of the studies (hospital-based studies are thought to underestimate PD as most patients are in the community and are not in a hospital or clinical environment) and

the fact that many patients are either misdiagnosed or undiagnosed (Dotchin and Walker 2012).

1.3 Genetic aetiology of PD

PD was long considered to be the direct result of environmental factors. Until 1997, the concept that PD carries a genetic component and subsequently the notion of heritability in PD was contentious – it was considered by many as a "nongenic disorder" (Farrer 2006). Interestingly, the factors supporting PD as a result of environmental influences occurred after the epidemic of postencephalitic Parkinsonism (PEP) after World War I (Casals, Elizan, and Yahr 1998). PEP is thought to be a viral disease which initiates the degeneration of the neurons in the SNpc, thus leading to Parkinsonism, defined as the clinical manifestation of PD symptoms but the predominant phenotype is atypical (Klein, Schneider, and Lang 2009). Two additional factors that suggested PD is an environmental disease was firstly the discovery that MPTP, a by-product of synthetic heroin production, could induce features of PD (Dauer and Przedborski 2003), and secondly a lack of disease concordance in monozygotic twin studies (Tanner et al. 1999).

Over the past 17 years or so, the advances in molecular biology have provided the necessary platform and supporting evidence that PD has a strong genetic component. To date, at least eleven genes have been implicated in PD pathogenesis, each of them contributing independently to the development of the disease or interacting with one another in various molecular processes (Figure 1.4).

Mutations within *LRRK2, SNCA, VPS35, EIF4G1, Parkin, ATP13A2, DJ-1, CHCHD2* and *PINK1* have all been identified in cases of autosomal dominant and autosomal recessive PD (Table 1.1) (Trinh and Farrer 2013). Genes such as *GBA* (glucocerebrosidase), *MAPT* and *DNAJC6* and *DNAJC13* have also been identified as key role players in PD. Homozygous or compound heterozygous mutations in *GBA* have been linked to Gaucher disease – and patients with Gaucher disease type III have often reported Parkinsonism and Lewy body disease post mortem (Sidransky and Lopez 2012). Glucocerebrosidase activity can modulate ceramide metabolism and α-synuclein processing and therefore theoretically α-synucleinopathy, and for this reason has become a potential therapeutic target (Spencer et al. 2011). *MAPT* (Microtubule associated protein Tau) produces a protein product commonly known as tau, which is found to be highly expressed in neurons and is essential in the maintenance of cell structures through microtubule modulation (McMillan et al. 2014).

Genetic association studies have provided significant evidence that there is a relationship between *MAPT* defects and idiopathic PD (Vandrovcova et al. 2010). Aggregation of the tau protein results in so-called "tauopathies" which have been observed in numerous neurodegenerative disorders such as cortico-basal degeneration, frontotemporal dementia with Parkinsonian features, Pick disease and progressive supranuclear palsy (Vandrovcova et al. 2010). The gene most recently identified to be associated with PD is *DNAJC6* and a point mutation within the *DNAJC6* gene was identified in a Dutch-German-Russian Mennonite kindred with late onset PD and the presence of LBs in the autopsies of these individuals (Edvardson et al. 2012). It has been suggested that further research is warranted on the *DNAJC* genes to ensure that this was not a mutation unique to this family (Trinh and Farrer 2013). In addition to this, there are numerous disorders of multiple system degeneration, more commonly known as Parkinson-plus syndromes. These are a group of neurodegenerative diseases that feature the classical symptoms of PD (tremor, rigidity, postural instability and bradykinesia) but with additional features (Trinh and Farrer 2013; Verstraeten, Theuns, and Van Broeckhoven 2015). The most common Parkinson-plus syndromes are progressive supranuclear palsy, multiple system atrophy, cortical-basal ganglionic degeneration and dementia with Lewy bodies. However, there are recessively inherited Parkinson-plus conditions for which genes and variants have been identified. These include a loss of *PLA2G6*, thereby resulting in neuroaxonal dystrophy and a loss of *FBXO7* which results in juvenile onset pallido-pyramidal Parkinsonism (Trinh and Farrer 2013; Verstraeten, Theuns, and Van Broeckhoven 2015)..

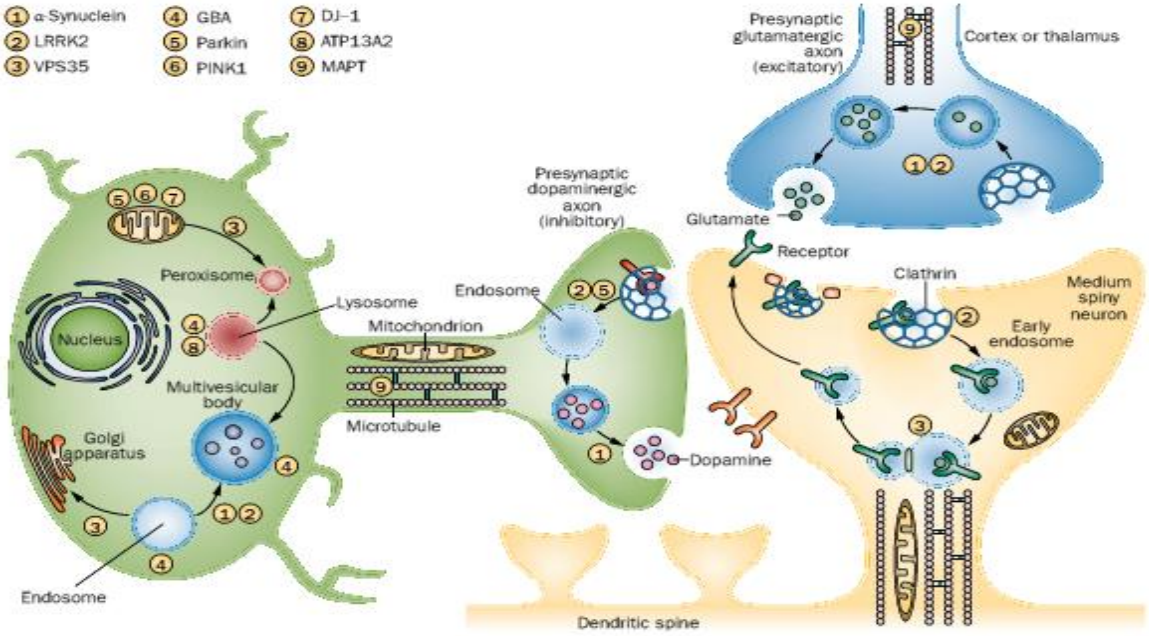


Figure 1.4 Key molecular processes implicated in Parkinsonism as identified through genetic findings and exploratory models of disease. An axon of a presynaptic glutamatergic cortical neuron (in blue), a dendritic spine of a medium spiny neuron (in yellow) and a dopaminergic SNpc neuron (in green) are shown. In presynaptic terminals, α-synuclein (1) promotes exocytosis and aids endocytosis. *LRRK2* (2) regulates phosphorylation of endophilin A, neuronal polarity and arborisation (all postsynaptically). Moreover *LRRK2* also plays a role in chaperone-mediated autophagy, microtubule stabilization and MAPT phosphorylation. *VPS35* (3) is a vital part of the retromer complex that facilitates cargo recognition early endosomes and membrane recruitment in order to form a clathrin-independent carrier. Cargoes may be destined for lysosomal degradation or exosome secretion. *VPS35* facilitates recycling from endosomes to the Golgi apparatus or plasma membrane and vesicle transport between perioxisomes and mitochondria. *GBA* (4) and additional lysosomal acid hydrolases also require the retromer complex for receptor cycling. Loss-of-function mutations in *PINK1* (6), *DJ-1* (7) and *Parkin* (5) affect mitochondrial biogenesis and the induction of autophagy. *Parkin* is directly involved in proteasomal function and ubiquitination and *Parkin* and *PINK1* are involved in mitochondrial maintenance. *ATP13A2* (8) has a role in lysosome mediated autophagy while *MAPT* (9) regulates cargo trafficking and delivery (primarily in the axons) (taken from Trinh and Farrer 2013). Abbreviations: GBA, glucocerebrosidase; LRRK2, leucine-rich repeat kinase 2; VPS35, vacuolar protein sorting 35.

Table 1.1 List of genes involved in Parkinson's disease and how they were first identified.

Genes	Mutations	Clinical Features	How was the gene discovered
Juvenile and Early Onset PD			
Parkin	Various point mutations; exonic rearrangements	LD responsive PD; slowly progressive	Linkage analysis
PINK1	Various point mutations; rare, large deletions	LD responsive PD; akinetic with postural instability and gait disturbance; slow progression	Linkage analysis
DJ-1	Point mutations; large deletions	LD responsive PD; psychological and behavioural disturbances, amyotrophy and cognitive impairment	Linkage analysis
ATP13A2	Point mutations	LD responsive atypical PD associated with supranuclear gaze palsy, spasticity and dementia	Linkage analysis
Late Onset PD			
VPS35	Point mutations	Inconclusive – possibly Lewy body disease	Whole exome sequencing and linkage analysis
LRRK2	Point mutations	Brainstem Lewy body disease, neurofibrillary tangle of TDP43 pathology as well as nigral neuronal loss	Linkage analysis
SNCA	Four point mutations; gene duplications and triplications	Diffuse Lewy body disease with protuberant nigral and hippocampal neuronal loss	Linkage analysis
EIF4G1	Point mutations	Loss of dopaminergic neurons in the substantia nigra and diffuse Lewy body disease	Whole exome sequencing and linkage analysis
CHCHD2	Point mutations	-	Whole exome sequencing and linkage analysis

Genes	Mutations	Clinical Features	How was the gene discovered
Genes associated with PD			
GBA	Point mutations	Glucosidase is a lysosomal hydrolysing glucosylceramide, the penultimate intermediate in degradation of complex glycolipids	Linkage analysis
MAPT	Two distinct haplotypes can be associated with PD (H1 and H2)	Promotion of microtubule assembly and stability	Linkage analysis
DNAJC6	Point mutations	Regulates the transport of target proteins from the endoplasmic reticulum to the cell surface	Whole exome sequencing
DNAJC13	Point mutations	Regulates the transport of target proteins from the endoplasmic reticulum to the cell surface	Whole exome sequencing
FBOX7	Point mutations	Substrate recognition component of a SKP1-CUL1 F-box protein E3 ubiquitin ligase complex which mediates the ubiquitination and proteasomal degradation of target proteins	Linkage analysis
UCHL1	Point mutations	A thiol protease that hydrolyses a peptide bond at the C-terminal glycine of ubiquitin	Linkage analysis
PLA2G6	Point mutations	Catalyses the release of fatty acids from phospholipids.	Homozygosity mapping

Adapted from Trinh and Farrer 2013; Abbreviations: AR - Autosomal Recessive; AD - Autosomal Dominant; *GBA* - glucocerebrosidase; *LRRK2* - leucine-rich repeat kinase 2; *VPS35* - vacuolar protein sorting 35; *EIF4G1* - eukaryotic translation initiation factor 4G1; *PINK1* - PTEN-induced kinase 1; *SNCA* - α synuclein; *UCHL1* - ubiquitin carboxyterminal hydrolase 1; *MAPT* - microtubule-associated protein tau; *FBOX7* - F-box only protein 7; *DNAJC* - DNAJ- Homolog Subfamily C; *CHCHD2* – Coiled-coil-helix-coiled-coil-helix domain-containing protein 2.

1.3.1 Genes directly implicated in PD

Parkin, PINK1* and *DJ-1

Parkin, PINK1 and *DJ-1* have been referred to as the "Three Musketeers of Neuroprotection" (Trempe and Fon 2013). These genes encode very specific proteins, each with distinct enzyme activities whose separate functions and combined interactions appear to confer a role in neuroprotection (Trempe and Fon 2013). For this reason, mutations found in these three genes contribute to neurodegenerative disorders such as PD. PINK1 and Parkin are active role players in mitophagy (the selective degradation of mitochondria through autophagy), while DJ-1 acts as a redox sensor against oxidative stress.

Parkin was the first gene to be associated with autosomal recessive PD (Kitada et al. 1998; Luecking et al. 2000). It encodes a 465 amino acid protein that belongs to the E3 ubiquitin ligase family (Beasley, Hristova, and Shaw 2007). *Parkin* has five specific domains that enable it to carry out its function. These domains are the N-terminal ubiquitin-like domain (UBL), a cysteine-rich unique parkin domain and two C-terminal RING domains that are separated by an in-between-RING domain (IBR). It should be noted that E3 ligases are of particular importance within the cell as an integral part of the Ubiquitin Proteasome System (UPS), which is responsible for removal and recycling of dysfunctional and damaged proteins. E3 ligases catalyse the transfer of ubiquitin from an E2 ubiquitin conjugating enzyme to a protein substrate, tagging the protein for degradation via the 26S proteasome (Trempe and Fon 2013). Parkin therefore plays an essential role as an E3 ligase in protein degradation via the UPS by tagging proteins with ubiquitin (Beasley, Hristova, and Shaw 2007).

PTEN–induced putative kinase 1 (*PINK1)* was the first gene that effectively linked PD to the mitochondria (Valente et al. 2004) and was then further identified in autosomal recessive PD. *PINK1* encodes a 581 amino acid protein which is cytoplasmic, but associates with the mitochondria and is composed of an N–terminal mitochondrial targeting sequence, a serine/threonine kinase domain, a C-terminal domain (function is unknown) and a transmembrane helix (Valente et al. 2004; Trinh and Farrer 2013). Studies support the concept that PINK1 has significant neuroprotective roles within the cell and protects the cell from oxidative stress, mitochondrial dysfunction and cell apoptosis (Matsuda, Kitagishi, and Kobayashi 2013). Mutations in this protein have differential effects on its ability to phosphorylate protein substrates and more specifically, PINK1 is thought to prevent

apoptosis as well as mitochondrial dysfunction that is a direct consequence of protein inhibition (Rohe et al. 2004; Trempe and Fon 2013; Trinh and Farrer 2013).

The *DJ-1* gene was initially identified as an oncogene but mutations in this gene have now been linked to autosomal recessive early onset PD (Nagakubo et al. 1997; Bonifati et al. 2003). The protein product of this gene is 189 amino acids in length and is located in the cytoplasm (van Duijn et al. 2001; Bonifati et al. 2003). This protein belongs to the DJ-1/Thi/PfpI protein superfamily. All proteins belonging to this family are oligomers that are responsible for the maintenance of cellular biochemical activity and stability (Wilson et al. 2004). DJ-1 has neuroprotective activity and directly affects cell sensitivity to oxidative stress (Canet-Avilés et al. 2004; Martinat et al. 2004). However, it remains unclear as to how DJ-1 carries out these functions – it is hypothesised that the neuroprotective effects as well as oxidative stress sensitivity is mediated through the localization of the mitochondria where oxidative stress reduction is induced through the inhibition of components (one such component is rotenone, a pesticide that inhibits mitochondrial complex I) within the respiratory chain (Canet-Avilés et al. 2004; Blackinton et al. 2009; Trempe and Fon 2013). Although the complete mechanism by which DJ-1 functions within the cell is not yet understood, it has been documented that DJ-1 deficiency leads to altered mitochondrial morphology, and increases in reactive oxygen species (ROS) due to the changes in mitochondrial dynamics (Irrcher et al. 2010).

To summarize, PINK 1 and Parkin play an essential role in mitophagy while DJ-1 is a redox sensor of oxidative stress*; PINK1* (a mitochondrial-associated protein kinase that is located at outer mitochondrial membrane) acts upstream of Parkin (an E3 ubiquitin ligase that facilitates the degeneration of damaged mitochondria) and together, the trio plays an essential role in the maintenance of healthy mitochondria (Narendra et al. 2008; Kahle, Waak, and Gasser 2009). Early onset PD (age at onset younger than 50) as well as juvenile Parkinsonism (age at onset younger than 20 years) accounts for less than 4% the total PD cases. However, a loss of function in Parkin contributes to an approximated total of 15% of the sporadic, early onset and juvenile cases (Bonifati 2014). At autopsy, patients that have been identified with *Parkin*–associated PD do not have LB pathology, but significant nigral neuronal loss is present; on the other hand patients with compound heterozygous mutations (these patients therefore have two different disease associated alleles at a specific locus) have been documented to carry LB or tau pathologies (van de Warrenburg et al. 2001; Bonifati 2014). As yet, there has been only one documentation of *PINK1*–related PD with LB disease

whereas the pathology for *DJ-1*-related PD has yet to be determined (Trinh and Farrer 2013). Mutations (either compound heterozygous or homozygous) which result in autosomal recessive forms of the disorder can be identified most commonly in *Parkin* (Kitada et al. 1998; Abbas et al. 1999), intermittently in *PINK1* (Rohe et al. 2004; Trempe and Fon 2013) and seldom in *DJ-1* (Bonifati et al. 2003; Annesi et al. 2005).

ATP13A2

The gene *ATP13A2* is an infrequent cause of PD and was first reported in 2006 when mutations were identified in Chilean and Jordanian families that had been reported to have Kufor Rakeb Syndrome (KRS) (Ramirez et al. 2006). KRS is significant as it is a form of autosomal recessive PD, which has a significantly lower age at onset and more extensive neurodegenerative features, which include dementia (Ramirez et al. 2006; Vilariño- Güell et al. 2008; Bras et al. 2012). The protein encoded by this gene is relatively large and is comprised of 1 180 amino acids spanning ten transmembrane domains that are located in the lysosomal membranes (Ramirez et al. 2006; Vilariño- Güell et al. 2008). ATP13A2 belongs to the P-type superfamily of ATPases that are directly involved in the conveyance of substrates (some of which include inorganic cations) across the cell membrane (Fan et al. 2013). The protein is universally expressed and is also found in the brain – with the highest levels identified in the SNpc. Remarkably, the protein has also been reported to be up-regulated in late-onset sporadic PD patients (Vilariño- Güell et al. 2008; Fan et al. 2013).

SNCA

SNCA was the first gene to be directly linked to PD, thus paving the way for further investigation into the genetic aetiology of PD (Polymeropoulos et al. 1996). *SNCA* encodes a small, 140 amino acid protein called α-synuclein that is composed of three major regions: a C-terminal region, an amphipathic N terminal region and a non-amyloid B component domain (Fortin et al. 2004; Bisaglia et al. 2009). α-Synuclein is a member of the synuclein family and is one of three proteins that are structurally related to one another. The additional proteins that can be found in this family include β- Synuclein (implicated as an antagonist to α- Synuclein) and γ- Synuclein (implicated in neurodegeneration as well as cancer) (Surguchov and Jeon 2008; Devine et al. 2011). All proteins belonging to the synuclein family are small and soluble and are expressed in neural tissues. Structurally, these

molecules have two noteworthy characteristics - the presence of a degenerative, repetitive KTKEGV motif throughout the first 87 residues as well as acidic stretches throughout the C-terminal region (Surguchov and Jeon 2008). α-synuclein is a notable protein as it is the major component of LBs. Linkage analyses - the classic study of genetic markers and recombination events in pedigrees with multiple effects – have associated point mutations as well as genomic multiplications (duplications and triplications) with familial, late onset PD (Chartier-Harlin et al. 2004; Devine et al. 2011). It should be noted, however that patients harbouring *SNCA* whole gene duplications or triplications lead to prominent LB formation, earlier onset and dementia (Devine et al. 2011).

LRRK2

PD associated mutations in *LRRK2* result in the development of autosomal dominant forms of the disease. This gene encodes a large, multi-domain protein that is 2 527 amino acids in length (Zimprich et al. 2004). *LRRK2* is composed of six domains namely the mitogen-activated protein kinase kinase kinase (MAPKKK), Ras of complex proteins (ROC), armadillo domain (ARM) carboxy terminal of ROC (COR), ankyrin repeat domain (ANK) and a leucine-rich repeat domain (LRR). The LRRK2 protein has been studied in depth and has very well defined GTPase and kinase functions within the cell (Anand and Braithwaite 2009). Moreover, LRRK2 possesses multiple roles in autophagy, immunity, neurotransmission and endocytosis (Cookson 2012). To date, there are seven PD-associated mutations in *LRRK2;* these mutations include N1437H, R1441C/G/H, Y1699C, G2019S and I2020T and patients with *LRRK2* mutations have a clinical presentation of idiopathic PD (Cookson 2012). What is most interesting about patients that carry *LRRK2* mutations, is the fact that in many of the cases, patients have some form of LB disease or at the very least, neurofibrillary tangle pathology coupled to gliosis and nigral neuronal loss (Trinh and Farrer 2013). This is of particular interest to researchers as LBs and Lewy Neurites (LN) are by definition the pathological trademarks of PD, but these abnormal aggregations are largely comprised of α-synuclein. This then challenges the doctrine that pathogenesis should be defined according to end-stage neuropathology (Trinh and Farrer 2013).

EIF4G1

EIF4G1 is one of the most recent genes that has been implicated in autosomal dominant PD with LB disease. *EIF4G1* produces a protein product of 1 396 amino acids in length which is an active component of the multi-subunit protein complex EIF4G1 that expedites the recruitment of mRNA to the ribosome (Siitonen et al. 2013). A dominantly inherited point mutation R1205H, has been linked to late onset PD (Chartier-Harlin et al. 2011). It should be noted, however, that several unaffected carriers of this mutation have been identified – it is hypothesised that this may be due to reduced or incomplete penetrance (Chartier-Harlin et al. 2011; Trinh and Farrer 2013). For this reason, the role of *EIF4G1* in PD remains unclear and further studies to support or disprove the hypothesis of its role are therefore necessary. Numerous studies have subsequently been conducted in various global PD cohorts in an attempt to provide supportive evidence that *EIF4G*1 mutations are involved in PD (Lesage et al. 2012; Tucci et al. 2012; Siitonen et al. 2013; Blanckenberg et al. 2014; Nishioka et al. 2014). Each of these studies failed to find an association between variations in *EIF4G1* and PD. In a large-scale meta-analysis of genome-wide association studies (GWAS) using a custom designed genotyping array NeuroX, three of the known sequence variants in *EIF4G1*, namely R1205H, R1197W and A502V were assessed (Nalls et al. 2014; Nichols et al. 2015). Here, a total of 6 249 PD patients and 6 032 control individuals were screened using the array. The data revealed an excess of the heterozygous R1205H variant as it was present in five control individuals compared to one PD patient, thereby suggesting that this variant is a benign polymorphism as opposed to a mutation. Moreover, the A502V variant was identified in a heterozygous state in one control and five PD cases and the R1197W variant was not identified in any cases but was found in a heterozygous state in a single control individual (Nichols et al. 2015). For these reasons, it has been concluded that variations in *EIF4G1* are not a cause of PD.

VPS35

VPS35 encodes a 796 amino acid residue known as vacuolar sorting protein 35 (VPS35). VPS35 plays an essential role in the retromer system that mediates intracellular retrograde transport of endosomes to the trans-Golgi network. The discovery of the D620N mutation in *VPS35* is noteworthy as it was the first gene implicated in PD using next generation sequencing (NGS), more specifically whole exome sequencing (WES). The mutation was first identified in a Swiss kindred with autosomal dominant late-onset PD (Vilariño-Güell et

al. 2011). WES performed on an affected pair of first degree cousins identified the mutations (Vilariño-Güell et al. 2011) and subsequently GWAS performed on 4,326 PD patients and 3,309 unaffected controls; only four additional patients were identified as carriers of the novel variant in *VPS35*. None of the controls were found to carry the *VPS35* variant thus identifying it as a novel disease-causing mutation in PD (Vilariño-Güell et al. 2011). Not only did the discovery of *VPS35* provide significant insights into PD aetiology, it also highlighted the effectiveness of WES in novel gene discovery for complex diseases such as PD.

CHCHD2

The *CHCHD2* gene encodes a small protein of 150 amino acid residues known as the coiled-coil-helix-coiled-coil-helix domain-containing protein 2 (Funayama et al. 2015). It is a small protein that is localised to the mitochondria thereby providing evidence that CHCHD2 may fit into the disease-related network that is associated with PINK1, Parkin and DJ-1. A novel missense mutation was identified in the *CHCHD2* gene in a Japanese family with autosomal dominant PD. Through the use of WES, whole genome sequencing and linkage, a heterozygous T61I mutation was identified in the *CHCHD2* gene as the possible cause for disease. Moreover, Funayama and colleagues then screened a total of 341 patients with familial PD and 517 with sporadic PD as well as 559 control individuals. Three additional families were identified as carriers of *CHCHD2* mutations; one family carried the same T61I mutation, a family with R145Q mutation and a family with a splice site mutation (300+5G>A). The two families that carry the same mutation were found to be unrelated and this mutation arose independently in each family.

1.4 Pathways implicated in PD

Neurodegeneration requires an alteration in neuronal structure as well as a change in function. Disease modification and neuroprotection to decrease and possibly even stop PD progression and hence provide a cure, requires a detailed understanding of PD pathogenesis as well as the molecular aetiology of the disease (Trinh and Farrer 2013). In PD, as is the case with most brain disorders, genetic analysis of blood samples provides a non-invasive and unbiased means by which to identify genes and pathways that can be targeted in the disease.

1.4.1. Mitochondrial dysfunction and oxidative stress

As previously discussed in section 1.3, the initial identification of MPTP and its effects led some researchers to develop an opinion that PD is a result of environmental stimuli due to Parkinsonian features presented by some heroin addicts (Langston 1983). However, the discovery of MPTP simultaneously highlighted the role of mitochondria in PD. The active metabolite of MPTP is 1-methyl-4-phenyl-pyridinium ion (MPP^+) and it is selectively transported into the dopaminergic neurons, thereby causing irreparable damage to these neurons. Interestingly, MPP^+ is an active inhibitor of mitochondrial complex I (Nicklas 1987) and the inhibition of this specific mitochondrial complex is directly related to an increase in free radical generation such as reactive oxygen species (ROS). Free radical generation results in an increase in oxidative stress through changes in the electron transport chain (Schapira et al. 1997; Schapira 2010). This discovery is of relevance to PD as some studies have shown that PD patients have significantly lower activity of complex I but that this lack of activity is not due to levodopa treatment administered to the patients (Mann et al. 1994; Haas et al. 1995; Cooper et al. 1995).

Oxidative stress results in significant damage to numerous cellular structures, both intra and extra-cellular as well as major damage to nucleic acids and proteins – because of the excess ROS that is produced (Storz and Imlayt 1999). Increases in ROS within the cells are beneficial to the immune system and may play a role in cell signalling (Zhou, Ma, and Sun 2008). However, it is important that ROS levels are carefully maintained within the cell or damage may occur – if ROS levels increase to beyond a certain point, the cells can no longer neutralize and eliminate them from the targeted cells, thereby causing structural damage to the cells as well as causing damage to DNA, lipids and proteins (Zhou, Ma, and Sun 2008).

Research conducted on transgenic mice suggests that an overexpression of α-synuclein significantly impairs mitochondrial function and may heighten the toxicity of MPTP as the levels of oxidative stress within the cell increase (Song et al. 2004). The protein products of *PINK1*, *Parkin* and *DJ-1* all interact with one another during oxidative stress – Parkin associates with the outer mitochondrial membrane, where it prevents the activation of caspases and the release of cytochrome c (Darios et al. 2003). DJ-1 translocates to the mitochondrial intermembrane space and matrix where the PTEN-tumour suppressor protein is down-regulated thereby protecting the cells against oxidative stress induced apoptosis (Kim et al. 2005). Finally, PINK1 is capable of localising to the mitochondrial matrix and is

hypothesised to protect against apoptosis (Petit et al. 2005). The knowledge gained through the identification and analysis of each of these genes strengthens the importance of mitochondrial dysfunction and oxidative stress as a vital mechanism in PD pathogenesis.

1.4.2 The Ubiquitin-Proteasome System

The UPS is a pathway that is conserved from yeast to mammals and is necessary for the degradation of most short-lived proteins (cytosolic, secretory and membrane) in the eukaryotic cell (Hershko and Ciechanover 1998). Some of the targets of the UPS include cell regulatory proteins, whose judicious destruction is essential for controlled cell division as well as proteins that are unable to fold properly within the endoplasmic reticulum. Other networks on which the UPS functions include cell cycle regulation, cellular differentiation and cell development, morphogenesis of neuronal networks, intra-cellular stress responses and extra-cellular effectors and most importantly, DNA repair (Glickman and Ciechanover 2002; Dawson and Dawson 2003). In short, the purpose of the pathway is to tag proteins with ubiquitin so that they can be recognised by the 26S proteasome for degradation.

Parkin plays a pivotal role in the UPS. Parkin belongs to the E3 ubiquitin ligase family due to the fact that it has an in-between-ring domain. This domain is important as it is the region that interacts with the ubiquitin–conjugating enzymes (E2) and catalyses the attachment of ubiquitin molecules to specific protein targets (Moore et al. 2005). This process allows for 'ubiquitin tagging' to take place in order to specify the destruction of specific proteins by the proteasome (Shimura et al. 2000). Ubiquitination results from the consecutive actions of the ubiquitin activating E1, E2 and E3 enzymes. Subsequent cycles of ubiquitination result in the formation of a poly–ubiquitin chain that can then be recognised by the 26S proteasome (Moore et al. 2005). E3 ubiquitin ligases provide substrate specificity to the ubiquitination process as each ligase binds to specific subsets of proteins (Figure 1.5). Defects in Parkin may therefore interfere with the proteolytic pathway that could lead to the deleterious accumulation of particular proteins, in turn contributing to the death of nigral neurons (Matsumine et al. 1997; Kitada et al. 1998). The tagging of proteins with ubiquitin may also occur for processes that are proteosome-independent: some of these roles include signal transduction and protein trafficking (Kahle and Haass 2004). Moreover, it has been established that Parkin is associated with mitochondrial DNA in a neuroblastoma cell line as well as in cells that are undergoing proliferation (Rothfuss et al. 2009). The conclusions

reached through various studies are that Parkin protects the mitochondrial DNA from oxidative damage and may act to stimulate mitochondrial repair (Rothfuss et al. 2009; da Costa et al. 2009). Parkin acts together with PINK1 in a pathway which promotes the maintenance of mitochondrial functioning and integrity (Rothfuss et al. 2009).

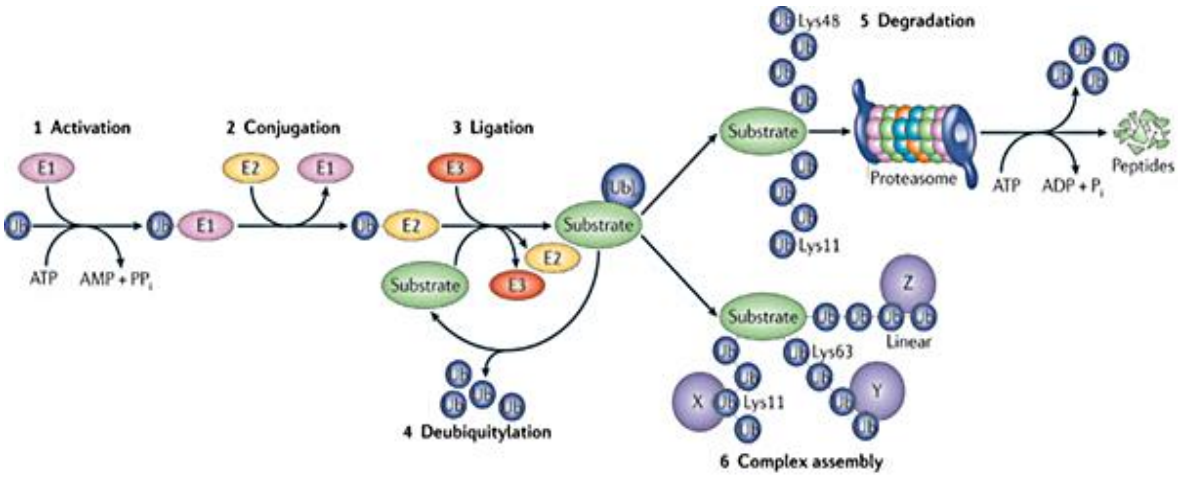


Figure 1.5 The Ubiquitin Proteaseome System. The ubiquitylation and degradation of substrate proteins occurs by a series of reactions that are controlled by the enzymes of the ubiquitin–proteasome system (UPS). During the activation reaction, ubiquitin is transferred to an E1 enzyme in process that is ATP dependant (step 1). The activated ubiquitin is subsequently transferred to an E2 enzyme in the conjugating reaction (step 2). The E2 enzyme then carries the ubiquitin to the E3 enzyme – this is known as an ubiquitin ligase (step 3). The E3 is important not only because it covalently ligates ubiquitin to Lys residues on the substrate protein, but also because it mediates substrate specificity. This process of ubiquitin ligation may be repeated with a Lys of the ubiquitin protein itself serving as the substrate, which leads to the formation of a polyubiquitin chain on the target protein. Deubiquitylating enzymes may reverse substrate protein ubiquitylation (step 4). Ligation of polyubiquitin has diverse biological consequences for the recipient protein. For example, Lys11- and Lys48-linked polyubiquitin chains serve as tags to target substrate proteins for proteasomal degradation (step 5). Conversely, linear, Lys63- and Lys11-linked chains promote the assembly of signalling complexes (step 6). X, Y and Z indicate ubiquitin-binding proteins. Pi, inorganic phosphate; PPi, inorganic diphosphate; Ub, ubiquitin (taken from Vucic, Dixit, and Wertz 2011).

1.4.3 The Autophagy-Lysosomal Pathway

The autophagy-lysosomal pathway (ALP) is another system by which unwanted proteins are removed from the cell. The ALP can be divided into three different pathways, each of which is based on the substrates that will reach the lysosomal lumen: macroautophagy, microautophagy and chaperone-mediated autophagy (CMA) (Figure 1.6) (Cuervo et al. 2005; Levine, Mizushima, and Virgin 2011). The induction of autophagy can occur within relatively short periods of nutrient deprivation, CMA is the result of prolonged nutrient deprivation while the induction of microautophagy is not dependent on any form nutritional deprivation or stress (Pan et al. 2008).

In contrast to the UPS, macroautophagy (the most inducible pathway) is hypothesised to be the primary mechanism by which entire organelles such as mitochondria are recycled. Large membrane proteins and complexes that are unable to pass through the tapered proteasome barrel can thus be degraded (Cuervo et al. 2005; Levine, Mizushima, and Virgin 2011). ALP dysfunction may result from the failure of autophagosome formation or autophagosome fusion with lysosomes, dysfunction of molecular chaperones or lysosomal membrane receptors or deficiency of enzymes in the lysosomes (Pan et al. 2008). Moreover, it has been shown that the ALP clears α-synuclein from cells (Cuervo et al. 2005; Levine 2005) and abnormal functioning of the ALP could therefore lead to the toxic accumulation of this protein and subsequently neurodegeneration. The role of the ALP in diseases such as PD has been strengthened by the fact that mutations in the *ATP13A2* gene lead to insufficient protein degradation (Ramirez et al. 2006; Pan et al. 2008). ATP13A2 is a lysosomal protein and it is thought to be responsible for cation transport and the regulation of manganese levels. Loss of ATP13A2 levels result in lysosomal dysfunction with an accumulation of lysosomes and autophagosomes as well as a decrease in proteolytic activity (Manzoni and Lewis 2013)

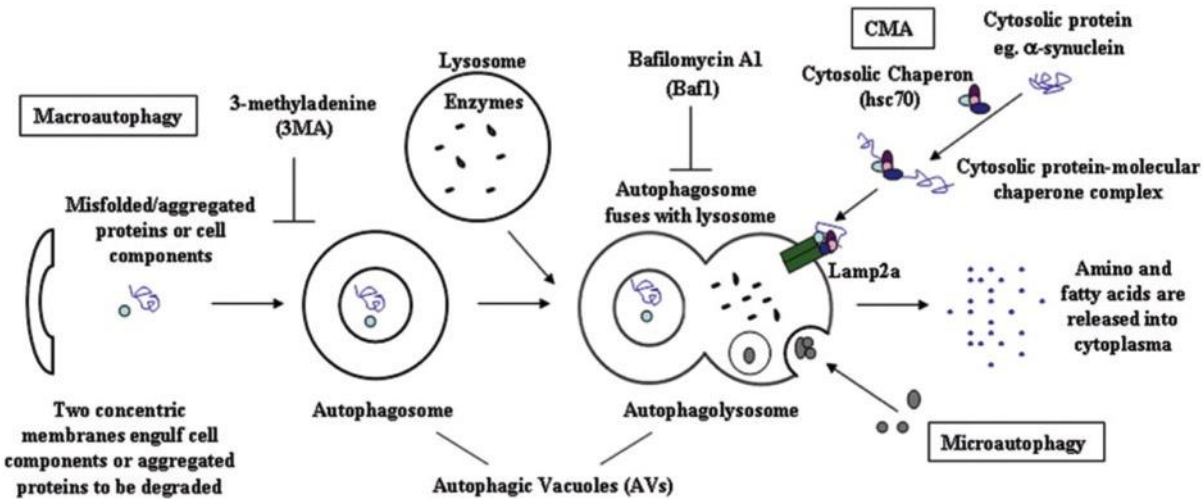


Figure 1.6 The Autophagy Lysosomal Pathway. The macroautophagy pathway is the most common pathway by which cytosolic proteins and cellular components are degraded. Inhibition of the autophagosome formation by 3-methyladenine (3MA) without affecting ATP levels or protein synthesis, or the inhibition of the fusion of the autophagosome with the lysosome by bafiloycin A1 (BafA1) may lead to dysfunction of macroautophagy. In microautophagy, the lysosomal membrane will deform so as to engulf the cytosolic substrates. Specific cytosolic proteins that can be identified by cytosolic chaperone, the heat-shock cognate protein of 70kDA (hsc70) which targets them specifically to the surface of lysosomes, are degraded through the CMA pathway (taken from Pan et al. 2008).

1.5 Next generation sequencing and whole exome sequencing

Since this PhD thesis focuses on the use of WES as a means for novel gene discovery for PD, the next section with provide an introduction to this methodology. The discovery of dideoxy nucleotides for the use in the "chain terminator sequencing method" by Sanger marked a massive breakthrough in the history of DNA sequencing (Cawley 2005). This concept revolutionized DNA sequencing and drove the development of *automated* Sanger sequencing – the choice method for DNA sequencing over the course of about 20 years. Throughout this time, the technological advancements have allowed for longer fragments of DNA to be sequenced as well as a greater degree of parallelism – interestingly, the technology that currently exists supports the concurrent sequencing of 1000 base pairs per DNA fragment in 96 capillaries (Cawley 2005; Metzenberg 2008). Although Sanger sequencing provided a means by which to analyse DNA, this approach has not allowed for the analysis of DNA in a high throughput manner (http://genome.tugraz.at/Theses/AbstractFischerM2010.pdf).

Automated Sanger sequencing was the fundamental principle on which the Human Genome Project (HGP) was based. This project was initiated in 1990 and aimed at determining the full genomic sequence (all three billion base pairs) of the human genome. The project took a total of 13 years to complete. Initial draft results were produced within the first ten years (Lander et al. 2001; Venter et al. 2001) and was completed in its entirety in the following three years (Jasny and Roberts 2003). The HGP was not the only outcome of 13 years of work; numerous spin-off projects have since been developed, but two of the most notable ones are the International HapMap Project and the 1000 Genomes Project. The 1000 Genomes Project focussed on the sequencing of the genomes of at least 1000 individuals of various ethnicities so as to provide a comprehensive resource/reference on genetic variation in humans (http://www.1000genomes.org/about). The International HapMap Project, on the other hand, intended to develop a haplotype map of the human genome, that describes common patterns in the human genome as well as those regions particularly prone to sequence variations (Gibbs et al. 2003).

The HGP, 1000 Genomes Project and the International HapMap Project required a large amount of time and resources and one of the major outcomes of these projects was the conclusion that faster, cheaper and high throughput platforms were necessary (Schloss 2008). This was one of the major contributing factors that led to the development of the National

Human Genome Research Institute (NHGRI) funding scheme, with the goal of reducing the cost of human genome sequencing to US$1000 within ten years (Schloss 2008; van Dijk et al. 2014). This drove the development of Next Generation Sequencing (NGS) technologies, each of which share the following significant improvements: NGS libraries are prepared in a so-called free cell system and therefore bacterial cloning of DNA fragments is no longer required; secondly sequencing output is directly detected and electrophoresis is no longer necessary as the base pair interrogation is performed cyclically and in parallel. Lastly, hundreds of thousands of sequencing reactions can be produced in parallel. The effectiveness of large scale, high throughput sequencing generated large amounts of sequencing data but with a notable drawback. NGS technologies produce short sequencing reads, thereby making assembly challenging and in turn, driving the development of novel alignment algorithms (van Dijk et al. 2014). The advent of NGS has enabled researchers to study biological systems at a level that has never before been possible. Moreover, NGS technologies are an effective strategy for the discovery of genetic causes underlying various disorders; more specifically, those disorders for which a genetic basis was intractable using conventional approaches such as positional cloning and linkage analysis (Bras, Guerreiro, and Hardy 2012; Grada and Weinbrecht 2013; van Dijk et al. 2014) NGS is fast becoming an important technology in basic science and is becoming an reputable tool in translational research (Grada and Weinbrecht 2013). The continual reduction in cost of sequencing as well as the development of standardized methodologies for both alignment and data analysis is making NGS an appealing tool for routine applications, even in small scale laboratories (van Dijk et al. 2014).

Two arms of NGS are whole genome sequencing (WGS) and the aforementioned WES (Hedges et al. 2009). WGS examines the entire genome of an individual – large volumes of data are obtained per single sample analysed, making this form of NGS very complex and costly (Ng et al. 2009; Robinson, Krawitz, and Mundlos 2011). Due to the complexity and volumes of data obtained, bioinformatics infrastructure and proficiency is necessary for data processing; the result being that WGS is beyond the scope of most laboratories (Bras and Singleton 2011).

WES on the other hand, is a so-called targeted sequencing approach, where only approximately 1.2% of the human genome is examined for genetic variation in order to identify potential disease-causing variants. WES involves the sequencing of only the protein

coding regions of the genome (exons), more commonly referred to as the exome. WES is considered to be an effective strategy for novel variant discovery because:

- It has long been hypothesised that most functional variants are in the coding regions of the genome i.e. the variants in these regions are most likely to have a direct effect on the protein (Botstein and Risch 2003; Ng et al. 2009),
- There are approximately 180 000 exons in the human genome and therefore only 30 mega bases (Mb) of the genome need to be sequenced (Ng et al. 2009),
- It has been documented that the most common causes for Mendelian disorders are single nucleotide variants (SNVs) found in coding regions – these are a source for the mapping of complex genetic traits (Horner et al. 2009).

WES has already been used in numerous applications some of which include the elucidation of disorders that are genetically heterogeneous, the identification of molecular defects within single gene disorders and improving diagnostics (Ng et al. 2009; Hedges et al. 2009; Robinson, Krawitz, and Mundlos 2011). Given the region that is examined using WES, it is not surprising that the volume of both the raw and processed data is significantly smaller than with WGS (Pabinger et al. 2013). However, it is important to bear in mind that each sequencing run will identify a large number of SNVs, including single nucleotide polymorphisms (SNPs) as well as insertions and deletions (indels). The number of variants that are identified may lie anywhere between 50 000 and 100 000, of which approximately 90% are hypothesised to already have been recorded in the public online databases (Ng et al. 2009; Bamshad et al. 2011; Ng et al. 2010). In contrast, WGS generates approximately 5 million SNVs per individual and for this reason, it is understandable that WES has become the more favourable approach in the exploration of Mendelian disorders, thereby contributing to the functional annotation of the human genome and providing insights into disease development and mechanisms (Pabinger et al. 2013).

1.6 Whole exome sequencing platforms and bioinformatics

Currently, a formidable challenge is faced when analysing WES data is the difficulty of identifying a single pathogenic mutation amongst the background of polymorphisms and possible sequencing errors that are generated for each sequenced individual. For this reason, it becomes imperative that a detailed understanding of the WES platform as well as the data

obtained from each sequencing reaction is understood so as to aid in post-sequencing variant processing as well as variant prioritization.

1.6.1 Commonly used whole exome sequencing platforms

(i) Illumina Hi-Seq

Illumina, like most NGS platforms, relies on the chain termination method and both the Hi-Seq 2000 and 2500 integrate improvements in engineering that currently produce the highest output available on the market. Before sequencing is conducted, the DNA is sheared into fragments to generate a library and is run on a gel in order to separate each of the fragments based on size. A fragment with an average length between 200 and 300bp is then selected for further replication through PCR. This so-called automated cluster generation is used to distribute the fragment library to the surface of a flow cell amongst a sea of adaptors. Each fragment will then bind to a complementary adaptor and a process known as bridge amplification will then occur – thereby allowing for the generation of copies of a specific molecule to be made on the surface of the flow cell. As each individual base is added, a camera records the location of each cluster through the capture of the fluorescent signal emitted by the cluster. The combination of these images creates the sequence. It should be noted that although the addition and sequencing of a single base at a time seems slow, each flow cell is capable of analysing approximately 150 million of these clusters, thereby making this an extremely efficient system (http://systems.illumina.com/systems/sequencing.ilmn).

Although the Hi-Seq 2000/2500™ is still the most favoured Illumina sequencing platform to use for WES, January 2015 brought exciting news that Illumina had developed two new machines, the Hi-Seq 3000/4000™. These systems are both developed off the success of the Hi-Seq 2000 / 2500™ and include a leveraging pattern flow technology, providing unparalleled sequencing speeds and multiple applications for high throughout sequencing. The development of these improved systems means that the Hi-Seq 4000™ is capable of sequencing up to 12 whole genomes, 100 whole transcriptome samples or a total of 180 full human exomes in 3.5 days or less – thereby currently making Illumina the biggest driver in the field of NGS (http://www.illumina.com/company/news-center/press-releases/press-release-details.html?newsid=2006979).

(ii) Roche 454 Genome Sequencer FLX™

Roche's 454 Genome Sequencer FLX became commercially available in 2005 and is a modification of the Sanger sequencing method – more specifically, Roche managed to significantly simplify the preparation process. The principle behind this platform is the use of custom design fibre chips that house adaptor-flanked fragments in order to hold the primers and polymerase enzymes and start the synthesis of complementary strands. Moreover, the 454 sequencer also employs emulsion PCR amplification that replicates the strands attaching to beads – this is of particular significance as it ensures that the reaction can be detected at a specific light intensity. The sample is then loaded onto a picotiterplate, where the beads enter individual wells. Packing beads are then also added to the plate in order to assist with the spectrophotometric reading of the sample. The result of this method is that the 454 is capable of analysing numerous samples in parallel – a significant improvement on the Sanger sequencing method. The 454 system does, however have a significant drawback as it is incapable of managing homopolymers, thereby producing a significant error rate (http://www.454.com/).

(iii) Applied Biosystems SOLiD System™

The Applied Biosystems SOLiD System is considered to be an extremely flexible system in that it allows for genome sequencing in numerous applications. There are five major primary steps namely (1) enzyme and sample preparation; (2) PCR and substrate preparation; (3) ligation; (4) imaging and (5) data analysis. Interestingly, enzyme and sample preparation are the only samples that need to change based on the desired application. SOLiD uses emulsion PCR, that is similar to Roche, but the fragment library is distributed onto microbeads that can vary in size and richness of slides that they are on. Slides containing one, four or eight sections can be used, based on the required application. Fluorescence is then emitted when each fragment is ligated onto a single strand sequence. Data analysis occurs through Exact Call Chemistry, which relies on an eight base pair interrogation system with four different coloured primers so as to map out possible combinations within the sequence. This system is effective in the detection of SNVs (http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing.html).

(iv) IonTorrent™ by Life Technologies

The Ion Torrent is a long read, high-density semiconductor sequencing platform that was developed by Roche 454 Life Sciences in partnership with DNA Electronics and

is largely considered to be the "new kid on the block" when it comes to NGS. It is based on the detection of hydrogen ions that are released during the polymerization of DNA. As the deoxyribonucleotide triphosphate (dNTP) is incorporated into the DNA strand that is complementary to the template strand, the release of a hydrogen ion triggers an in-sensitive field-effect transistor (ISFET) ion sensor that will record the reaction. The Ion Torrent differs from other NGS sequencing technologies as no modified nucleotides or optics are made use of; rather a single species of dNTP is used per unit time as opposed to all four dNTPs that are used on other platforms. Should there be no complementarity between the dNTP and the template nucleotide, there will be no reaction. It has been estimated that the Ion Torrent has a per base accuracy of 99.6% based on 50 base reads with 100Mb per run, with read lengths averaging 100 bases. One of the disadvantages of the Ion Torrent is that the enumeration of long repeats is a difficult exercise as multiple ions will be released as multiple nucleotides are incorporated, therefore making signal differentiation different (http://www.lifetechnologies.com/za/en/home/brands/ion-torrent.html).

1.6.2 Data analysis strategies

The current achievements in molecular biology specifically pertaining to WES have contributed greatly to the understanding of various diseases, but there have been significant bioinformatics challenges that may limit the efficiency and application of WES as a whole (D'Antonio et al. 2013). One such example is the fact that the interpretation and manipulation of sequencing data that is obtained through WES presents notable computational challenges (D'Antonio et al. 2013; Pabinger et al. 2013). As already mentioned, the first major obstacle to overcome when analysing WES data is the management of the actual volume of data that is obtained through various sequencing platforms. It is estimated that a researcher may be confronted with a vast number of variants that may range anywhere from between 50 000 and 100 000 per sample sequenced, dependant on the sequencing platform that is used. Of these, approximately 10 000 will be predicted as insertions, deletions, splice-site alterations or non-synonymous amino acid substitutions (Clark et al. 2011). Analysis of the raw data can be cumbersome due to the volume of data as well as the read lengths that are obtained. Moreover, erudite informatics tools are necessary and the management and storage of the data files is very often impractical. To date, there is no "gold standard" that can be applied to WES data due to the

array of file formats, software and analytical tools that are available– more importantly, it has become common knowledge that scientists are required to have the necessary skills in computational biology to analyse, mine and interpret the data (Bras, Guerreiro, and Hardy 2012; D'Antonio et al. 2013; Pabinger et al. 2013). For this reason there has been much focus on the development of a streamlined and highly automated pipeline for WES analysis. An example of one such pipeline is illustrated in Figure 1.7.

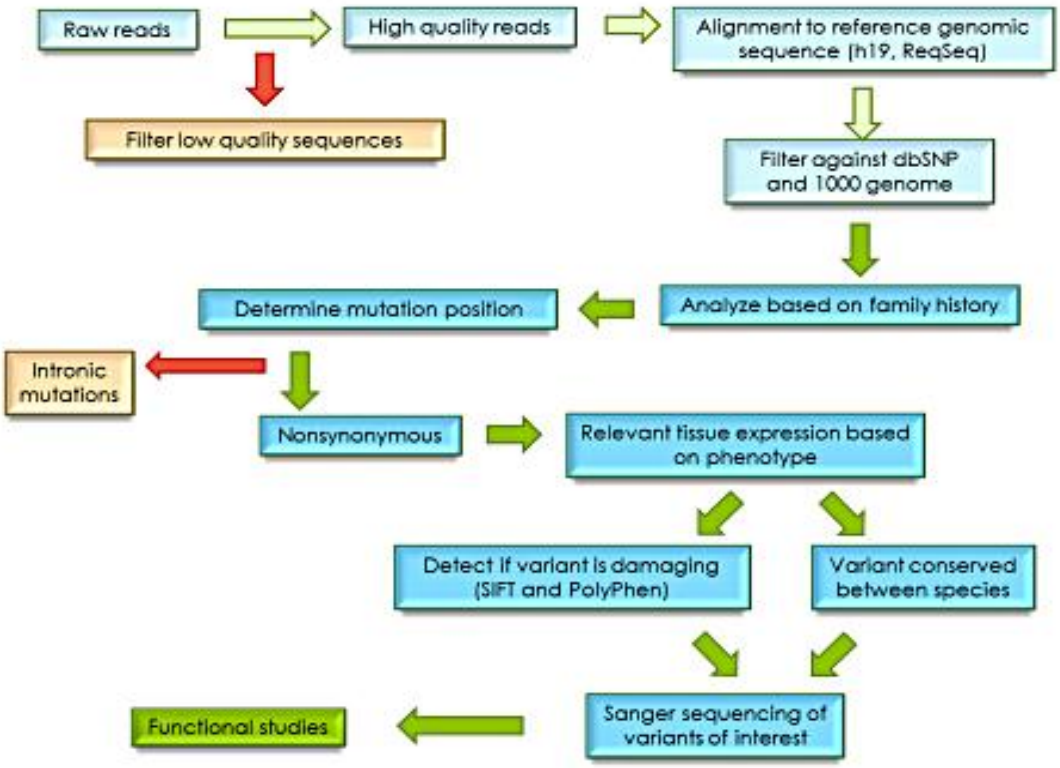


Figure 1.7 Sample pipeline for whole exome sequencing result filtration (taken from http://d-scholarship.pitt.edu/18107/1/WesterfieldL_Thesis_2013etd.pdf).

It should be noted that variants that are consistently found to be common in the general population are unlikely to be attributed to a specific disease. Variants that occur at high frequencies (very often greater than 5%) are found in databases such as dbSNP, the 1000 Genomes Project, and various exome databases. One of the major drawbacks for making use of these databases is the fact that there may be a change or error in information pertaining to specific genes and variants; there are approximately 17 million SNPs that have been recorded in the human genome but the false positive rates for these SNPs is estimated to lie between 15 and 17% (Ku, Naidoo, and Pawitan 2011). Computational algorithms have been incorporated into WES data analysis so as to identify genes that have been consistently shown to have high false positive or negative results and can be applied to the data analysis

so as to narrow the list of candidate genes to a more manageable size (Robinson, Krawitz, and Mundlos 2011). WES variant prioritisation is further supported by numerous additional computational algorithms, most of which will predict the pathogenicity of a particular variant. The two most commonly used tools for pathogenicity predictions are SIFT (Sorting Intolerant From Tolerant) (http://sift.bii.a-star.edu.sg/) and PolyPhen2 (genetics.bwh.harvard.edu/pph2/dokuwiki/start). The removal of SNPs that are predicted to be functionally benign or tolerated allows for further trimming of the list of variants that require further analysis. Once a list of variants of interest has been identified, validation of the selected variants becomes imperative – variants of the greatest interest are confirmed through the use of Sanger sequencing and should the variant be identified as a plausible candidate, functional studies should be conducted on the variant so as to determine the possible physiological effects of said mutation.

1.6.3 Proof of concept: use of WES to identify PD-causing genes

Diseases such as PD are amenable to WES approaches with both rare Mendelian as well as common sporadic forms of the disorder being suitable for this type of analysis (Bras and Singleton 2011). For recessive forms of PD, as few as three individuals may provide significant insight into the disease when using WES; for clearly dominant disorders, as few as four or five individuals may be sufficient to identify novel mutations (Wang et al. 2010; Glazov et al. 2011).

The success of this approach in identifying novel mutations in diseases such as PD has been shown by the identification of a novel gene *VPS35* (vacuolar sorting protein associated protein 35) in a Swiss kindred with autosomal dominant late-onset PD (Vilariño-Güell et al. 2011). WES was performed on an affected pair of first degree cousins (Vilariño-Güell et al. 2011). The NimbleGen Sequence Arrays were used for exonic capture and sequencing performed on the Illumina Genome Analyser and the number of variants identified in each patient was 34,754 and 29,952 respectively. Filtering was carried out using HapMap to filter the results further by eliminating additional polymorphisms. Structural alterations such as CNVs were eliminated using the Database of Genomic Variants (version 6) and a total of 4,265 candidate variants remained. Upon further filtering, where variants found on the X and Y chromosomes as well as synonymous and non-coding variants that were already present in dbSNP (version 130) were excluded, a preliminary candidate list of 69 disease-causing

variants was identified (Vilariño-Güell et al. 2011). Notably, of these, 36 were found to be artefacts using Sanger sequencing, leaving 33 validated variants. Only two variants were identified as novel - namely A1012V found in Integrin alpha X *(ITGAX)* and D620N, found in *VPS35*. Upon further screening of 4,326 PD patients and 3,309 controls, four additional patients were identified as carriers of the novel variant in *VPS35* and none of the patients carried the *ITGAX* variant, but it was identified in one of the controls. None of the controls were found to carry the *VPS35* variant thus validating it as a novel disease-causing mutation in PD (Vilariño-Güell et al. 2011). The use of first-degree cousins and the specific filtering strategy employed was a proof of principle that WES could be used to successfully identify novel PD-causing genes.

Notably, this same gene and mutation was identified by an independent group (Zimprich et al. 2011). They studied an Austrian family and in this case, two second degree cousins were selected for WES under the assumption that any shared rare variants identified in these patients would be plausible disease-causing mutations (Zimprich et al. 2011). Once the sequencing results were obtained and the sequences aligned, the SNVs were identified using dbSNP (version 131). Further filtering made use of SAMtools (version 0.1.7), which eliminated SNVs recorded in dbSNP as well as known indels (Zimprich et al. 2011). This approach resulted in only ten non-synonymous coding variants to be short-listed as candidates, possibly as more distantly-related individuals had been used (second degree cousins) as opposed to the first degree cousins which had been used for the first study (Vilariño-Güell et al. 2011). The D620N change in the *VPS35* gene was observed in all eight patients available for genetic study but was not found in any of the 2,783 controls screened (Zimprich et al. 2011). These two studies provided further evidence that WES is an effective tool that can be used in the identification of novel disease genes even if the filtration processes to identify the mutations differs from study to study.

1.7 Parkinson's disease research in South Africa

Numerous investigations into both the clinical and genetic characteristics of PD have been conducted in the Japanese, European and North American populations but such investigations into the clinical and genetic characteristics of PD in the Sub-Saharan African (SSA) populations are very limited (Okubadejo 2008; Blanckenberg et al. 2013).

There is an increasing urgency in the need to identify novel genes that play a direct role in the development of PD or may contribute directly to the development or protection against the progress of the disease (Farrer 2006; Gasser 2010; Trinh and Farrer 2013). Africa is experiencing a demographic transition; some of these changes include an increase in fertility rates - the population on the African continent is expected to peak at 1.6 billion in 2030, thereby representing 19% of the global population; urbanization. It is hypothesised that there will be considerable migration from rural areas into urban areas which poses formidable challenges pertaining to land access, infrastructure and service delivery; finally, decreases in mortality rates -thus increasing the average life expectancy and increasing the incidence of neurodegenerative disorders such as PD (Africa's Demographic Trends, http://www.afdb.org) There is currently a very limited knowledge about clinical presentations of SSA patients or the genetic aetiology of PD in these individuals; it remains important that mutations in the known PD genes or mutations in novel genes be investigated as the root cause for the disease in this population.

To our knowledge, the PD genetics research group at the Division of Molecular Biology and Human Genetics at Stellenbosch University in Cape Town, South Africa is currently the only group studying the genetic aetiology of PD in South African patients. It has been determined that the known PD genes do not appear to play a significant role in these affected patients (Bardien et al. 2009; Keyser et al. 2010; Haylett et al. 2012).

A total of 458 South African PD patients from diverse ethnic groups have been recruited for genetic analysis (Table 1.2). For the purposes of the study, the PD cohort was split into various ethnic groups. This is due to the fact that there are a number of diseases that may be specifically related to a particular ethnic group. Moreover, these diseases may share overlapping clinical features but the genetic cause for the disease is different (Klein, Schneider, and Lang 2009). In addition, the English-speaking Whites and the White Afrikaans-speaking patients were analysed independently from each other due to the unique ancestry of the Afrikaner.

The various ethnic groups can be defined as follows:

- The English-speaking White population is composed of individuals of European descent.

- The Afrikaner population is unique to South African and is composed only White Afrikaans speaking individuals. These individuals are mainly of Dutch and German decent but also have French ancestral lines (Greeff 2007).
- The Mixed Ancestry population is defined as an admixture or a combination of various ethnic groups. These various combinations include immigrants from Western Europe, India, Malaysia and Madagascar as well as combination of ethnic groups which are indigenous to South Africa, such as San and Khoi–Khoi (de Wit et al. 2010).
- The Black African population is composed of individuals whose ancestry can be directly traced to the African continent. This ethnic group is comprised of individuals who speak traditional African languages such as Zulu, Xhosa, Ndebele, Tsonga, Venda, Swazi, Northern Sotho, Tswana and Sesotho.
- The Indian population is composed of individuals who migrated from colonial India to the African continent in the latter part of the 19th century.

Table 1.2 Ethnic breakdowns of 458 South African Parkinson's disease patients recruited for genetic studies.

Ethnicity	**n (% of 458)**	**Positive family history of PD n (%)**
White (non-Afrikaner)	175 (38.2)	46 (26.3)
White Afrikaner	148 (32.3)	36 (24.5)
Mixed Ancestry	104 (22.7)	16 (15.4)
Black	26 (5.7)	5 (19.2)
Indian	5 (1.1)	2 (33.3)

The numbers of PD patients from each of the ethnic groups and their disease inheritance patterns are illustrated in Figure 1.8. From this, it is clear that for the majority of South African PD patients, the familial inheritance pattern is not known but dominant inheritance patterns appear more common than recessive inheritance patterns. Moreover, 70.5% (323/458) of the cohort are white and of these, 45.8% are Afrikaner.

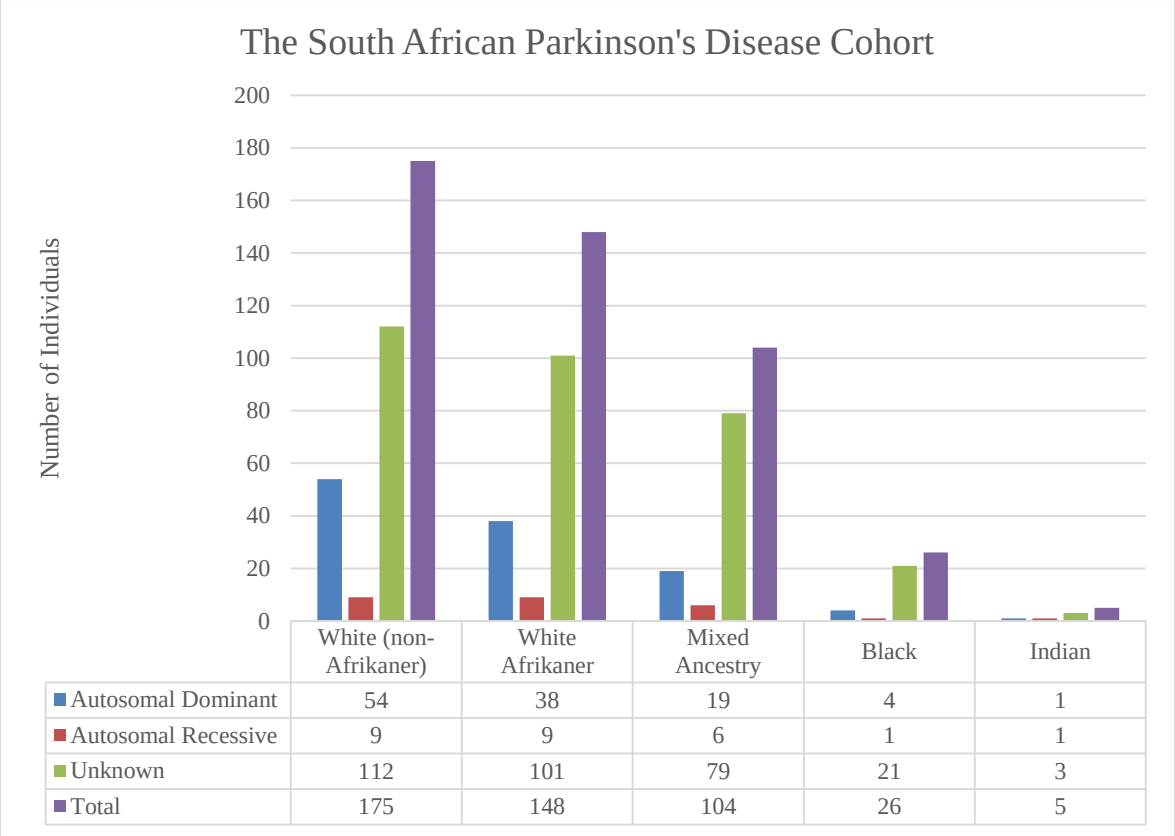


Figure 1.8 Graphic representation of the South African PD patient group according to ethnicity and disease inheritance pattern.

In the following section, a more detailed account of the Afrikaner will be provided, as this particular ethnic group is the focus of the present study.

1.7.1 The South African Afrikaner population

The South African Afrikaners are a unique group of individuals that are mainly descended from Western Europeans who came to settle on the southern tip of Africa from the middle of the 17th century (Greeff 2007; http://www.sahistory.org.za/people-south-africa/afrikaans). The Dutch East India Company (DEIC) established a refreshment station in Table Bay, which is today known as Cape Town. In 1657 executives from the DEIC were allowed to retire from the Company's service and become "Free Burghers" (independent farmers) – these retirees were mainly of Dutch and German descent – and subsequent settlers in the Cape (Heese, 1971). Moreover, in 1688 a group of French Protestants, who were intent on obtaining religious freedom, fled from France and also settled in the Cape. For this reason, the French, German and Dutch are considered to be the forefathers of the Afrikaner nation. It is estimated that between 1652 and 1806, approximately 4000 emigrants had arrived at the Cape of Good Hope. Interestingly, according to J.A. Heese by the year 1867 the 'Afrikaners'

were composed of a mixture of Dutch (34.8%), Germans (33.7%), French (13.2%), People of Colour (7%), British (5.2%), Unknown origin (3.5%) and Other Europeans (2.6%) (Heese, 1971).

The colonization of the Cape by the British resulted in numerous consequences, one of which was the drive for the Afrikaners to become an independent entity. It is understood that this is what led to the Great Trek, a north eastward and eastward migration away from British control into the interior of South Africa – over a period of 18 years (1836–1854). For this reason, groups of Afrikaner settlers became geographically isolated; more importantly consanguinity was common especially in the early generations (Hall et al. 2002). Due to the suggested inbreeding and genetic isolation, the Afrikaners have become a widely used example of founder effects (Botha and Beighton 1983a; Botha and Beighton 1983b). A founder effect results when there is reduced genetic diversity due to the fact that a population is descended from a small number of colonising ancestors. Although in recent years some admixture has occurred in this population, for approximately the first 15 generations, population growth was almost entirely attributed to reproduction as immigration prior to the founding was minimal (Hall et al. 2002).

The demographic account of this population is reflected in the unusually high frequency of specific rare Mendelian disorders which has been recorded to be between 5-10 times higher than in other population groups (Brink and Torrington 1977; Hayden et al. 1980; Botha and Beighton 1983a; Tipping et al. 2001; Hall et al. 2002). In addition to this, this population carries an unusually low allelic diversity at associated loci. At any susceptibility locus, all affected individuals in the specific population may carry a limited set of alleles that are identical by descent from a few common ancestors. Moreover, due to the fact that the origins of the Afrikaner are relatively recent, the chromosomal regions that surround the disease allele are distinctly larger than outbred populations therefore sparse genetic maps are informative (Roos, Pretorius, and Karayiorgou 2009). As a result of this, founder populations are more amenable to linkage disequilibrium (LD) approaches, which is considered to be more informative than traditional linkage for the analysis of complex traits (Karayiorgou et al. 2004) In summary, due to the high LD in the Afrikaner population, they are considered to be important for the identification of genes that are associated with disease.

There are numerous disorders that occur in the Afrikaners at high frequencies because of founder effects. These include long QT syndrome (Brink et al. 2005), Fanconi anaemia (Tipping et al. 2001), pseudoxanthoma elasticum (PXE) (Saux et al. 2002), schizophrenia (Karayiorgou et al. 2004), Huntington's Disease (HD) (Hayden 1980), progressive familial heart block 1 (Brink and Torrington 1977), familial colonic polyposis, porphyria variegate, osteogenesis imperfecta (Knoll, de Vries and de Wet 1988) and familial hypercholestrolemia (FH) (Brink et al. 1987). Studies on FH revealed that it is found at a prevalence of 1 in 100, which is approximately five times more common in a South African Afrikaner population than in populations from Europe and North America. In addition to this, studies on FH identified three founder mutations that could be associated with 95% FH cases in the Afrikaner population (Roos, Pretorius, and Karayiorgou 2009). These mutations were subsequently identified in FH cases from the Netherlands (Defesche et al. 1993) and one of the mutations was specifically shown to have originated in the Netherlands and introduced to South Africa by a single individual (Defesche et al. 1996).

The present day Afrikaners are an identifiable group with a relatively small gene pool and more importantly extremely well kept family records over a period of more than 350 years thereby allowing for accurate historical tracing and examination (Prof. Geldenhuys, personal communication). The family records that are available for inspection include annals of christenings, marriages, deaths and membership records of the Reform Church (Prof. Geldenhuys, personal communication; http://www.gisa.org.za).

1.8 The present study

As it was observed that approximately one third (32.3%; 148/458) of our study participants are Afrikaner, and due to the high incidence of founder effects for other disorders in the Afrikaners, it was postulated that a founder effect for PD may exist in this ethnic group. For this reason, extensive genealogical analysis was conducted on all recruited Afrikaner PD families with a positive family history of the disorder. A genealogist specialising in the Afrikaner (Prof Gerhard Geldenhuys, Department of Applied Mathematics, Stellenbosch University) constructed comprehensive genealogical charts, and it was determined that the genealogical information for most of the families partaking in the study could be traced back at least eight generations. Theoretically, a single ancestral chart could contain 511 individuals, equivalent to the eight generations that are ancestral to the proband.

The complete genealogical trees for six PD affected individuals were constructed and it was determined that there was a single ancestral couple that was common to each of these six families – a so-called founder couple (Figure 1.9). The couple identified was married in South Africa in the year 1668 – the husband was originally from the Netherlands and arrived on South African shores in 1661, while the wife was a German, having arrived in South Africa in the late 1650s (Prof. Geldenhuys, personal communication).

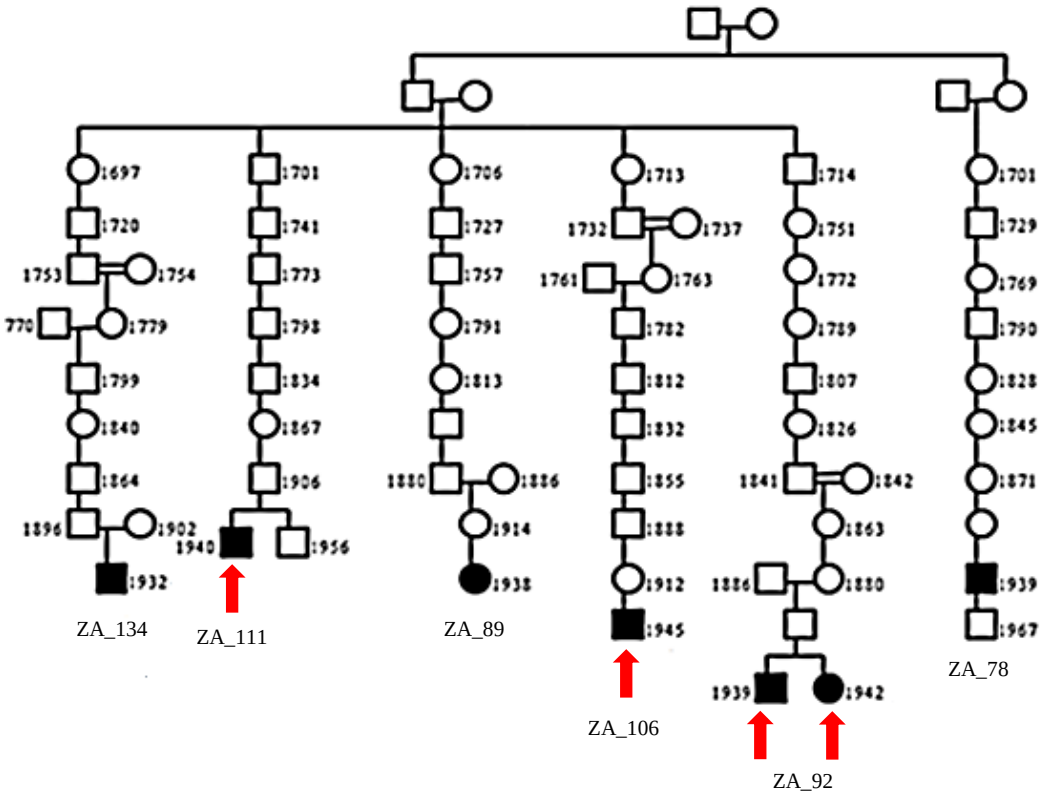


Figure 1.9 Pedigree of the six Afrikaner PD probands shown to be distantly related through genealogical studies. The probands used in this study are coloured in black and the three probands and affected sibling selected for exome sequencing are numbered ZA92_proband, ZA_92_sibling, ZA106 and ZA111 and these are indicated by a red arrow. Males are denoted as squares and females are denoted as circles in the pedigree. Numerical values indicate date of birth and the double lines indicate consanguinity.

1.8.1 Hypothesis

These findings led to the hypothesis for the present study, that these six PD probands share a common pathogenic mutation as the cause of their disease due to a founder effect. Furthermore, it was hypothesised that the gene involved would be a novel gene for PD, as all known PD-causing genes had previously been excluded as the cause in these six individuals.

1.8.2 Aims and objectives

The aim of the present study was therefore to identify a PD-causing gene in the South African Afrikaner population using the WES approach and further to identify whether or not this putative disease-causing mutation could be attributed to the development of PD in other South African ethnic groups.

The objectives of the study are as follows:

1. To identify any additional Afrikaner probands that may be related to the common founder couple as identified.
2. To determine the degree of relatedness of Afrikaner probands may have been linked to a common founder couple using a whole genome SNP array.
3. To identify a novel PD-causing gene in the South African Afrikaner population through the use of WES and bioinformatics approaches.
4. To develop a bioinformatics pipeline for the analysis of data that potentially can be universally applied to any WES project.
5. To determine the frequency of the novel mutation(s) in the PD patients recruited to date from the Afrikaner population as well as the other ethnic groups (English-speaking Whites, Black and Mixed Ancestry).
6. To determine the *in silico* functional effect of the novel mutation(s).

CHAPTER 2: MATERIALS AND METHODS

INDEX **PAGE**

2.1 Study participants 39
2.2 Genealogical analysis 39
2.3. Multiplex ligation-dependent probe amplification (MLPA) 40
2.4 Whole genome SNP array 42
2.4.1 Identification of regions of identity by descent (IBD) 42
2.5 Whole exome sequencing 44
2.6 Bioinformatics analysis 46
2.6.1 Variant prioritisation using the hypothesis-based approach 49
2.6.1.1 Variant prioritisation using Ingenuity Variant Analysis 50
2.6.1.2 Prioritisation of filtered WES results 51
2.6.2 Analysis of WES data using a hypothesis-free approach and the construction of TAPER™ 52
2.7 Sanger validation of prioritised variants 56
2.8 TaqMan® SNP genotyping 56
2.8.1 Real time PCR amplification conditions 57
2.8.2 Allelic discrimination 58
2.9 SNP genotyping using High Resolution Melt 58
2.9.1 HRM real time amplification conditions 61
2.10 *In silico* prediction of prioritised variants 61

CHAPTER 2: Materials and Methods

2.1 Study participants

Ethics approval was obtained from the Committee for Human Research at Stellenbosch University, Cape Town (Protocol number: 2002/C059). A total of 458 PD patients had been recruited from the Movement Disorders Clinic at Tygerberg Hospital in Cape Town, as well as from the Parkinson's Association of South Africa. The patients were diagnosed according to the UK Brain Bank Diagnostic criterion which requires that patients present with bradykinesia as well as at least one of the following symptoms: resting tremor, rigidity and postural instability (Gibb and Lees 1988). All study participants met the criteria. The cohort included 277 (60.5%) male and 181 (39.5%) female patients. The average age at onset (AAO) of the patients was 56.8 years of age. The standard deviation (SD) is 12.7 years and the range of the AAO falls between 13 and 82. A total of 35% of these patients reported a positive family history while 65% could either not provide any information regarding possible family history, or had no known reported history of PD.

Written, informed consent was obtained from each of the patients and a blood sample was taken in order to obtain a DNA sample for the genetic analysis. A total of 690 controls were recruited from the Western Province Blood Transfusion Services, the Geriatric Clinic as well as from other sources. These individuals were not examined for PD, but were used as a means to assess the frequency of specific sequence variants in each ethnic group. The controls were ethnically matched and were made up of 184 white Afrikaners, 160 white individuals, 180 mixed ancestry individuals and 166 black individuals.

2.2 Genealogical analysis

Extensive genealogical analysis was conducted by a genealogist, Prof. Gerhard Geldenhuys, on all recruited Afrikaner PD families with a positive family history of the disorder and an early age at onset (mostly ≤ 60 years of age). Upon the initiation of the genealogical study in 2009, a total of 193 PD probands had been recruited and of these, around one third (62/193) were self-identified as Afrikaner. Subsequently, Afrikaner probands that met the criteria of early onset PD and a positive family history of the disease (at least one first, second or third degree relative that presented with the disease), were subjected to genealogical analysis and six of these individuals were traced back to a common founder couple (B. Glanzmann, MSc Thesis, March 2013, Prof. Geldenhuys, personal communication). In February 2013 a total

of 48 Afrikaner families had been investigated and for each of these, a proband was chosen for whom an ancestral chart could be constructed. Methods for the accurate genealogical tracing of individuals included interviews with the probands as well as their relatives and in depth searches into various sources such as state archives, marriage and baptismal records, death notices and certificates, published genealogies, tombstone inscriptions, voter's rolls and telephone directories, as well as the internet. Moreover, it is well known that the three mainstream Afrikaner churches, the Nederduitse Gereformeerde Kerk (Nether-Dutch Reformed Church), Gereformeered Kerk (Reformed Church) and the Nederduitsch Hervormde Kerk van Afrika (Nether-Dutch Reformed Church of Africa) keep concise and complete records of both marriages and baptisms. Many of these are available in either film or microfiche form at the Genealogical Institute of South Africa in Stellenbosch (Prof. Geldenhuys, personal communication).

2.3 Multiplex ligation-dependent probe amplification (MLPA)

Exonic rearrangements, whole gene duplications and triplications are common in PD patients (Hedrich et al. 2002) but are not detected by techniques such as High Resolution Melt (HRM) analysis or Sanger sequencing. All patients should thus be subjected to MLPA analysis to exclude copy number variations or rearrangements within the known PD genes. Moreover, WES does not identify copy number variations (CNVs) such as duplications, triplications and deletions. CNVs are thought to influence gene expression and can be directly associated with a range of phenotypes and diseases. Also as new genes are discovered, CNVs analysis of these genes should be included in the mutation screening strategies. Two commercially available probe kits namely SALSA P051-B1 and P052-B1 Parkinson MLPA kits (MRC Holland, Netherlands) were used to detect possible copy number variations in the 40 South African Afrikaner probands. Each of the probe kits contains oligonucleotides for the ligation to the exons that are known to cause PD - more specifically to *SNCA, GCH1, UCHL1, ATP13A2, LRRK2, DJ-1, PINK1* and *Parkin* as well as specific reference probes.

Patient and control DNA was diluted to a final concentration of 30ng/µl for each MLPA reaction. Control samples that were included in the MLPA were known samples with no CNVs in the PD genes. Samples were denatured at 95°C for 5 min in the GeneAmp® PCR system 2720 Thermal Cycler (Applied Biosystems, Foster City, CA, USA), and were subsequently cooled for 5 min at a temperature of 25°C. A hybridisation master mix

consisting of 0.75µl of MLPA buffer and 0.75µl probe mix was prepared for each sample (both patients and controls) and mixed gently. Each tube was then placed into the thermocycler for incubation at 95°C for 1 min and then for 16 hours at 60°C. Thereafter the thermocycler was paused at 54°C. The following day, a Ligase-65 master mix was prepared using 1.5µl Ligase buffer A and B respectively, 0.5µl Ligase-65 enzyme and 12.5µl dH_2O for each sample and a total of 16µl of the master mix was added to each sample while in the thermocycler at 54°C. Ligation of the probes to the DNA sample was initiated by running the thermocycler at 54°C for 15 min, followed by heat inactivation of the ligase enzyme at 98°C for 5 min and cooling at 20°C for 5 min.

Finally, the polymerase master mix consisting of 0.25µl SALSA polymerase, 1µl SALSA PCR primer mix and 3.75µl dH_2O was prepared. Sample tubes were removed from the thermocycler after ligation and 5µl of the polymerase master mix was added to each tube. Tubes were returned to the thermocycler for PCR amplification using the following conditions: 30s at 95°C, 30s at 60°C and 60s at 72°C for 35 cycles, followed by 20 min at 72°C and cooling at 15°C for 5 min. Fragment separation by capillary electrophoresis was then performed on the PCR products at the Central Analytical Facility of the Department of Genetics, Stellenbosch University on the ABI 3130xl® Genetic analyser (Applied Biosystems, Foster City, CA, USA). The raw data was then analysed using the Coffalyser.Net software, version .131211 (http://coffalyser.software.informer.com/download/).

Verification of the MLPA results was performed using quantitative PCR (qPCR) on the Lightcycler 96 (Roche Diagnostics, Mannheim, Germany). Primers were available for each gene of interest, as well as HBB (haemoglobin beta), the housekeeping gene that was used in all the aforementioned experiments. Primers were diluted to a final working concentration of 20µM and 30ng/µl DNA was used. A master mix consisting of 0.5µl forward primer, 0.5µl reverse primer, 10µl Lightcycler 480 SYBR Green I Master Mix and 7µl dH_2O was prepared. A total of 18µl of master mix was added to each using the *epMotion™ 5070* (Brinkmann Instruments, Canada) which allows for automated pipetting and preparation of PCR reactions. Thereafter, 2µl of sample DNA was added to each well. All samples and controls were prepared in triplicate. PCR was performed under the following conditions: pre-incubation at 95°C for 10 min, followed by 45 cycles of a three step amplification that included 95°C for 10s, 60°C for 10s and followed by a touchdown to 55°C after the second cycle and 72°C for

10s and finally a melting period of 95°C for 10s, 65°C for 60s and 97°C for 1s – the final step of the reaction was a cooling period of 37°C for 30s. Results were then analysed on the Lightcycler 96 software version 1.1 (www.roche.com), which performs the ΔΔCt method automatically.

2.4 Whole genome SNP array

In order to determine whether or not the Afrikaner probands that traced back to the common founder couple were genetically related, thereby supporting the genealogical data, a whole genome SNP array was performed on all of these patients. This was done using the Illumina® Infinium® Human Core-24 BeadChip (Illumina, San Diego, California, USA) through the collaborative efforts of Professor André Franke at the Institut für Klinische Molekularbiologie (IKMB; Institute for Clinical Molecular Biology) in Kiel, Germany. This is a customizable BeadChip that contains more than 240 000 highly informative genome-wide SNPs and over 20 000 high value markers, including indels. A total of 306 670 markers were screened for in each of the related Afrikaner probands. The whole genome SNP array results were run on the Illumina® Hi-Scan™ System and the results visualized using the GenomeStudio™ Genotyping Module v1.0 (Illumina, San Diego, California, USA). The final output of results to be analysed were in .ped and .map formats for easy manipulation.

2.4.1. Identification of regions of identity by descent (IBD)

The information obtained from the whole genome SNP array was used to identify regions of identity by descent (IBD) in the related Afrikaner probands thereby ascertaining a measure of relatedness in these individuals. The IBD was performed using the open source software package PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/contact.shtml#cite). Quality control was performed on the .ped files according to the following criteria:

1. Calculate the genotyping call rate. The per sample rate is calculated. This is done through the use of the following calculation:
    - (Total number of non-missing genotypes) / (Total number of markers genotyped)
    - It should be noted that a low genotyping call rate is an indication of a sample issue such as low DNA concentration.

- Thresholds may vary between 3 – 7%.

2. Calculate the heterozygosity call rate. The per sample rate is calculated. The following calculation is used:
   - Number of (Total non-missing genotypes (N) – homozygous genotypes (O)) / (Total number of non-missing genotypes (N))
   - Excess heterozygosity could be an indication that there is possible sample contamination while less than expected heterozygosity is an indication that there is possible inbreeding among individuals.
   - Thresholds for the inclusions are approximately the mean ±3 from the standard deviation across all samples.

Genotyping call rate and heterozygosity rate are plotted against each other so as to determine appropriate cut-off values for the minor allele frequencies (MAF), genotyping call rate and overall percentage of data genotyped rather than using arbitrary values. For the purposes of the current study, the following criteria were established:

- Remove all individuals who have less than 95% of data genotyped;
- Remove all individuals who have SNPs that have less than 1% MAF;
- Remove all SNPs that have less than 95% genotype call rate (or greater than 5% genotype error)

Initial investigation into IBD looked only at the original six probands that had initially been related back to the common founder couple. Subsequently, the same quality control was performed on the additional Afrikaner probands. Through the calculation of IBD, PI(hat) scores were generated in order to identify the degree of relatedness between these individuals. In addition to the calculation of PI(hat) scores as well as IBD, segmental sharing was also calculated. High levels of segmental sharing are indicative of relatedness. The identification of segments also allows for specific regions of chromosomal overlap between the related individuals to be identified. In order to identify regions of segmental sharing, PLINK was used and results verified using GERMLINE (http://www1.cs.columbia.edu/~gusev/germline/). The following steps were utilized to identify shared segments:

1. Prune the set of SNPs by performing linkage disequilibrium based SNP pruning. This function recursively removes SNPs within a sliding window. SNPs are therefore pruned on the variance inflation factor (VIF). VIF is a measure of how much

variance of the estimated regression coefficients are inflated when compared to predictor variables that are not linearly related.

2. Identify regions of IBD and segmental sharing. IBD is calculated in order to determine the degree of relatedness across individuals. Subsequently, the degree of segmental sharing is calculated in order to identify which segments of specific chromosomes are identical to one another and therefore further infer IBD as these segments will be identical as a result of inheritance as opposed to sequence similarity. Siblings are expected to share anything between 0-100% of IBD regions – at certain loci the sibling may not share anything, whereas there may be 100% sharing at other loci.

2.5 Whole exome sequencing

Genomic DNA of three Afrikaner patients and one affected sibling belonging to both the original pedigree, which identified the six Afrikaner probands as related to a common founder as well as to the large pedigree of affected individuals, were selected and subjected to WES. The individuals that were selected for WES are shown in Table 2.1 (Pedigree pg 36).

Table 2.1 Afrikaner Parkinson's disease patients selected for whole exome sequencing.

Family ID	Individuals sequenced	Part of the original six pedigree	Reason for selection for WES	Part of the original six pedigree
ZA92	ZA92 (proband) ZA92 (affected sibling) ZA92 (unaffected sibling)	Yes	Proband had both an affected and unaffected sibling	Yes
ZA106	ZA106 (proband)	Yes	Proband had early AAO and positive family history	Yes
ZA111	ZA111 (proband)	Yes	Proband had early AAO and positive family history	Yes

AAO = age at onset

In addition to the three families that form part of the larger pedigree, an unaffected, unrelated control (Control_1) was included for WES as a means to discern which of the variants are present in a control individual that is unrelated to the affected patients and which of the variants may be shared across the sibling pair as a result of inheritance. Three of the probands (ZA92, ZA106 and ZA111) had already been sequenced in the laboratory of our collaborator, Prof. Owen Ross at the Department of Neuroscience at the Mayo Clinic College of Medicine in Florida, USA.

For the purposes of the current study, one of the probands (ZA92) that had already been subjected to WES, was resequenced along with the affected sibling and unaffected sibling for both quality control methods and as a means to exclude common variants across siblings.

WES was performed at Otogenetics Corporation in Norcross, United States. Exome capture was performed using the Agilent SureSelect Human All Exon Kit, a liquid-phase hybridization method that covers 1.22% of the human genome. This coverage includes all known genes, over 700 human miRNAs and over 300 non-coding RNAs, which include small nucleolar RNAs (snoRNAs) and small Cajal body-specific RNAs (scaRNAs). WES was performed using an Illumina Genome Hiseq 2000™, by paired end reads. The input DNA was diluted and the DNA sheared (Agilent Technologies, Santa Clara, California, USA). Samples were purified using the QIAquick PCR Purification Kit (Agilent Technologies) and the quality of the DNA subsequently checked through the use of the Agilent 2100 Bioanalyser™ - DNA quality could be observed in the form of an electropherogram and samples with a distribution peak at a height of 150 ± 10% nucleotides were selected for further analysis. Further purification of the sheared DNA then took place and 'A' bases were then added to the 3' end of the fragments. The samples were then purified through the use of Qiagen MinElute PCR Purification Column (Qiagen, Hilden, Germany). The paired end adaptors were then ligated to the fragments and the samples further purified through the use of the AMPure DNA Purification Kit (Agilent Technologies). An adaptor ligated library was then generated, purified and the quality assessed and at this stage, a minimum of 500ng of library was needed for the hybridization amplification. The sequencing was then carried out following cluster amplification of the library (http://www.chem.agilent.com/Library/datasheets/Public/5990-6319en_lo.pdf).

This design covers approximately 50Mb of the genome with a minimum of thirty-fold redundancy that provides coverage of more than 99% and a concordance of 99.9%. Raw data

was obtained from Otogenetics Corporation in order to perform in-house bioinformatics analysis.

2.6 Bioinformatics analysis

The raw, unaligned sequences were obtained in FASTQ format. Quality control was performed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and aligned to the NCBI Human Reference Genome hg19 using NovoAlign (http://www.novocraft.com/main/page.php?s=novoalign). Due to the fact that the sequence alignment is a computationally expensive endeavour, the alignment for each proband that was sequenced was performed at the South African National Bioinformatics Institute (SANBI) as well as in the laboratory of our collaborators at UBC to ensure concordance – we obtained 100% concordance. Following sequence alignment, the output file is in Sequence Alignment/Map (SAM) format, a text format for the storage of sequencing data in a series of tab delimited ASCII columns. SAM files were subsequently converted to Binary Alignment/Map (BAM) files, which carries the same data as a SAM file, but is compressed, indexed and in binary form. Thereafter, local rearrangements around the indels were performed, duplicates removed using the Genome Analysis Toolkit (GATK), quality scores recalibrated and a Variant Call Format (VCF) file was generated, containing so-called analysis ready reads. The workflow is illustrated in Figure 2.1.

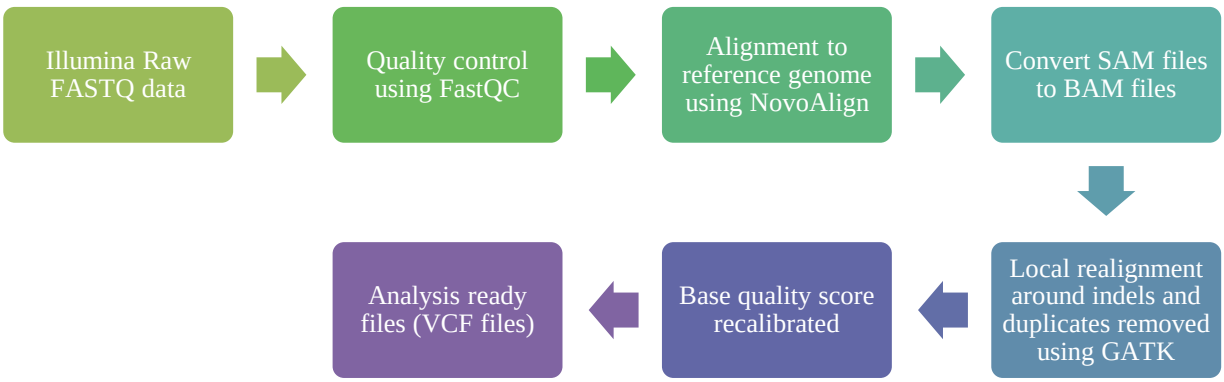


Figure 2.1 Basic workflow for the generation of a Variant Call File (VCF) for further analysis

The VCF files are important as the files themselves store gene sequence variations in such a way that each sequence variant can be analysed in multiple ways. In the case of the current project, the aim was to determine which of the novel variants overlap across all four affected patients namely ZA92, ZA92 affected sibling, ZA106 and ZA111. The VCF files that were generated were subjected to two modes of analyses namely hypothesis based analyses and alternatively, non-hypothesis based analysis (Figure 2.2).

The first method of variant analysis made use of a combination of an in-house or custom method, whereby variants across all four affected individuals were compared and analysed using open source software and basic scripting methods coupled to variants called by commercially available software programs such as Ingenuity Variant Analysis (IVA). The second made use of a custom-designed program called TAPER™ (Tool for Automated selection and Prioritisation for Efficient Retrieval of sequence variants).

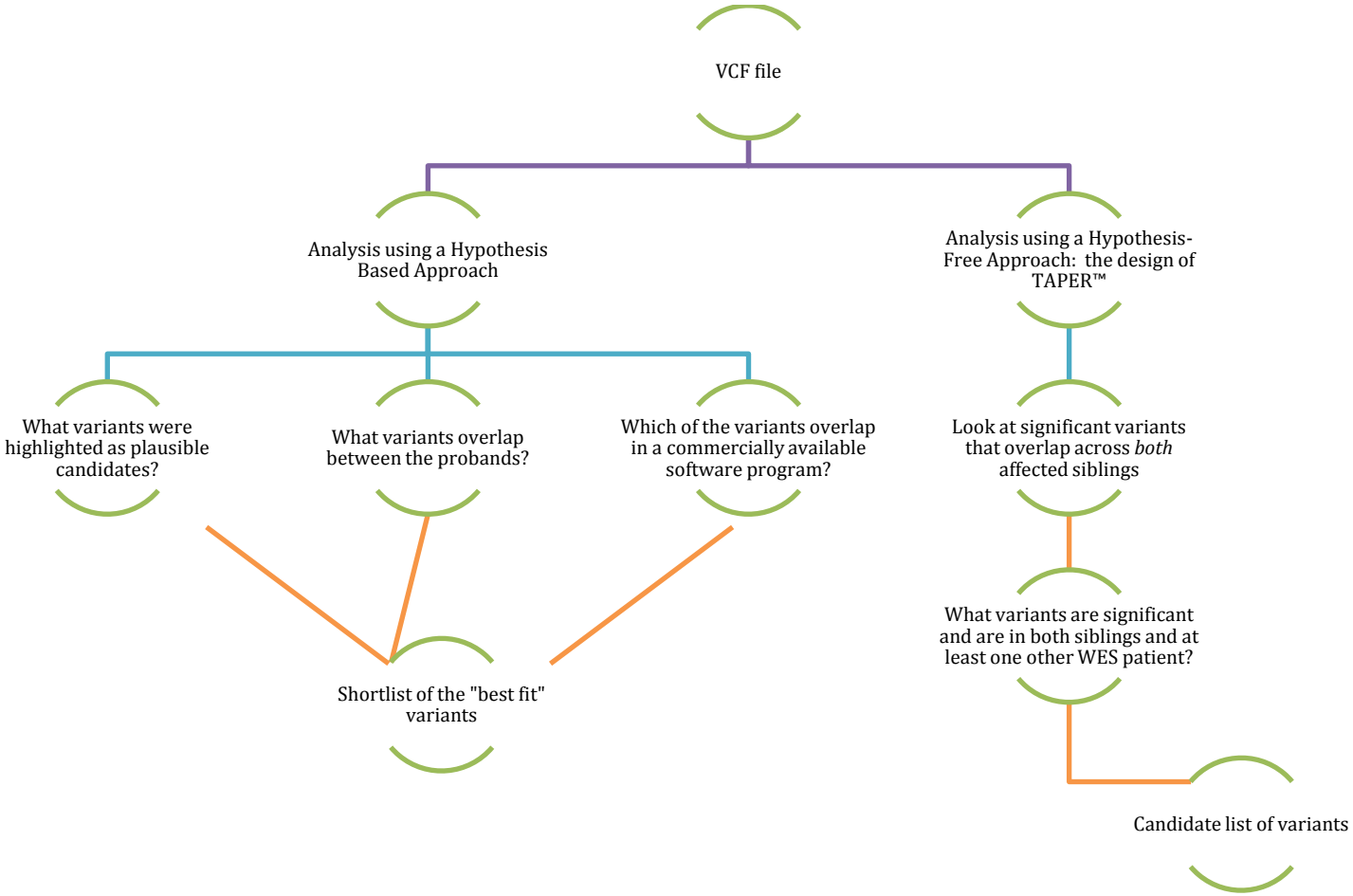


Figure 2.2 Flow diagram of the two approaches used for variant prioritisation

2.6.1 Variant prioritisation using the hypothesis based approach

For the custom-designed bioinformatics analysis pipeline, VCF files were submitted to the web interface of ANNOVAR, namely wANNOVAR (http://wannovar2.usc/). ANNOVAR is a fast and efficient tool that can be used for the annotation of functional consequences of SNPs as well as insertions and deletions that are identified through high-throughput sequencing data. ANNOVAR can therefore be used to generate a shortlist of variants for further analysis by examining the functional consequence of SNPs and indels by examining their functional consequence on genes, reporting functional significance scores, the identification of variants in conserved regions, inferring cytogenetic bands and identifying variants that are already listed in databases such as the 1000 Genomes Project (1KGP), Exome Sequencing Project (ESP6500) and dbSNP. The output of the wANNOVAR analysis is a tab delimited file that was spilt into a number of smaller files according to the following criteria:

- Novel insertions, deletions, frameshifts, substitutions, gain of function and loss of function (in both heterozygous and homozygous forms) – these sequence variations are considered to be novel as there is no record of them in any of the available online databases;
- Known insertions, deletions, frameshifts, substitutions, gain of function and loss of function (in both heterozygous and homozygous forms).

Note that for any of the variations to be considered for further analysis, the following conditions were imposed: no record of the variant in dbSNP, 1KGP or HapMap (therefore novel); if the variant is present, the minor allele frequency (MAF) must be lower than 3%, the read depth must be greater than or equal to 50, call quality must be greater than or equal to 30 and the variant must have predicted functional significance across multiple variant prediction tools.

Following the file split, SNPs of interest were identified. This was done by identifying SNPs of interest that met all of the above mentioned criteria in all of the patients. Finally, in order to identify a shortlist of prioritised variants for further scrutiny, each of the outputs from the individual patients was compared in order to determine which of the variants did, in fact, overlap across all four individuals. A shortlist of prioritised variants was generated for further analysis and comparison with a commercially available prioritization tool, Ingenuity Variant Analysis (IVA) (www.ingenuity.com).

2.6.1.1 Variant prioritisation using Ingenuity Variant Analysis

Variant identification and prioritization is in itself is extremely variable and versatile because of the various ways in which the data can be analysed. For this reason it was decided that a commercially available variant analysis software package be used as a means for comparison so as to determine which of the variants overlap between the two filtration methods. The package used was IVA. VCF files were uploaded onto the server and the following filtering cascade was used for variant prioritization:

1. Confidence – this filter allows for the filtering and subsequent disregard of variants that are of low quality. This is based on variant call quality and read depth. Variants that pass this filter must satisfy all selected criteria. Confidence settings were as follows:
    - Call quality is at greater than or equal to 30 in all probands;
    - Read depth is at greater than or equal to 50 in all probands.
2. Common variants – this filter allows for the inclusion or exclusion of variants that may be commonly observed in a particular population. It is considered extremely valuable for the identification of novel, potentially disease-causing variants as one would not expect such a variant to be present at high frequencies in the general population. Common variants were filtered according to the following criteria:
    - Variants with a MAF of greater than or equal to 3% in 1KGP, Complete Genomics genomes as well as those present in the ESP6500 were excluded – however variants that were present in these databases as well as in dbSNP were included if the MAF was less than or equal to 3%.
3. Predicted deleterious – this filter allows for the identification of variants that have either predicted or observed evidence that may suggest that gene expression and function may be disrupted. The benchmarks for this filter were as follows:
    - Variants that were associated with loss and gain of function and have been experimentally observed to be associated with a specific phenotype were kept for further scrutiny.
4. Genetic analyses – this filter allows for filtration that is based on genotypes as well as inheritance models. Here, the filter can be used to filter and test variants for particular inheritance patterns; the comparison and filtration of variants in one sample set from another (examples include affected vs. unaffected, responders vs. non-responders etc.) and finally for the comparison of frequency of specific variants either to other

samples or to control samples. The criteria for this filtration step was set up as follows:

- Variants that were associated with a loss or gain of function were examined as well as heterozygous, homozygous, compound heterozygous, hemizygous, haploinsufficient or het-ambiguous variants.

5. Biological significance – this filter allows for the selection of biological or clinical concepts and phenotypes that may be of interest to the particular disorder that is examined – genes of interest or known genes can be selected and scrutinized.

2.6.1.2 Prioritisation of filtered WES results

Cumulatively, the shortlist of variants identified through IVA's filtration cascade was compared to that obtained through the custom method. The reason for this was two-fold: in order to determine whether or not the custom filtration cascade could be deemed as a true reflection of results i.e. were the same prioritised variants obtained from both methods, the variant list was compared to that obtained from the commercially available IVA; moreover, the cross platform comparison was used to further whittle down the lists into a workable number of variants as opposed to the investigation of each variant identified independently. Initial analysis of WES results focussed only on the variants that were found to be common in both of the siblings across both filtering methods. Extensive bioinformatics was employed in order to determine the frequency of each variant in the online databases such as dbSNP, 1KGP, HapMap and the ESP6500. Additionally, a total of 20 genes that have been flagged as false positives by numerous WES studies were excluded (Fajardo et al. 2012). This generated a list of variants that are considered to be common in the databases (greater than 5%), which could then be excluded from further analysis and uncommon variants (less than 5%) thus selected for further analysis. Following the prioritization of uncommon variants in the affected siblings, a comparison of variants in all four affected individuals was then done – uncommon variants across all four individuals were identified. Furthermore, theoretical functional prediction tools were used to determine whether or not the selected variants may be predicted to cause disease. The tools that were made use of are summarized as follows:

1. SIFT (Sorting Intolerant From Tolerant) – SIFT predicts whether or not an amino acid substitution will affect protein function. This is based on sequence homology and the actual physical properties of the amino acids. SIFT can be applied to spontaneously occurring non-synonymous polymorphisms as well as laboratory induced missense mutations (http://sift.bii.a-star.edu.sg/);

2. PolyPhen (Polymorphism Phenotyping) – PolyPhen is a software tool that predicts the possible impact of amino acid substitutions on both the structure and function of human proteins but making use of straightforward physical as well as evolutionary comparative considerations (genetics.bwh.harvard.edu/pph2/dokuwiki/start);
3. MutationTaster – this is another web server that predicts the possible effects that a SNP in a particular position may have on protein structure and function (http://www.mutationtaster.org/info/documentation.html). MutationTaster, is however different as pathogenicity scores are calculated within a range of 0 and 1. The closer a variant is to 1, the more likely it is to be pathogenic; the converse is also true: the closer the score is to 0, the more likely the SNP is to be benign and/or tolerated;
4. Project Hope – this is an interactive web server that is capable of analysing the structural effects of a mutation of interest. The server allows the user to submit a protein sequence as well as the mutation. All available information is then accessed and analysed in order to determine the whether or not there is a significant functional effect on the protein (http://www.cmbi.ru.nl/hope/home).

The use of theoretical functional prediction tools allows for the prioritization of variants based on the likelihood that they will affect protein function or folding, thereby moving the specific variant up or down the list for further examination.

2.6.2 Analysis of WES data using a hypothesis-free approach and the construction of TAPER™

Continual identification of sequencing artefacts and high frequencies of variants in control populations as well as no formal, established means by which to filter the data, led to a re-analysis of WES data. A so-called "hypothesis-free" approach was developed in which the approach to variant analysis was aimed at identifying significant variants that were in the affected sibling pair and in at least one of the other two samples that had been subjected to WES. Moreover, for the intention of excluding variants that are non-pathogenic, but that may be familial or possibly commonly occurring in healthy individuals, the two WES controls were considered to be pertinent to the data analysis. VCF files were generated in the same manner as described in Chapter 2, Section 2.6, page 46.

As described previously, VCF files were submitted to the web interface of ANNOVAR, namely wANNOVAR (http://wannovar.usc.edu/). However, following the release of the

updated version of the human reference genome, wANNOVAR was also updated (http://wannovar.usc.edu/). VCF files were submitted to wANNOVAR produced tab delimited files that were very similar to those from the original version, but with additional features such as MAFs for the 1KGP, ESP6500, Exome Aggregation Consortium (ExAC) database as opposed to the conventional dbSNP frequencies. Moreover, the tab delimited file also contained information specifically pertaining to functional and clinical relevance and significance of the variants. Given the volume of the information contained in the wANNOVAR output, it was decided that files would not be split into known or novel outputs, but .csv files would be generated for each patient and control in order to do an overlap comparison at the end of variant analysis.

The generation of the wANNOVAR files produced output files with significant information and an approach that did not possibly exclude the candidate variants had to be developed. The first step in the targeted approach involved the removal of all variants that were synonymous and those that did not result in frameshift mutations. This was then followed by an intricate eight step approach for the prioritization of variants as possibly disease-causing candidates (Figure 2.3). It should be noted that this was not done manually, but a custom design bioinformatics pipeline program, TAPER™ (Tool for Automated selection and Prioritisation for Efficient Retrieval of sequence variants**).** TAPER™ is composed of a number of steps to filter and prioritise candidate variants across individual patients that have been subjected to WES. It was constructed using Microsoft Visual Studio Professional 2013 (Microsoft Corporation, Microsoft Redmond Campus, Redmond, Washington, United States) and additional packages downloaded in order to support the development of TAPER™ included Visual C#, CSV Helper (http://joshclose.github.io/CsvHelper/) and HTML Agility Pack (https://htmlagilitypack.codeplex.com/). It should be noted that TAPER™ has been designed in such a way as to filter all prioritised variants according to a default setting, by which all of the predetermined filtration criteria are implemented or filtered according to a so-called custom method, whereby the user is able to filter the data independently and optionally according to specific parameters. TAPER™ therefore creates a user-friendly environment for a wet bench scientist to analyse and identify variants of interest independently.

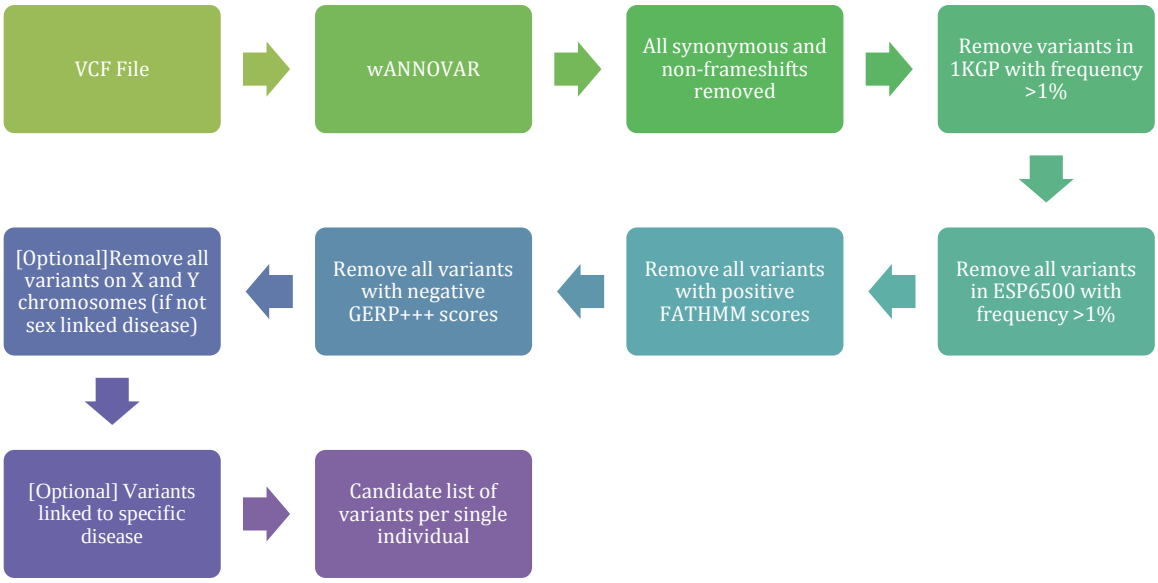


Figure 2.3 A diagrammatical representation of the approach used for the hypothesis-free approach to novel variant discovery and the backbone for TAPER™.

1. The submission of the processed .vcf files to an online variant caller such as wANNOVAR (http://wannovar.usc.edu/), allows for the annotation of functional consequences of genetic variants from high-throughput sequencing data.
2. Removal of all synonymous variants as well as variants that do not cause frameshifts – synonymous variations are defined as codon substitutions that do not change the amino acid and are unlikely to be the underlying cause for rare diseases – for this reason, these variants, along with those that do not cause frameshifts, are removed from the list of prioritised variants.
3. Removal of variants in the 1KGP that are found at a frequency of greater than 1% - any variant that is found in the 1KGP database at a frequency of 1% or less is considered to be rare. It is hypothesised that rare variants are likely to cause disease and for this reason, variants with very low or no available frequency data are prioritised.
4. Removal of variants in the Exome Sequencing Project (ESP) 6500 (http://evs.gs.washington.edu/EVS/) with a frequency of greater than 1% - this step is based on that of the 1KGP data. Rare, possible disease-causing variants are likely to be at an extremely low frequency in this database and any variant with a frequency that is higher than 1% is removed from the list of interest.
5. Removal all variants with positive FATHMM scores - functional analysis through hidden Markov Models (FATHMM) scores are used to determine the species-specific

weightings for the predictions of the functional effects of protein missense variants. The use of FATHMM scores have been shown to outperform the conventional prediction methods such as SIFT, PolyPhen2 and MutationTaster (Rackham et al. 2014). Positive FATHMM scores predict a tolerance to the variation while negative FATHMM scores predict an intolerance to the variation, and is subsequently considered to be pathogenic.

6. Removal of variants with negative GERP+++ scores – Genomic Evolutionary Rate Profiling (GERP) +++ scores are the conservation scores from dbNSFP (database for nonsynonymous SNPs functional predictions); higher scores are indicative of greater conservation; scores of > 0 are considered to be conserved.
7. Remove all variants on X and Y chromosomes – this step is incorporated as an additional, independent step. This is to allow the researcher the freedom to determine whether or not a particular disease has been sex-linked. Should the disease not have previously been identified as a sex-linked disease, the variants may be removed so as to decrease the overall number of candidate variants.
8. Variants linked to a specific disease – the final step of TAPER™ determines whether the genes of interest have been linked to any other disorders using OMIM (Online Mendelian Inheritance in Man) (http://www.omim.org/) database as well as the DISEASES (http://diseases.jensenlab.org) database. If any of these disorders are similar to the disorder under study then that gene and variant(s) becomes top candidates for further study.

The implementation of the new filtering criteria generated a shortlist of potential candidate variants that warranted further examination. Following the prioritization of variants per individual, the following comparative approach was employed: the affected sibling pair was analysed and any variants that overlapped across these two individuals were identified. Following the identification of overlapping variants between the siblings, an additional comparison was performed so as to exclude any variants that were present in either of the control samples (both related and unrelated unaffected controls). Subsequently, the sib pair shortlist of variants was compared to ZA106 and ZA111 independently and three candidate lists of variants were obtained. Candidate variants were therefore present in at least three of the four probands from the original six pedigree that was constructed. Additionally, any

variants that were identified in either of the two control individuals were excluded from further analysis.

2.7 Sanger validation of prioritised variants

NGS technologies such as WES have emerged as extremely effective and powerful tools for the investigation of the genetic aetiology of diseases and the use of such technologies has proven invaluable in both clinical treatment of disease as well as bench research (Patel et al. 2014). Typically WES generates between 50 and 70 million bases of sequence and it is hypothesised that 99.99% of the bases will effectively realign accurately to the reference genome. However, the remaining 0.01% of bases that differ from the reference will be identified as variants (Patel et al. 2014). However it should be noted that the majority of these 0.01% of bases that are called as variants are actually sequencing artefacts as opposed to actual sequence variants. For this reason, each of the prioritised variants were Sanger sequenced in all four of the probands that had originally been subjected to WES as a means to determine whether or not the variants were real or rather a sequencing artefact.

2.8 TaqMan® SNP genotyping

Genotyping was performed in patient and control samples in order to determine the frequency of candidate variants in both patients and controls. A total of 458 patients (all PD probands available at the time of the study) and 690 controls were included for genotyping. The controls were ethnically matched and made up of 184 white Afrikaners, 160 white individuals, 180 mixed ancestry individuals and 166 black individuals. The DNA samples were subjected to TaqMan® allelic discrimination technology using the ABI TaqMan® Custom SNP Genotyping Assay (Applied Biosystems, Foster City, CA, USA). The genotyping was outsourced to a commercial company, IKMB in Kiel, Germany. Each ABI TaqMan® Custom SNP Genotyping Assay consists of two primers in order to amplify the sequence of interest, as well as two TaqMan® MGB (minor groove binding) probes for allele detection. Each probe contains a reporter dye at the 5' end of each allele specific probe (the first allele contains the VIC reporter dye and the second allele probe contains the FAM reporter dye). It should be noted that each probe also contains the MGB as well as a non-fluorescent quencher (NQF) at the 3' end of the probe. The MGB will increase the probe's melting temperature (T_m) without increasing the length of the probe, thereby generating greater differences in T_m values between matched and mismatched probes, thereby improving

allelic discrimination (Beaucage et al. 2001). Figure 2.4 provides a diagrammatic explanation of how detection is achieved with proven 5' proven nuclease chemistry by means of exonuclease cleavage of a 5' allele specific dye label, thereby producing a stable assay signal by removing the effect of the 3' non-fluorescent quencher.

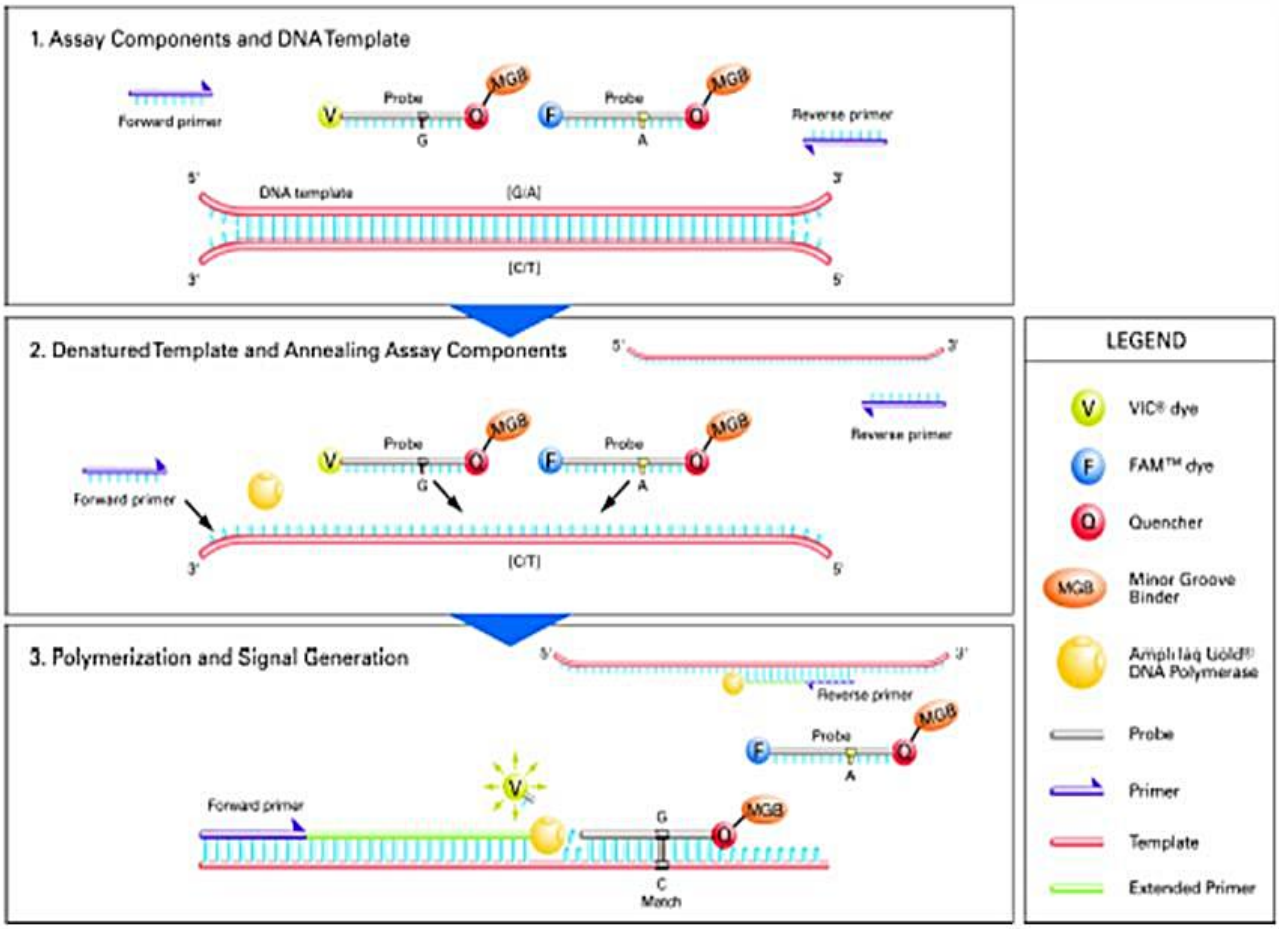


Figure 2.4 Overview of TaqMan® allelic discrimination technology. Selective annealing of the TaqMan® probes as well as the exonuclease cleavage of a 5' allele specific dye label generates the assay signal, thereby enabling allelic discrimination (taken from www.dnavision.com).

2.8.1 Real time PCR amplification conditions

A total of 1148 PD patients and controls were selected and subjected to SNP genotyping at IKMB. A total of three thermostable 384-well plates were prepared, with each well containing 5ng of DNA as per the service provider's instruction. SNP amplification was done by polymerase chain reaction (PCR) in a single reaction tube on an ABI Prism 7900HT Sequence Detection System (Applied Biosystems, Foster City, CA, USA). An EpMotion® liquid handling robot was used to dispense the PCR reagents into the 384-well plates (Eppendorf, Hamburg, Germany). Each PCR reaction consisted of 2.5µl ABI TaqMan®

Universal PCR Master Mix, the stipulated 5ng of genomic DNA, 0.25µl ABI TaqMan® primer and probe dye mix and 1.25µl Dnase-free sterile water so as to generate a total reaction volume of 5µl. Each of the thermostable 384-well plates consisted of a total of 383 samples (either patients or controls) and one non-template control. Each reaction was subsequently subjected to the following PCR conditions: 2 min at 50°C, 10 min at 95°C followed by 40 cycles of 15s at 92°C and 90s at 55°C each.

2.8.2 Allelic discrimination

Allelic discrimination was performed on the ABI Prism 7900HT using the end-point analysis which was carried out using the Sequence Detection System (SDS) 2.4 software that has a 95% confidence level. This software allows for the fluorescence of the samples to be detected and calibrated and subsequently performs automatic allele calling through the generation of allelic discrimination plots.

2.9 SNP genotyping using High Resolution Melt

High Resolution Melt (HRM) is an analytical method in which DNA fragments are distinguished from each other through their melting behaviour. It is an expansion of existing DNA dissociation methods that allow for the characterization of DNA fragments according to the way in which they dissociate ('melt'). Double stranded DNA (dsDNA) (pre-melt phase) is converted to single stranded (ss) DNA (post-melt phase) as it is subjected to increases in temperatures (Figure 2.5). This is analysed or monitored by adding a fluorescent dye (e.g. EvaGreen, Syto 9 and Sybr Green) to the PCR reaction mixture that is allowed to intercalate within the dsDNA of the PCR products. As the strands separate, the dye is released, causing a decrease in fluorescence as the temperature increases. HRM instrumentation collects and analyses fluorescent signals in real time, thereby characterising the different DNA fragments.

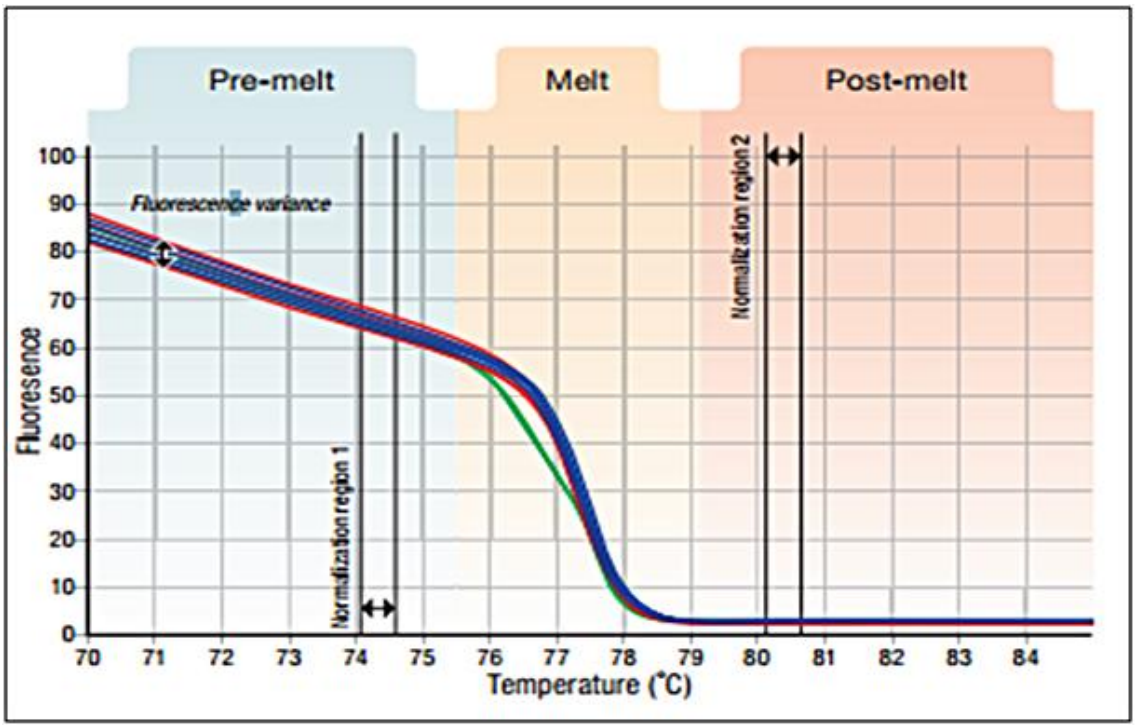


Figure 2.5 Illustration of the principle underlying high resolution melt. With an increase in temperature double stranded (ds) DNA melts and becomes single stranded (ss) DNA. As the melt progresses, an intercalating dye is released - the fluorescence produced is used to create a thermal denaturation profile that is unique for each DNA sequence. As the temperature increases, more of the dsDNA is converted to ssDNA. Fluorescence is plotted against temperature (Taken from Introduction to HRM Analysis http://www.kapabiosystems.com/public/pdfs/kapa-hrm-fast-pcr-kits/Introduction_to_High_Resolution_Melt_Analysis_Guide.pdf).

A thermal denaturation profile can be constructed in which fluorescence is plotted against the temperature; this profile is specific for the PCR product as well as the length of the sequence, base and GC content (Ye et al. 2010). Alterations in the nucleotide sequence will affect the way in which the fragment melts. This will allow fragments with nucleotide sequence alterations to be identified when they are compared to the wild type sample; these can then be sequenced in order to characterize the sequence variant (Ye et al. 2010). HRM data can be analysed as either normalised graphs (Figure 2.6) or as difference graphs (Figure 2.7).

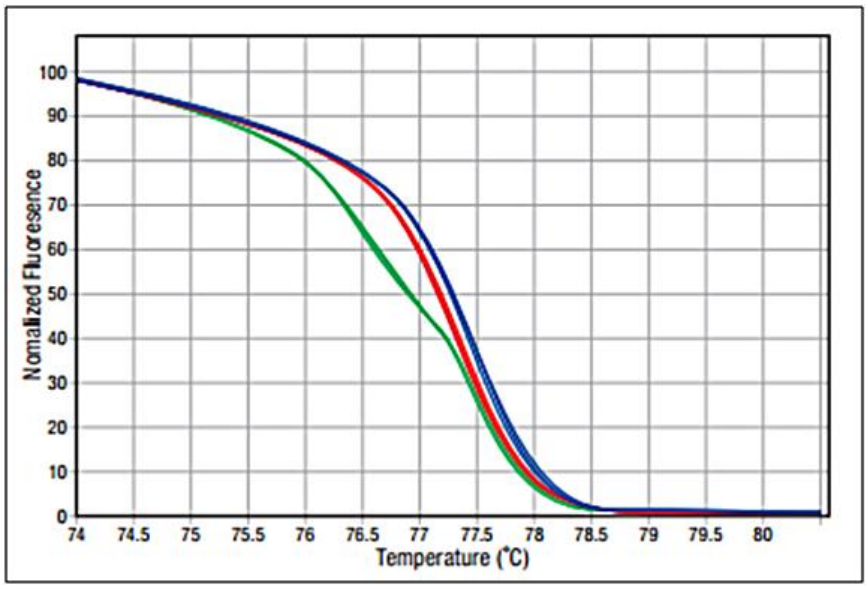


Figure 2.6 Example of a HRM normalised graph. Temperature range in a specific region is selected to allow for identification of variation between different wild type and mutant samples respectively. Blue and red lines are homozygous wild type and homozygous mutant respectively; green line is a heterozygous mutant sample (Taken from Introduction to HRM Analysis http://www.kapabiosystems.com/public/pdfs/kapa-hrm-fast-pcr-kits/Introduction_to_High_Resolution_Melt_Analysis_Guide.pdf).

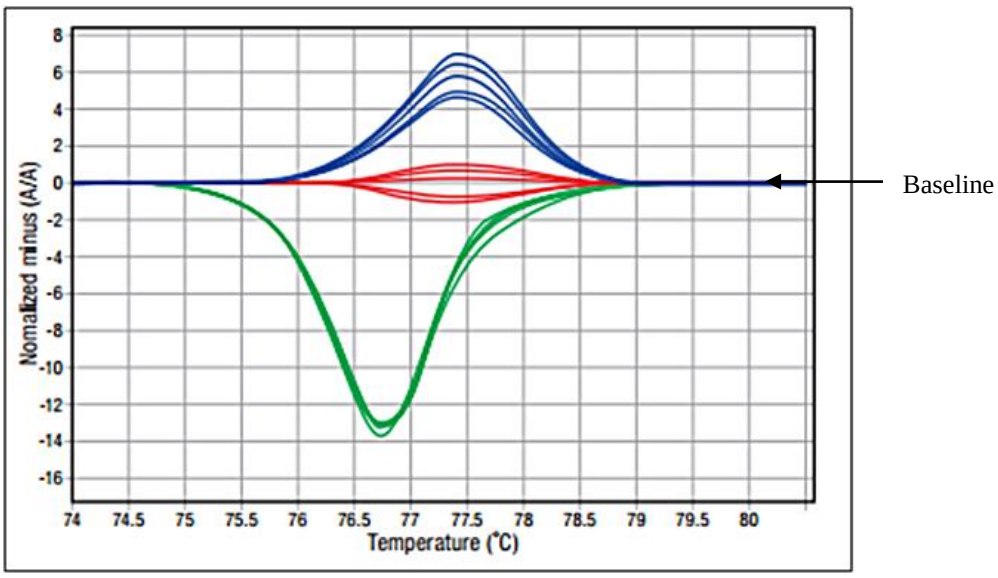


Figure 2.7 Example of a HRM difference graph. This graph is used when a particular genotype is identified and used as a reference/baseline for the other DNA samples. The position of each sample relative to the reference is plotted against the temperature thereby showing differences between various samples. The homozygous wild type is used as a reference (red) and the other samples are compared to it. Blue is the homozygous mutant and green is the heterozygous mutant (Taken from Introduction to HRM Analysis http://www.kapabiosystems.com/public/pdfs/kapa-hrm-fast-pcr-kits/Introduction_to_High_Resolution_Melt_Analysis_Guide.pdf).

HRM is a relatively simple and cost effective means to screen patients for known mutations. It is also a means of identifying novel as well as rare variants and is sufficiently sensitive to identify single base pair changes. An additional advantage is that it is a closed-assay system, and no post PCR processing is therefore necessary.

2.9.1 HRM real time amplification conditions

The real-time PCR and HRM analysis was set up and carried out on a RotorGene 6000 instrument (Corbett Life Science, Australia) with the following cycling conditions: an initial step at 95^0 C for 5 min; 40 cycles with conditions of denaturation at 95°C for 15 s, varying annealing temperatures for each SNP for 15 s and extension at 72°C for 20 s. Thereafter, two additional holding steps were included: 95°C for 1 min to allow for complete denaturation of the double stranded DNA and then at 50°C for 1 min to allow for renaturation of the DNA. HRM analysis was performed with melt temperatures ranging from 65°C to 99°C with the temperature increasing by 0.1°C increments at each step. A wild-type (WT) reference sample was included in every run and samples with altered HRM profiles were selected and Sanger sequenced in order to identify the sequence variant.

2.10 *In silico* prediction of prioritised variants

In order to determine the effect of the selected variants on the structural integrity of the selected proteins, *in silico* modelling was performed on selected candidate variants. For each gene and variant that was analysed, information about the sequences, protein product of the gene, isoforms of a particular protein and the domains for each of the proteins was obtained through UniProt (www.uniprot.org; Consortium 2015). UniProt is an important collection of protein sequences as well as their annotations. Moreover, additional domain information was obtained using the NCBI Conserved Domains Database (www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). For those genes for which sufficient information could be obtained, modelling was performed using SWISS-MODEL (www.swissmodel.expasy.org; Benkert, Biasini, and Schwede 2011). SWISS-MODEL consists of six major steps used to build the protein model and these are summarized below:

1. Template search – performed with Basic Local Alignment Search Tool (BLAST) (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and HHBlits (http://toolkit.genzentrum.lmu.de/hhblits/) against the SWISS-MODEL template

library (SMTL). The target sequence was searched with BLAST against the primary amino acid sequences that were contained in the STML;

2. Template selection – multiple templates were identified and the quality of each of the templates was predicted from the features of the target template alignment. Only the templates with the highest quality were selected;
3. Model building –target templates are used to build specific models and Promod-II was used for the modelling. Coordinates that were observed between the target and the template were subsequently copied to the model. Insertions and deletions were remodelled using a fragment library and side chains were rebuilt;
4. Model quality estimation – the global and per-residue model quality was assessed using the QMEAN scoring function (Benkert, Biasini, and Schwede 2011);
5. Ligand modelling – ligands that were present in the template structure were transferred by homology to the model when an established using the following criteria:
    a. The ligands were annotated as biologically relevant in the template library;
    b. The ligand was in contact with the model;
    c. The ligand did not clash with the protein;
    d. The residues that were in contact with the ligand that were conserved between the template and the target.
6. Oligomeric state conservation – homo-oligomeric structure of the target protein was predicted and this was based on the analysis of pairwise interfaces of the identified template structures.

Subsequent to the modelling of the protein, the protein structures were visualized using two separate visualization programs namely Swiss PDP Viewer and PyMol (www.schrodinger.com). The last part of the *in silico* analysis was to identify motifs in the sequences. This was performed using the Eukaryotic Linear Motif (www.elm.eu.org).

CHAPTER 3: RESULTS

INDEX **PAGE**

3.1 Genealogical analysis to identify additional families 64

3.2 Whole genome SNP array 66

3.2.1 Original six probands 66

3.2.1.1 Identification of segmental sharing between the six original probands 68

3.2.2 All 40 probands 71

3.2.2.1 Calculation of IBD between the 40 probands 71

3.2.2.2 Segmental sharing between the 40 probands 72

3.2.3 Comparison of the 40 probands to the controls 75

3.2.3.1 The identification of IBD and segmental sharing in four unaffected, unrelated Afrikaner individuals 75

3.3 Screening of the known PD genes 78

3.3.1 Sanger sequencing of the 12 exons of *Parkin* 78

3.3.2 MLPA results 80

3.4 Whole exome sequencing to identify a novel PD-causing gene 80

3.4.1 Analysis of WES data using the hypothesis-based approach 84

3.4.1.1 Sanger sequencing validation 87

3.4.1.2 Frequency in ethnically matched control samples 87

3.4.1.2.1 Genotyping of G23E in *TIMM23* 87

3.4.1.2.2 Genotyping of D172G in *IL32* 90

3.4.1.2.3 Genotyping of S301C in *KATNAL2* 90

3.4.2 Analysis of family ZA92 only using the hypothesis-based approach 91

3.4.2.1 Sanger sequencing validation 92

3.4.2.2 Genotyping of P1150S in *EFCAB6* 93

3.4.3 Analysis of WES data using the hypothesis-free approach 95

3.4.3.1 Sanger sequencing validation 98

3.4.3.2 Frequency in ethnically matched control samples 100

3.4.3.2.1 Genotyping of V1405I in *SYNJ1* 100

3.4.3.2.2 Genotyping of C357S in *USP17* 103

CHAPTER 3: Results

3.1 Genealogical analysis to identify additional families

As previously illustrated, the South African PD cohort is composed of individuals spanning numerous ethnic groups (Table 1.2 page 34). Notably, the South African Afrikaner makes up 32.3% (148/458) of the total PD cohort which prompted us to investigate evidence for a founder effect in these patients. Initial genealogical analysis identified the aforementioned six families that were related to one another through a common founder couple (Geldenhuys et al. 2014; B Glanzmann, MSc thesis 2013). While the complete family trees were constructed for the six families, the genealogical analyses for the other PD patients concentrated on finding at least one line of descent from the founder couple to the affected proband. A total of 42 additional probands met the criteria established (positive family history of PD, age at onset younger than 60 years) in order to be scrutinized further.

Intensive genealogical research revealed that for 40 (including the six original probands) of the PD probands, at least one line of descent was found that connected them. Selected lines of descent from the founder couple to each of the 40 probands are highlighted in Figure 3.1. Interestingly, it was determined that on average, a total of four ancestral lines could be identified for each of the probands (note that this may vary between one and fourteen lines per individual). This finding is significant from a genealogical standpoint as this gives strong indication for a founder effect for PD, with the founder couple plausibly identified. For the remaining eight families, the family history that had been provided was insufficient to determine their possible relationship to the founder couple (Geldenhuys et al. 2014). It has been documented that the founder couple had twelve children, but only the five children that have direct ancestral links to the PD probands have been shown. A summary of the available demographic and clinical information for these 40 families is provided in Supplementary Table 1, Appendix II. In these families both AD and AR patterns of inheritance are evident, based on currently available information. However, it is plausible that reduced penetrance and marriage to spouses with a family history of PD could influence the pattern of inheritance observed within specific families.

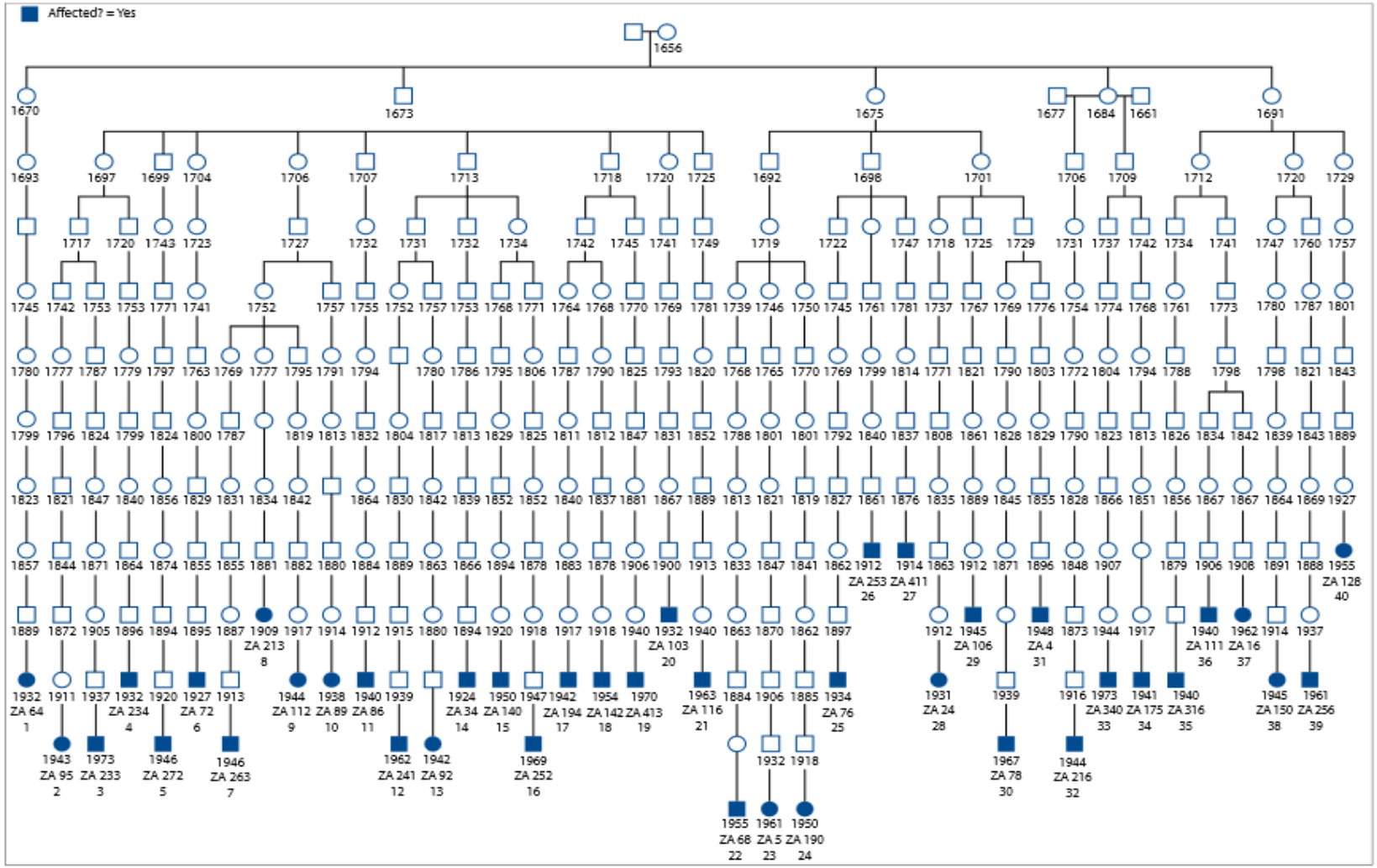


Figure 3.1 Pedigree of the 40 individuals affected with Parkinson's disease shown to be linked to a common founder couple. The pedigree illustrates the ancestral lines to a common founder couple (labels listed as year of birth, family number (ZA) and order on the pedigree 1-40). Males are denoted as squares and females are denoted as circles in the pedigree. The filled shapes are the probands that initially presented with the disease.

3.2 Whole genome SNP array

In order to establish whether or not the probands were in fact closely related to one another, a whole genome SNP array was performed on the probands. A total of 306 670 markers were included in the analysis in order to determine the identity by descent (IBD). Two alleles are IBD if they are copies of the same ancestral gene.

The data was analysed in three ways:

i. In the six original probands only
ii. Across all 40 probands
iii. Comparing the 40 probands to the healthy control individuals

Quality control (QC) was performed on the original six probands who were traced back to the common founder couple. The genotyping rate (also known as the F_MISS rate) and the heterozygosity rate are calculated for each individual in order to identify outlier individuals that are based on both of the statistics and should therefore be excluded from the analysis. These results are shown in Supplementary Table 2, Appendix III. The cut-off thresholds were determined to be as follows: individuals with low genotyping were removed from the dataset (--mind 0.05); minor allele frequencies of greater than 0.01 were removed from the dataset (--maf 0.01) and the removal of all SNPs with missing call rates greater than 0.05 (--geno 0.05). Sample ZA111 (78.95) was removed from the analysis as it had a large number of SNPs that were missing and an excess of heterozygosity. The inclusion of this individual could lead to skewed results particularly pertaining to the IBD shared regions.

3.2.1 Original six probands

Following the identification of appropriate cut-off thresholds for QC, IBD testing as well as the search for segmental sharing was conducted using PLINK (Purcell et al. 2007). In order to identify related individuals, IBD was calculated. This is based on the average proportion of alleles that are shared in common at specific genotyped SNPs. A preliminary analysis was conducted on the sibling pair of family ZA92 (68.27 and 68.46) as a means of secondary QC. Siblings are expected to show an IBD score of approximately 0.5 because they are first-degree relatives. The IBD scores can be seen in Table 3.1. PI(hat) scores are used as a measure to determine the probability of a pair of relatives being IBD and is therefore a measure of the overall IBD. PI(hat) scores are calculated through the use of the following formula: PI(hat) = Z2 + 0.5*(Z1) where Z1 is the probability that individuals will share one

allele at a specific marker position and Z2 is the probability that individuals will share two alleles at a specific marker position. Z0 is the probability that individuals will share no alleles at a specific marker position. While PI(hat) scores are an indication of the overall degree of relatedness, Z0, Z1 and Z2 are an indication of the type of relatedness between individuals. The affected siblings share a PI(hat) score of 0.4883, which can be rounded to 0.5, therefore producing an expected IBD. Note that IBD is calculated between two individuals at a time.

Table 3.1 Identity by descent (IBD) scores shared between the siblings of family ZA92.

FID1	IID1	FID2	IID2	Z0	Z1	Z2	PI(HAT)
ZA92	68.27	ZA92	68.46	0.2449	0.5335	0.2216	0.4883

FID – Family ID; IID – Individual ID; Z0 – probability that individuals at a specific marker will share no alleles; Z1 – probability that individuals at a specific marker will share 1 allele; Z2 – probability that individuals at a specific marker will share 2 alleles.

Note that due to quality control measures, ZA111 was excluded from the analysis due to the sample's high degree of missingness. Although there are varying degrees of relatedness, it was determined that all of the individuals are related to one another. It is estimated that each generation is approximately 25 years (Takahata 1993; Cavalli-Sforza and Feldman 2003; Medeiros-Domingo et al. 2007; Greeff and Erasmus 2015) and for this reason, probands ZA89 (68.16) and ZA92 (68.27 and 68.46) are the most distantly related (approximately 150 years; 6 generations) and ZA92 (68.27 and 68.46), ZA106 (78.67) and ZA134 (81.65) are the most closely related (approximately 100 years; 4 generations). The results from the IBD shared between the original six probands are shown in Table 3.2. From these results it is clear that each of the probands that were related to one another through genealogical tracking, are in fact genetically related to one another – although there are varying degrees of relatedness between various individuals.

Table 3.2 IBD shared between the original six probands and affected sibling traced back to a common founder couple.

FID1	IID1	FID2	IID2	Z0	Z1	Z2	PI(HAT)	Approximate Degree of Relatedness in Generations
ZA106	78.67	ZA134	81.65	0.8936	0.1064	0.0000	0.0532	4
ZA106	78.67	ZA78	67.82	0.9027	0.0973	0.0000	0.0487	5
ZA106	78.67	ZA89	68.16	0.9044	0.0956	0.0000	0.0478	5
ZA106	78.67	ZA92	68.27	0.9362	0.0638	0.0000	0.0319	5
ZA134	81.65	ZA78	67.82	0.9031	0.0969	0.0000	0.0485	5
ZA134	81.65	ZA89	68.16	0.9365	0.0635	0.0000	0.0317	5
ZA134	81.65	ZA92	68.27	0.8873	0.1127	0.0000	0.0564	4
ZA78	67.82	ZA89	68.16	0.9225	0.0775	0.0000	0.0388	5
ZA78	67.82	ZA92	68.27	0.9192	0.0808	0.0000	0.0404	5
ZA89	68.16	ZA92	68.27	0.9571	0.0429	0.0000	0.0214	6

FID – Family ID; IID – Individual ID; Z0 – probability that individuals at a specific marker will share no alleles; Z1 – probability that individuals at a specific marker will share 1 allele; Z2 – probability that individuals at a specific marker will share 2 alleles.

3.2.1.1 Identification of segmental sharing between the six original probands

As already mentioned, the degree of recent shared ancestry for a pair(s) of individuals can be measured through the calculation of IBD (or PI(hat) scores). Given that we identified that all of the aforementioned individuals are IBD, the regions that are shared by all of these individuals were then investigated. This was done using PLINK's segmental sharing algorithm which uses a Hidden Markov Model (HMM) to detect chromosomal segments that are shared by descent (Purcell et al. 2007). Segmental sharing was first calculated in the sibling pair to identify shared IBD stretches between the sibling pair before applying this to the six probands. There are 55-shared segments between the sibling pair, with the majority of them (8/55) on chromosome 3. The segmental sharing results between the sibling pair are shown in Supplementary Table 3, Appendix III. It is estimated that the sibling pair shares approximately 67% of the genome in IBD stretches. This is calculated by taking the sum of the physical length of the segments that are shared / sum of the total lengths of chromosomes. It is expected that the siblings will share approximately 50% of IBD stretches, but this figure can vary between 0% and 100% depending on the specific loci. This is due to the fact that there are hotspots for recombination at specific chromosomal regions.

The percentage of genome sharing between the original six probands was also subsequently calculated as a measure of "how related" individuals are to one another. This is done by calculating the degree of segmental sharing and subsequently what percentage of the genome is shared to give a more accurate indication of relatedness as opposed to PI(hat) scores. It is estimated that the original six probands share 1.54% of the genome with the overall lengths of the shared segments differing between 4192.60 KB and 14051.50 KB in length (Table 3.3). Individuals ZA78 and ZA89 (67.82 and 68.16) have the highest percentage of chromosomal sharing, as they share a total of 11.31% of chromosome 1. ZA134 shares the most number of segments with other probands (4/5) and can therefore be regarded as the most related proband, with ZA92 (68.27) the least related as it shares only one segment on chromosome 3 with ZA134.

Table 3.3 Highest percentage of the chromosomes shared across the six original probands.

FID1	IID1	FID2	IID2	CHR	BP1	BP2	SNP1	SNP2	NSNP	KB	Percentage of CHR shared
ZA134	8165	ZA92	6827	1	16001588	20423099	rs12746773	rs6699362	153	4421.51	3.56
ZA78	6782	ZA89	6816	1	168149142	182200659	rs4657741	rs10494545	330	14051.5	11.31
ZA106	7867	ZA134	8165	3	27253971	31488222	rs1445111	rs6766414	151	4234.25	2.12
ZA78	6782	ZA89	6816	6	148981394	153173999	rs7451498	rs9384046	163	4192.6	2.33
ZA106	7867	ZA89	6816	14	92280037	98263147	rs10134181	rs898927	244	5983.11	5.62
ZA134	8165	ZA78	6782	16	49932806	54343840	rs4785195	rs1004299	129	4411.03	4.97
ZA134	8165	ZA78	6782	16	59070616	64124789	rs990813	rs12596363	117	5054.17	5.69

CHR – chromosome; IID – Individual ID; FID – Family ID; BP1 – start of the physical position of the segment (base pair); BP2 - end of the physical position of the segment (base pair 2); SNP1 – start of the SNP segment; SNP2 – end of the SNP segment; NSNP – number of SNPs in the segment; KB – physical length of the segment

3.2.2 All 40 probands

3.2.2.1 Calculation of IBD between the 40 probands

Following the successful identification of IBD as well as shared segments between the siblings of ZA92 (68.27 and 68.46) as well as between the original six probands, the same quality control measures were implemented for the 40 probands that had been traced back to the common founder couple. The same QC parameters were used as previously defined in Section 3.2, page 66. Due to these parameters, individuals ZA340 (81.90) and ZA111 (78.95) were removed from the analysis due to a high degree of missingness. There are 702 different comparisons (this is each proband compared to the other respectively) for IBD and the PI(hat) scores are shown in Supplementary Table 4, Appendix III. It was determined that all of the 40 Afrikaner probands are related to one another with varying degrees of relatedness (Table 3.4, Figure 3.2). Estimates of IBD coefficients, namely Z0 and Z1 scores are plotted as a means to infer relatedness. Note that Z0 + Z1 = 1. This is the probability that individuals will share one allele at a specific marker (Z1) or none (Z0). Individuals that share no alleles and therefore have Z1 of less than 0.0100 are considered to be the least related to one another (and therefore the greatest number of generations between them) and have been circled in red (Figure 3.2). These individuals are ZA150 (82.39), ZA190 (83.16), ZA411 (10.110) and ZA34 (55.81). The approximate degree of relatedness calculated by the average number of PI(hat) scores – therefore probands with higher PI(hat) scores are related to more probands.

Table 3.4 Degrees of relatedness between the 40 Afrikaner probands.

Family ID	Number of unrelated individuals	Number of related individuals	Average PI(hat) scores	Approximate Degree of Relatedness in Generations
ZA140	14	23	0.01580000	6
ZA142	27	10	0.00147500	6
ZA103	19	18	0.00396486	8
ZA116	27	10	0.00157568	8
ZA128	15	23	0.00620811	8
ZA16	16	11	0.00385556	8
ZA175	29	8	0.00127027	8
ZA194	21	16	0.00328919	8
ZA213	15	22	0.00586216	8
ZA216	16	21	0.00644324	8
ZA233	12	25	0.00638649	8
ZA24	13	24	0.00595676	8

ZA241	12	25	0.00678378	8
ZA263	31	6	0.00130270	8
ZA272	28	9	0.00152973	8
ZA4	17	20	0.00443514	8
ZA413	18	19	0.00523784	8
ZA5	15	22	0.00561892	8
ZA64	35	2	0.00046175	8
ZA68	11	26	0.00762703	8
ZA72	15	22	0.00710270	8
ZA76	10	27	0.00644243	8
ZA86	15	22	0.00602703	8
ZA89	20	17	0.00453784	8
ZA95	17	20	0.00594167	8
ZA106	27	10	0.00251622	9
ZA112	25	12	0.00213243	9
ZA134	23	14	0.00295946	9
ZA252	23	14	0.00278108	9
ZA253	24	13	0.00285405	9
ZA256	25	12	0.00239189	9
ZA316	25	12	0.00225135	9
ZA78	28	9	0.00200270	9
ZA92	22	15	0.00297297	9
ZA190	31	6	0.00064865	11
ZA411	34	3	0.00061622	11
ZA150	36	1	0.00017568	12
ZA34	36	1	0.00017568	12

3.2.2.2 Segmental sharing between the 40 probands

Segmental sharing was calculated for all 40 of the related Afrikaner using PLINK, as described in Section 3.2.1, page 66. There are 1413 shared segments across the 40 related probands. For this reason, the shared segments across chromosomes rather than across individuals were calculated as there are too many shared segments to analyse properly. These results are shown in Table 3.5. The average percentage of chromosomal segment sharing is 5.17%, with the greatest degree of segmental sharing in chromosome 19 and the smallest degree of sharing in chromosome 2. The average length of shared segments in each chromosome is 5 748 298.77 base pairs (5 748.30KB). Moreover, there is an average of 238.32 SNPs that are shared per segment. The 40 probands do not share large regions of chromosomes, thereby supporting the genealogy that the individuals are distantly related to one another.

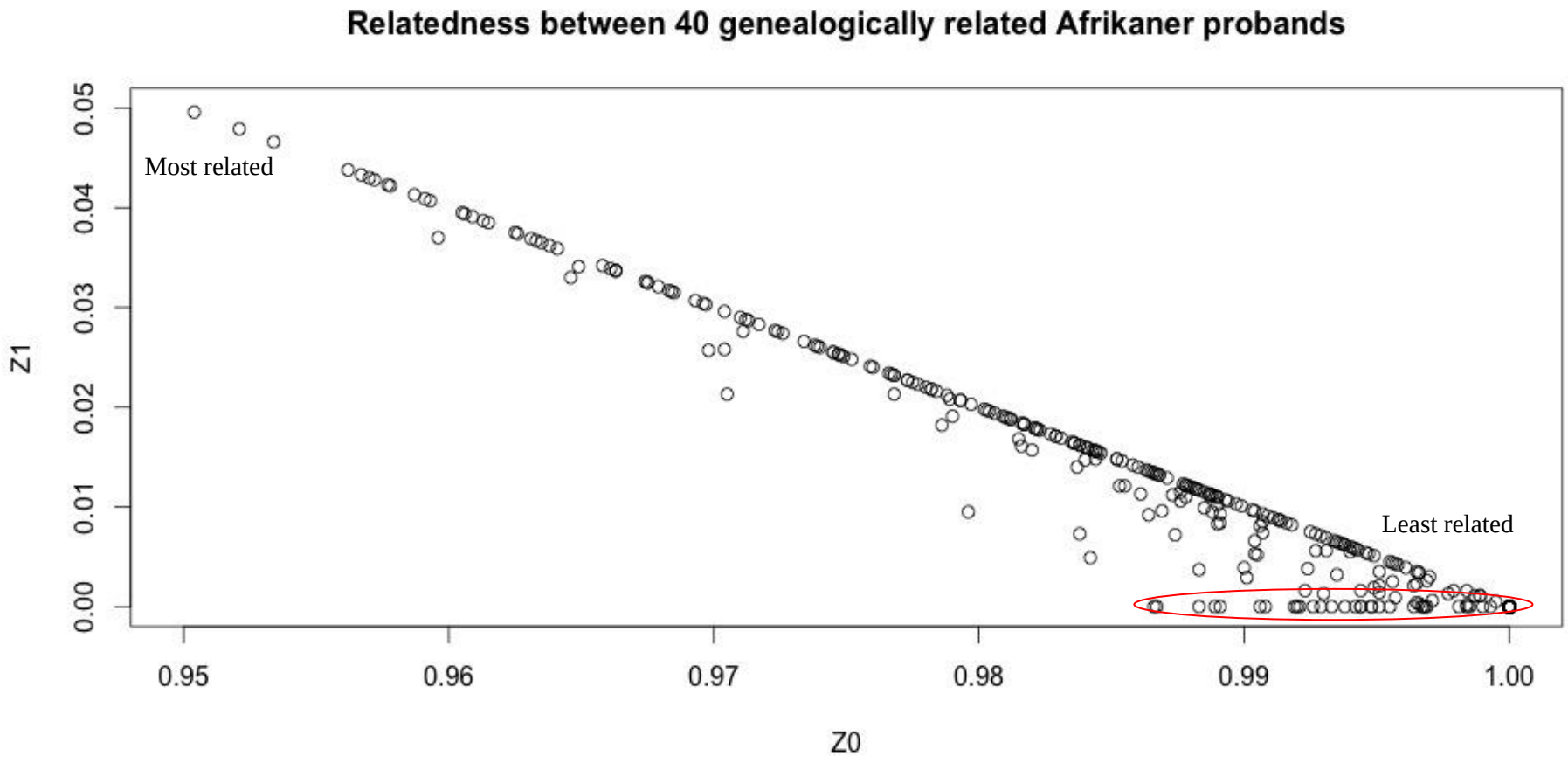


Figure 3.2 Relatedness inferences from IBD estimates. Estimates of the IBD coefficients, namely Z0 and Z1 are used to infer relatedness. Each point represents a pair of samples. Note that the sum of the values will give total variance. The pairs of samples that are circled in red are those pairs of samples that are the most unrelated to one another. All 40 probands were found to be distantly related to one another through IBD estimates.

Table 3.5 The number of shared segments across the 40 probands (chromosomally).

Chromosome	Length of Chromosome (bp)	Number of Shared Segments	Average Length of Shared Segment (bp)	Average Number of SNPs per Segment	Percentage of Chromosome Shared
1	247 249 719	93	8 541 490	282.82	3.35
2	242 951 149	125	6 065 207	237.98	2.50
3	199 501 827	86	6 822 114	264.71	3.41
4	191 273 063	107	6 398 454	253.24	3.35
5	180 857 866	87	5 557 209	204.47	3.07
6	170 899 992	139	5 178 270	234.02	3.03
7	158 821 424	80	6 030 635	240.60	3.80
8	146 274 826	94	5 435 839	233.85	3.72
9	140 273 252	59	5 512 568	227.24	3.93
10	135 374 737	87	5 143 802	236.03	3.80
11	134 452 384	68	7 175 158	235.29	5.33
12	132 349 534	61	6 409 982	263.28	4.84
13	114 142 980	47	4 637 914	215.06	4.06
14	106 368 585	50	5 898 843	250.72	5.55
15	100 338 915	33	5 725 393	241.36	5.71
16	88 827 254	43	5 281 153	235.07	5 95
17	78 774 742	40	7 575 798	288.73	9.61
18	76 117 153	41	4 739 005	225.44	6.22
19	63 811 651	20	7 123 259	286.95	11.16
20	62 435 964	18	2 540 611	149.39	4.07
21	46 944 323	19	4 591 567	247.42	9.78
22	49 691 432	10	4 078 302	189.30	8.21

3.2.3 Comparison of the 40 probands to the controls

3.2.3.1 The identification of IBD and segmental sharing in four unaffected, unrelated Afrikaner control individuals

Given the unique ethnic background of the South African Afrikaner, a total of four unaffected, unrelated Afrikaner control patients were selected and also analysed using the whole genome SNP array. This was to determine whether or not randomly selected, unrelated, healthy control individuals shared the same degree of IBD and segmental sharing as the genealogically related probands. Estimates of IBD coefficients, namely Z0 and Z1 scores are used plotted as a means to infer relatedness. Due to the fact that the control individuals were randomly selected, it is expected that none of them would share IBD with the probands. It was determined, however that control individual 11.937 shares a relationship with the original probands namely ZA134 (81.65), ZA78 (67.82), ZA89 (68.16) and ZA92 (68.27) (Table 3.6). Moreover, this individual shares a relationship with 21 other Afrikaner probands that are related to the founder couple but not to any of the other control individuals (Figure 3.3). This is an indication that this control may be potentially related to the probands.

Table 3.6 IBD shared between the original six probands and four randomly selected, unaffected Afrikaner controls.

FID1	IID1	FID2	IID2	Z0	Z1	Z2	PI(HAT)
ZA106	78.67	Control_1	12.057	1.0000	0.0000	0.0000	0.0000
ZA106	78.67	Control_2	11.987	1.0000	0.0000	0.0000	0.0000
ZA106	78.67	Control_3	11.976	1.0000	0.0000	0.0000	0.0000
ZA106	78.67	Control_4	11.937	1.0000	0.0000	0.0000	0.0000
ZA134	81.65	Control_1	12.057	1.0000	0.0000	0.0000	0.0000
ZA134	81.65	Control_2	11.987	1.0000	0.0000	0.0000	0.0000
ZA134	81.65	Control_3	11.976	1.0000	0.0000	0.0000	0.0000
ZA134	81.65	Control_4	11.937	0.9927	0.0073	0.0000	0.0037
ZA78	67.82	Control_1	12.057	1.0000	0.0000	0.0000	0.0066
ZA78	67.82	Control_2	11.987	1.0000	0.0000	0.0000	0.0000
ZA78	67.82	Control_3	11.976	1.0000	0.0000	0.0000	0.0000
ZA78	67.82	Control_4	11.937	0.9688	0.0312	0.0000	0.0156
ZA89	68.16	Control_1	12.057	1.0000	0.0000	0.0000	0.0000
ZA89	68.16	Control_2	11.987	1.0000	0.0000	0.0000	0.0000
ZA89	68.16	Control_3	11.976	1.0000	0.0000	0.0000	0.0000
ZA89	68.16	Control_4	11.937	0.9920	0.0080	0.0000	0.0040
ZA92	68.27	Control_1	12.057	1.0000	0.0000	0.0000	0.0000
ZA92	68.27	Control_2	11.987	1.0000	0.0000	0.0000	0.0000
ZA92	68.27	Control_3	11.976	1.0000	0.0000	0.0000	0.0000

ZA92	68.27	Control_4	11.937	0.9691	0.0309	0.0000	0.0155
Control_1	12.057	Control_2	11.987	1.0000	0.0000	0.0000	0.0000
Control_1	12.057	Control_3	11.976	1.0000	0.0000	0.0000	0.0000
Control_1	12.057	Control_4	11.937	1.0000	0.0000	0.0000	0.0000
Control_2	11.987	Control_3	11.976	1.0000	0.0000	0.0000	0.0000
Control_2	11.987	Control_4	11.937	1.0000	0.0000	0.0000	0.0000
Control_3	11.976	Control_4	11.937	1.0000	0.0000	0.0000	0.0000

FID – Family ID; IID – Individual ID; Z0 – probability that individuals at a specific marker will share no alleles; Z1 – probability that individuals at a specific marker will share 1 allele; Z2 – probability that individuals at a specific marker will share 2 alleles.

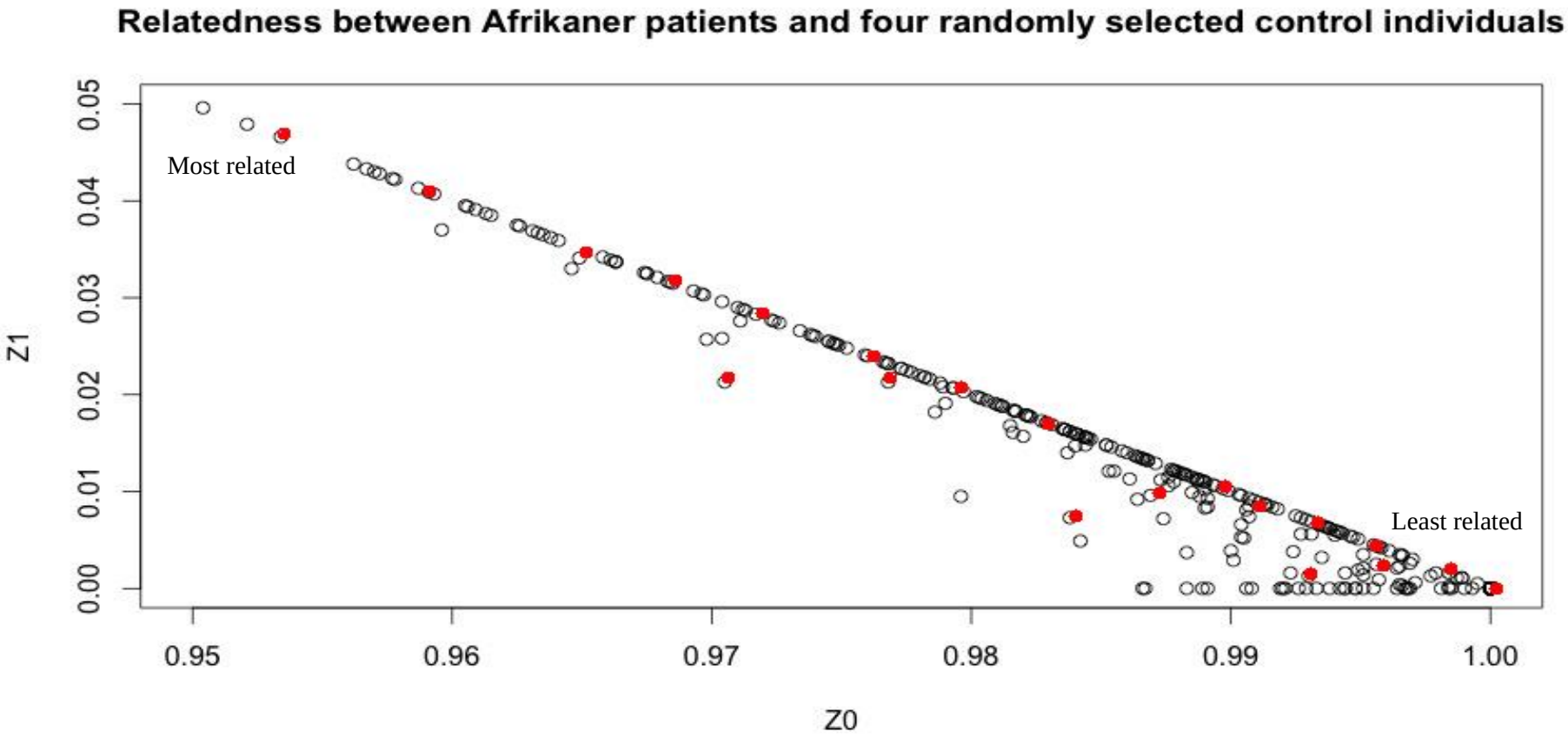


Figure 3.3 Relatedness inferences from IBD estimates including the control individuals. Estimates of the IBD coefficients, namely Z0 and Z1 are used to infer relatedness. Each point represents a pair of samples. Only one control (shown here in red – Control 11.937) was found to share any relation to the 40 Afrikaner probands through IBD estimates.

3.3 Screening of the known PD genes

3.3.1 Sanger sequencing of the 12 exons of *Parkin*

In order to identify the genetic cause of the disorder in these Afrikaner probands, we screened and excluded all of the common known pathogenic PD-causing mutations in these individuals. As the *Parkin* gene is a common cause of PD with more than 150 mutations identified, all 12 *Parkin* exons were screened using Sanger sequencing to determine whether or not this was the underlying cause of the disease in the 40 founder families. Numerous polymorphisms were identified in some of the patients (Table 3.7), but 18 of the 40 patients (45%) carried no variants in *Parkin*. However, it was determined that one patient, 95.94 (ZA340) carries two homozygous variants namely R275W in exon 7 (this is a known pathogenic mutation) and M432V in exon 12. Therefore, the remaining 39 probands were shown not to harbour pathogenic mutations in *Parkin.*

Table 3.7 Sequence variants found in *Parkin* in 22 Afrikaner patients.

Family Number	Patient	Exon	Variant	rs number	Frequency
ZA233	88.74	8	IVS8+48C>T	None	None
ZA272	94.58	10 11	V380L R402C	rs1801582 rs55830907	G, 0.682; C, 0.318; n=4550 C, 0.997; T, 0.003; n=4552
ZA112	78.97	2	IVS2-35G>A	None	None
ZA89	68.16	8	IVS8+48C>T	None	None
ZA241	90.87	10	V380L	rs1801582	G, 0.682; C, 0.318; n=4550
ZA140	81.74	11	D394N	rs1801334	G, 0.945; A, 0.055; n=4550
ZA252	92.00	8	IVS8+48C>T	None	None
ZA413	10.141	2 10	IVS2-35G>A V380L	None rs1801582	None G, 0.682; C, 0.318; n=4550
ZA116	79.29	6 8	A225T IVS8+48C>T	rs202212928 None	C, 0.998; A, 0.002; n=1323 None

		10	V380L	rs1801582	G, 0.682; C, 0.318; n=4550
ZA5	38.23	11	D394N	rs1801334	G, 0.945; A, 0.055; n=4550
ZA190	83.16	10	V380L	rs1801582	G, 0.682; C, 0.318; n=4550
ZA24	54.73	10	V380L	rs1801582	G, 0.682; C, 0.318; n=4550
ZA106	78.67	8	IVS8+48C>T	None	None
ZA5	38.36	8	IVS8+48C>T	None	None
ZA340	95.94	7 12	R275W* M432V*	rs34424996 None	None None
ZA175	83.01	10	V380L	rs1801582	G, 0.682; C, 0.318; n=4550
ZA316	95.64	8	IVS8+48C>T	None	None
ZA111	78.95	8 10	IVS8+48C>T V380L	None rs1801582	None G, 0.682; C, 0.318; n=4550
ZA16	56.45	8 10	IVS8+48C>T V380L	None rs1801582	None G, 0.682; C, 0.318; n=4550
ZA150	10.113	1 8	IVS1-61A>C IVS8+48C>T	None None	None None
ZA256	92.60	6 8	A225T V380L	rs202212928 rs1801582	C, 0.998; A, 0.002; n=1323 G, 0.682; C, 0.318; n=4550
ZA128	81.58	10	V380L	rs1801582	G, 0.682; C, 0.318; n=4550

All frequencies are determined in the Exome Sequencing Project cohort population available in dbSNP; n - number of chromosomes; IVS - intervening sequence; * -variant identified in homozygous state

3.3.2 MLPA results

The 40 PD probands were subjected to MLPA in order to eliminate CNVs in the known PD genes as the possible cause for PD in these patients. The two commercially available Parkinson's disease MLPA kits (SALSA P051-B1 and P052-B1) were used and it was determined that none of the 40 probands included in the genealogical analysis carried exonic duplications, triplications or deletions and this could thus also be excluded as a possible cause for PD in these patients.

3.4 Whole exome sequencing to identify a novel PD-causing gene

Given the substantial evidence of a possible founder effect for PD in the Afrikaner individuals as well as conclusive genealogical evidence that the 40 probands are related to one another, three of the probands (from the original six probands) from families ZA92, ZA106 and ZA111 were selected for WES based on their family history as well as the young age at onset of the disease. The individual pedigrees for each of the selected families are shown below in Figure 3.4, 3.5 and 3.6. Moreover, the affected sibling of ZA92 (herein referred to as ZA92_sib) was also selected for WES. The raw sequencing reads were realigned to the human reference genome hg19 as a means to identify SNPs, insertions and/or deletions. The variants identified through WES are summarized in Table 3.8.

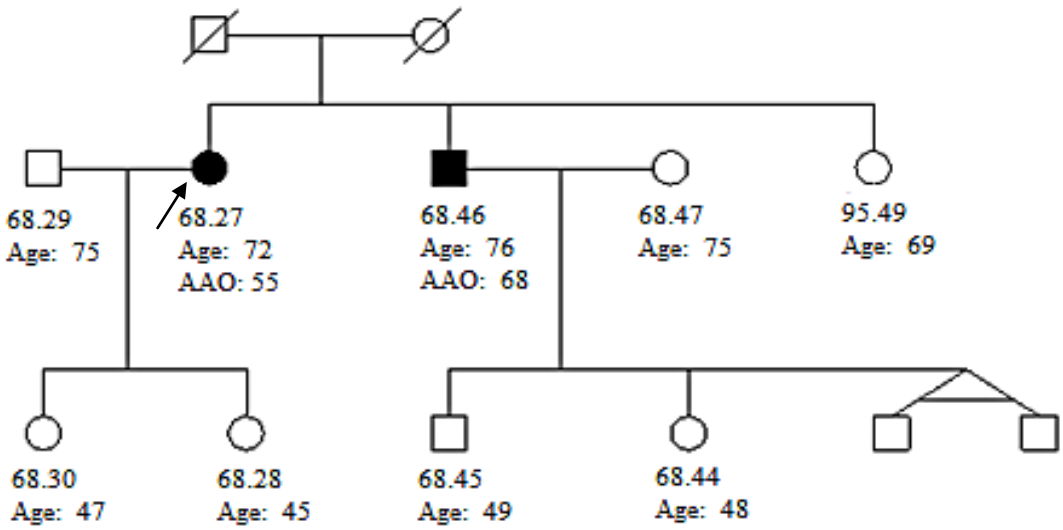


Figure 3.4 Pedigree of family ZA92. The pedigree shows the information available regarding this family. The arrow indicates the proband of the pedigree (68.27). Males are denoted as squares and females are denoted as circles in the pedigree. The filled shapes are the affected individuals.

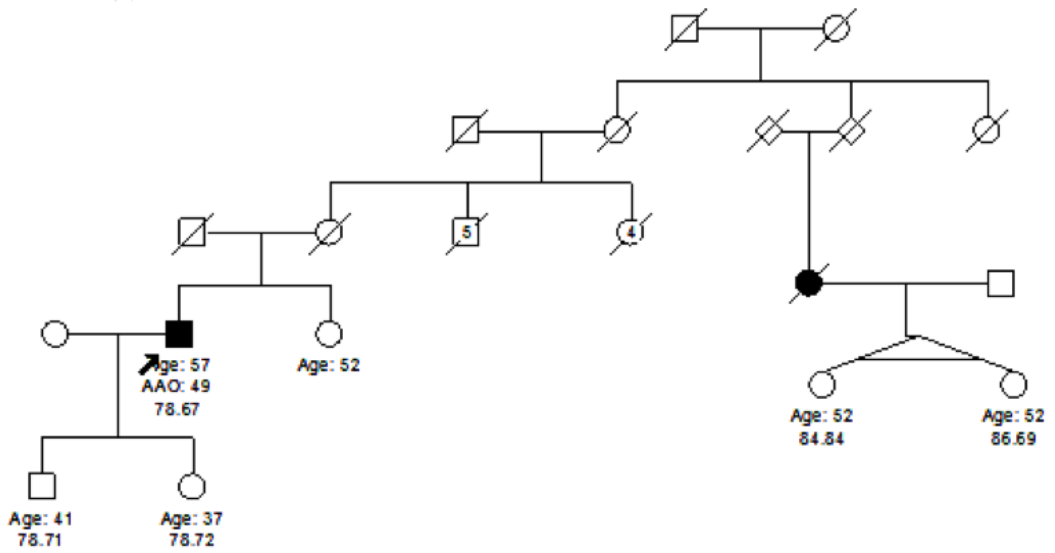


Figure 3.5 Pedigree of family ZA106. The pedigree shows the information available regarding this family. The arrow indicates the proband of the pedigree (78.67). Males are denoted as squares and females are denoted as circles in the pedigree. The filled shapes are the affected individuals.

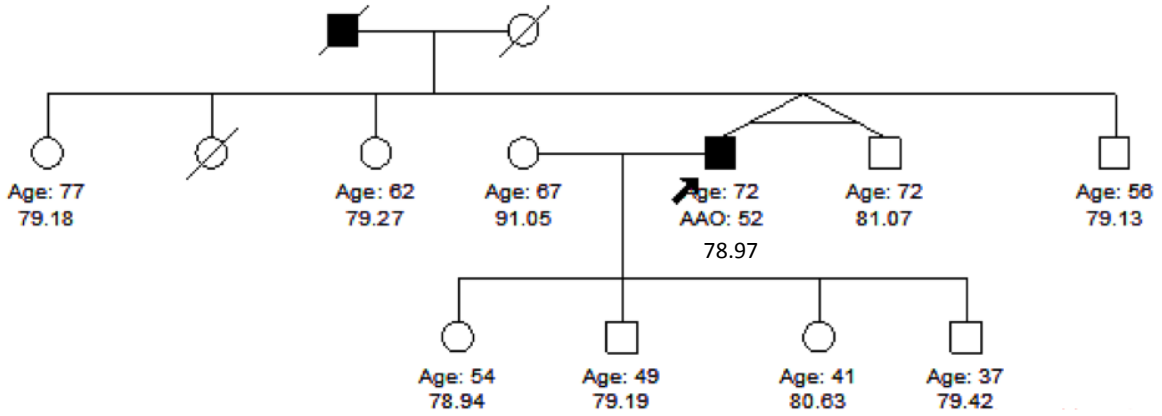


Figure 3.6 Pedigree of family ZA111. The pedigree shows the information available regarding this family. The arrow indicates the proband of the pedigree (78.97). Males are denoted as squares and females are denoted as circles in the pedigree. The filled shapes are affected individuals.

Table 3.8 Summary of WES results across three probands and one affected sibling.

	ZA92 proband	ZA92_sib	ZA106 proband	ZA111 proband
Total number of variants	74 938	76 117	77 700	79 070
Total number of SNPs	71 144	72 220	73 784	75 536
Total number of Indels	3 839	3 897	3 916	3 534

SNP – single nucleotide polymorphism

Although all of the known PD genes had previously been excluded as a cause for PD in these patients, we specifically analysed all of the known PD genes in the WES data, which revealed a number of known SNPs and one novel variant in the 5'UTR of *EIF4G1* (Table 3.9). No variants were found in *DJ-1*, *SNCA*, *CHCHD2*, *DNAJC13* or *UCHL1*. Notably, the findings confirm our previous findings and excluded all the known PD genes from causing the disorder in these individuals. This means that it is likely that they harbour a mutation(s) in a novel PD-causing gene.

Table 3.9 Variants detected in the known PD genes in the three PD probands ZA92, ZA106 and ZA111 as well as the affected sibling ZA92 (sibling).

PD Gene	Variant	In dbSNP	Frequency (n=no. chromosomes)	Present in			
				ZA92 proband	ZA92 sibling	ZA106 proband	ZA111 proband
Parkin	V189L	rs1801582	(ESP) C,0.174; G,0.826; n=4550	Yes	Yes	Yes	Yes
	3'UTR +118G>A	rs6812138	(ESP) A, 0.092; G,0.908; n=4550	No	Yes	No	Yes
	3'UTR +413 A>G	rs4709583	(ESP) A,0.085; G, 0.914; n=4550	No	Yes	No	No
PINK1	Non-genic*	rs3131713	(ESP) A, 0.883; G, 0.117; n=4550	Yes	Yes	Yes	Yes
	N521T	rs1042434	(ESP) C,0.677; G,0.333; n=4552	Yes	No	Yes	Yes
	3'UTR+37A>T	rs686658	(CSAgilent) A, 0.125; T,0.875; n=1324	No	No	Yes	Yes
	3'UTR+181 C>G	rs513414	(CEPH) C, 0.150; G, 0.850; n=184	Yes	No	Yes	No
	3'UTR+40A>G	rs6893	(CEPH) C, 0.890;	No	No	No	Yes

			T, 0.110; n=184				
LRRK2	R50H	rs2256408	(ESP) A, 0.923; G, 0.077; n=4550	Yes	Yes	Yes	Yes
	N551K	rs7308720	(ESP) C, 0.898; G0.102; n=4510	No	No	Yes	Yes
	R1398H	rs7133914	(ESP) A,0.100; G, 0.900; n=4540	No	No	Yes	No
	S1647T	rs3459182	No frequencies available	No	No	Yes	No
	M2397T	rs3761863	(ESP) A,0.384; G,0.616; n=4554	Yes	Yes	Yes	Yes
	3'UTR+140 C>T	rs6673790	No frequencies available	Yes	Yes	Yes	Yes
SNCA	-	-	-	-	-	-	-
DJ-1	-	-	-	-	-	-	-
ATP13A2	P1172P	rs3170740	(CEPH) A, 0.760; G, 0.240; n=184	Yes	Yes	Yes	No
	Non-genic*	rs7531163	(CSAgilent) A, 0.283; G,0.717 n=1324	No	Yes	No	Yes
VPS35	3'UTR+281 C>A	rs808078	No frequencies available	No	Yes	No	Yes
EIF4G1	T161A	rs1331914	(CEPH) C, 0.075; T, 0.925 n=184	No	No	No	Yes
	M432V	rs2178403	(ESP) C,0.841; T,0.159; n=4552	No	No	No	Yes
	5'UTR	Novel	No frequencies available	No	No	No	Yes
CHCHD2	-	-	-	-	-	-	-
PD-associated genes							
DNAJC13	-	-	-	-	-	-	-
GBA3	Y149X	rs358231	(CSAgilent) A, 0.816; T,0.184; n=569	Yes	Yes	No	No
UCHL1	-	-	-	-	-	-	-
FBXO7	M36I	rs11107	(ESP) C, 0.632; T,0.368; n=4540	No	Yes	Yes	No
PLA2G6	-	-	-	-	-	-	-
MAPT	-	-	-	-	-	-	-

*Non-genic refers to areas on contigs that do not contain any genes; ESP, Exome Sequencing Project; CSAgilent, this population includes 662 participants of European descent from the ClinSeq project, all of whom

have undergone whole exome sequencing using Agilent's 38Mb or 50Mb capture kit; CEPH; Genomic DNA samples were obtained for a panel of 92 unrelated individuals chosen from Centre d'Etude du Polymorphisme Human (CEPH) pedigrees. The genomic DNA comprised of UTAH (93%), French (4%), and Venezuelan (3%) samples were purchased from Coriell Cell Repository and pooled in equimolar amounts for use.

The exclusion of the known PD-causing genes indicates that the genetic cause for this disease is likely to be as a result of novel variations in a novel gene. For this reason, along with the genealogical information, all low frequency and novel variants that occur in all four affected patients were identified.

3.4.1 Analysis of WES data using the hypothesis based approach

As described in Section 2.6.1, page 49, the hypothesis-based approach for WES data analysis assumes a number of factors to be known about the disease including a detailed account of all possible phenotypic characteristics and the mode of inheritance. In the case of PD, phenotypic characteristics such as bradykinesia, resting tremor, postural instability and rigidity were all terms that were used to prioritise variants across affected individuals. The candidate list of variants was subjected to the Biological Database Network (http://biodbnet.abcc.ncifcrf.gov/db/db2db.php) and GO ontologies for cellular, biological and molecular processes were obtained. Genes that have been extensively studied and have either been associated with the disease phenotype and subsequently with various clinical presentations are moved to the top of the priority list, while lesser known genes that have been studied less extensively, move down the priority list. Moreover, genes that have variants with low frequencies and have been associated with an aspect of the disease phenotype are prioritised and placed higher up on the list as well.

The identification of common variants between the custom method and Ingenuity Variant Analysis were prioritised for further analysis. The results showed a total of 74 non-synonymous variants with low frequency in the databases across all four PD affected individuals. The results are summarized in Appendix IV. Variants with the lowest frequency or for which no frequency information is available have been prioritised and ordered at the top of the table. Further analysis of the variants removed all of the genes for which there were multiple SNPs per single gene. Moreover, a number of genes that have been flagged in numerous WES projects that recurrently produce SNPs, but which have been identified as false positives (Fajardo et al. 2012; http://massgenomics.org/2013/06/ngs-false-positives.html), were removed. The elimination of these genes resulted in a total of 22

genes (24 variants) remaining for further analysis. The results are summarized in Table 3.10. Variants with the lowest frequency or for which no frequency information is available have been prioritised and ordered at the top of the table.

Of these shared variants, it was necessary to determine which were likely to be disease-causing; the list of 24 variants was prioritised according to the following criteria:

1) were non-synonymous or nonsense variants, AND

2) the average allele frequency of the variant in the databases had to be less than or equal to 0.03, AND

3) the read depth had to be greater than or equal to 50, AND

4) if the variants were predicted as being potentially deleterious according to SIFT, PolyPhen2 or MutationTaster, AND

5) were predicted to be in a conserved domain using PhyloP, which considers evolutionary conservation across multiple species, thereby emphasising the fact that evolutionary conservation is essential in determining whether or not a non-synonymous variant is likely to be pathogenic (Li et al. 2013).

Read depth refers to the number of times each base was sequenced in total. This is predetermined by the platform used and for the Illumina Human All Exon Kit, it is predicted that each base will be covered a minimum of 50x - therefore anything below this coverage may be an artefact rather than a true variant (Charier et al. 2012). These criteria provided an ordered list of variants and 12 of the 24 were chosen for Sanger sequencing verification in all four of the patients who had been subjected to WES (Table 3.10). Verification with Sanger sequencing is a necessary and important step, as NGS technology is known to produce a significant number of artefacts mainly due to the short read lengths.

Table 3.10 Overlapping prioritised SNPs across four individuals affected with PD.

Gene	SNP	Frequency	rs number	SIFT	PolyPhen2	Mutation Taster	Selected variant for further analysis	Sanger sequencing result
TPSD1	H143R	0	rs72775466	T	B	0.102904	×	N/A
PRB2	R357Q	0	-	NA	NA	0.001769	×	N/A
PLIN4	A883T	0	rs80238130	-	-	-	×	N/A
NOTCH2NL	T158I	0	rs75586173	T	D	0.997294	✓	Artefact
KIR3DL1	S239N	0	-	T	NA	3.9E-5	×	N/A
IL28A	F137L	0	-	NA	B	1.9E-5	×	N/A
HBD	S87A	0	-	T	NA	0.162278	×	N/A
KATNAL2	S301C	0	rs76539063	D	P	0.9725871	✓	Real but in control individuals
CLIP1	L271F	0	rs79909185	D	P	0.974668	×	N/A
CASP1	G85E	0	rs2509649	T	B	0.097257	×	N/A
ANAPC1	Q451H	0	rs79100806	T	B	0.360565	✓	Artefact
IL32	D172G	0	rs2981599	D	P	0.981577	✓	Real but in control individuals
PRSS3	A145T	0	rs855581	D	P	0.999736	✓	Artefact
TIMM23	N9D	0	rs4935252	D	P	0.999742	✓	Artefact
TIMM23	G23E	0.50/0.50 (n = 2)	rs373071373	D	P	0.85697	✓	**Real**
YY1AP1	Q424R	0.998/0.002 (n = 1315)	rs113197997	T	B	0.294756	×	N/A
MYO5B	V1703A	0.981/0.019 (n = 259)	rs138128932	NA	B	0.145661	×	N/A
MAP2K3	Q73X	0.50/0.50 (n = 2)	rs55796947	NA	NA	1	✓	Artefact
KCNJ12 ***KCNJ18***	R118Q	0.50/0.50 (n = 2)	rs1657740	T	B	0.021368	×	N/A
GPRIN2	W91R	0.50/0.50 (n = 2)	rs3127820	D	B	4.0E-6	✓	Artefact
C22orf42	M120I	0.958/0.042 (n = 3420)	rs144597334	D	B	2.91E-4	✓	Artefact
BCLAF1	T837N	0.50/0.50 (n = 2)	rs62431284	D	P	5.0E-6	✓	Artefact
CNTN5	S23A	0.50/0.50 (n = 2)	rs10790978	-	-	-	×	N/A
CNTN5	L70R	0.50/0.50 (n = 2)	rs7125822	-	-	-	✓	Artefact

P - pathogenic; D - damaging; T - tolerated; B - benign; NA - stop/gain mutation; n - number of chromosomes;
✓ - selected for further analysis; ×- not selected for further analysis; N/A – not sequenced.

3.4.1.1 Sanger sequencing validation

In order to validate whether or not the prioritised SNPs were real or sequencing artefacts, primer pairs were designed for each SNP. The primer sequences as well as the individual annealing temperature for each of the 12 SNPs are summarized in Appendix V.

The results of the Sanger sequencing are shown in Supplementary Figure 1, Appendix VI; in total the three probands and the affected sibling were sequenced. Numerous variants (A145T in *PRSS3*, Q73X in *MAP2K3*, T158I in *NOTCH2NL*, Q451H in *ANAPC1*, N9D in *TIMM23*, W91R in *GPRIN2*, M120I in C22orf42, T837N in *BCLAF1* and L70R in *CNTN5*) were identified as sequencing artefacts as Sanger sequencing revealed the patients to be homozygous for the wild type allele. However, D172G in *IL32*, G23E in *TIMM23* and S301C in *KATNAL2* were found to be real, heterozygous variants and were analysed further through various genotyping techniques in order to determine the frequency of the variants in patients and control individuals.

3.4.1.2 Frequency in ethnically matched control samples

3.4.1.2.1 Genotyping of G23E in *TIMM23*

The G23E variant in *TIMM23* was prioritised for further analysis as it was found to be a real variant across all four individuals and was found to significantly alter the protein properties; glycine is converted into glutamic acid (Figure 3.7). Glutamic acid carries a negative charge that is hypothesised to affect protein folding; moreover, it is hypothesised that the mutation of glycine into glutamic acid is likely to abolish protein functioning.

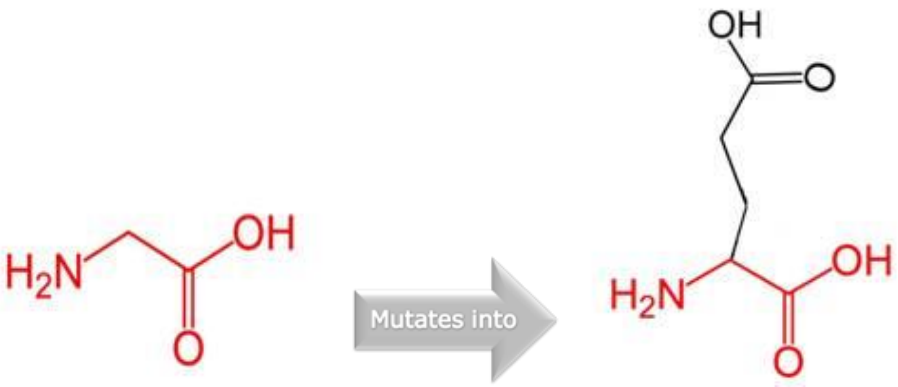


Figure 3.7 Diagrammatic representation of the amino acid change inducing the G23E variant in *TIMM23*.

The genotyping was performed by a commercial company, IKMB (Institut für Klinische Molekularbiologie – Institute for Clinical Molecular Biology) in Kiel, Germany in order to

determine the frequency of the variant in the South African PD patients as well as ethnically matched controls. A total of 1148 samples were genotyped – 458 of these were PD patients and 690 were ethnically matched controls (184 Afrikaner controls, 160 white controls, 180 mixed ancestry controls and 166 black controls). Figure 3.8 shows that the genotyping using TaqMan® SNP genotyping assay was unsuccessful; the result for the genotyping was considered to be "undetermined" as no positive genotyping calls could be made. The genotyping calls that were expected were either homozygous A/A (wild type); heterozygous A/G or homozygous G/G. It was determined that this was as a result of an erroneous sequence in dbSNP from which the primers and probes had been designed.

Given the problematic TaqMan® SNP genotyping assay, a total of ten Afrikaner control individuals were Sanger sequenced in order to determine whether or the G23E variant was, in fact, present in controls. The rationale behind this is that unaffected control patients should not harbour a variant that is carried by PD affected patients. The sequence alignment of the G23E variant in the controls shows that the SNP is in fact present in all of the controls (Figure 3.9) and the variant was thus excluded from the study for further analysis.

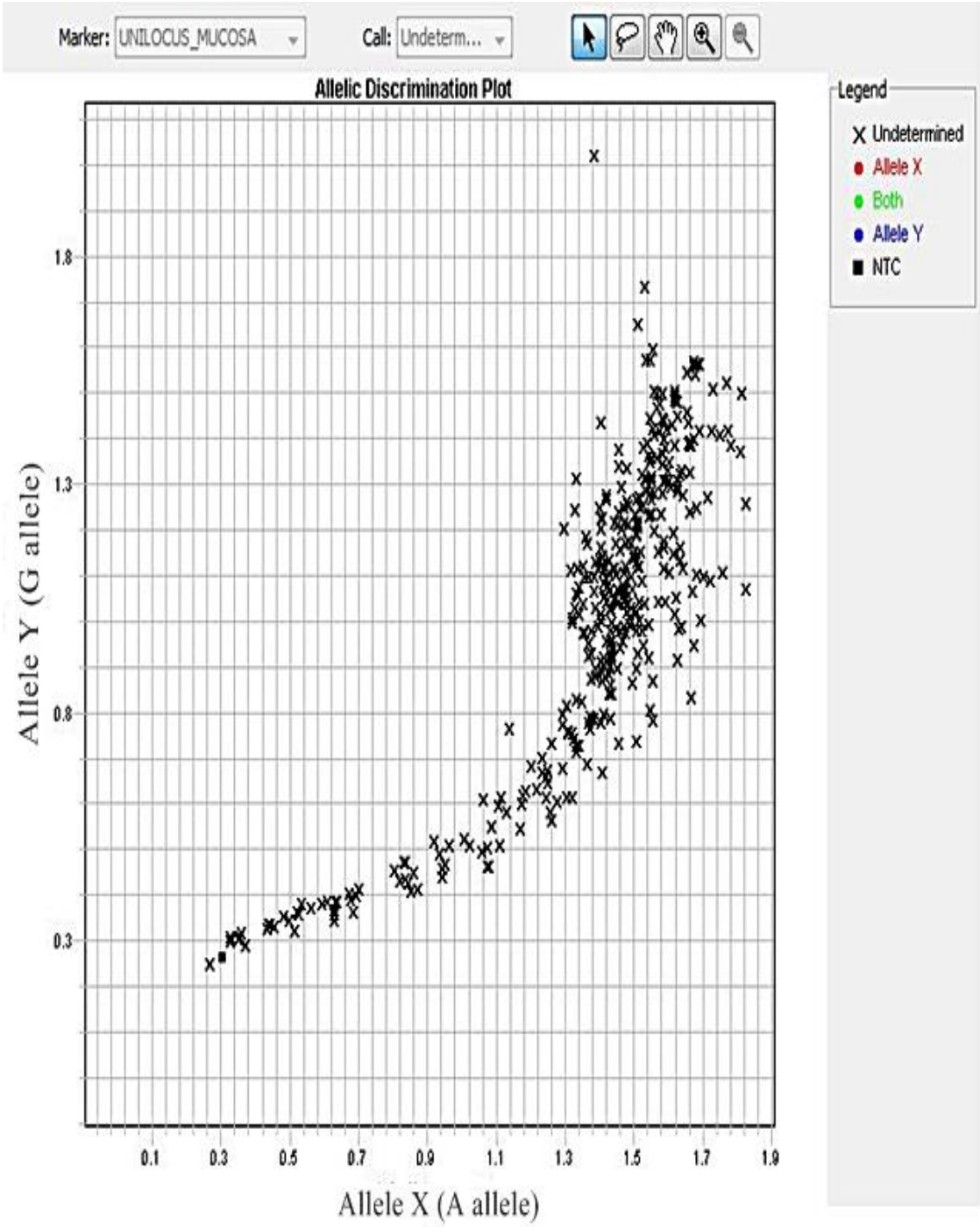


Figure 3.8 TaqMan® SNP genotyping assay result obtained from IKMB. Three heterozygous positive controls were included on the plate but no distinction could be made between any of the possible genotypes.

G23E in *TIMM23* (GGA > GAA)

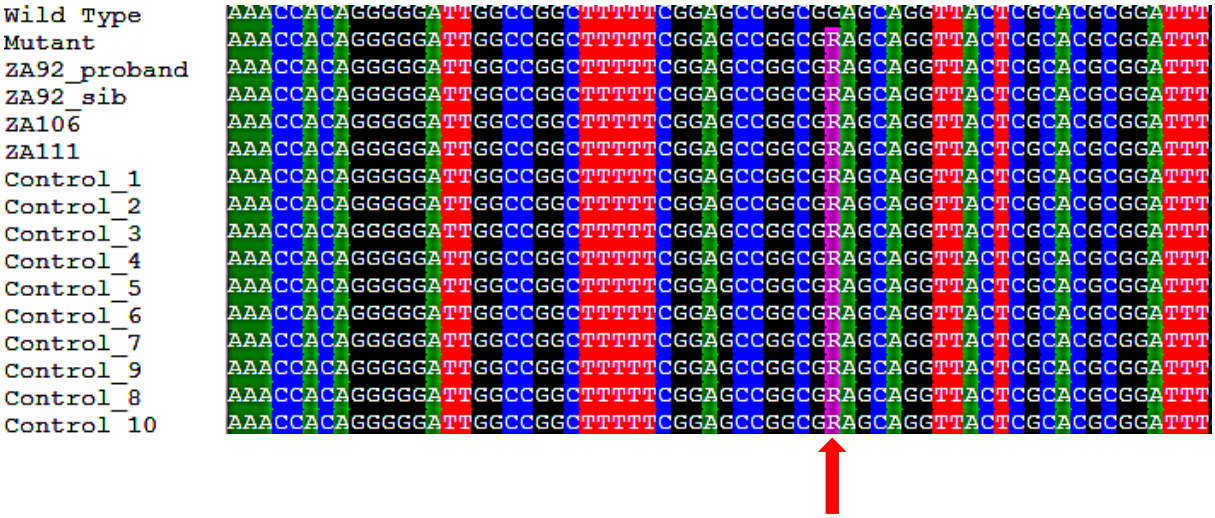


Figure 3.9 Sequence alignments of ten controls as well as the probands and affected sibling. The location of the identified SNP is indicated by the red arrow. The wild type is the reference sample, the mutant is the sample in which a heterozygous change would be present.

3.4.1.2.2 Genotyping of D172G in *IL32*

The D172G variation in *IL32* could be easily identified using both the normalized and difference graphs obtained from HRM analysis (Appendix VI, Supplementary Figure 2). Of the 31 Afrikaner controls that were screened, it was determined that ten individuals shared a melt profile that was the same as the positive controls that had been included in the run. For this reason, the ten individuals were selected and Sanger sequenced in order to determine whether or not the SNP was in fact present in these individuals. The results obtained from Sanger sequencing confirmed the presence of the D172G variant in the controls (in some cases in the homozygous form) and the variant was thus excluded from any further analysis (Appendix VI, Supplementary Figure 3).

3.4.1.2.3 Genotyping of S301C in *KATNAL2*

The S301C variation in *KATNAL2* could be easily identified using the normalized and difference graph obtained from HRM analysis (Appendix VI, Supplementary Figures 4 and 5). Of the 31 Afrikaner controls that were screened, it was determined that one sample shared a melt profile that was the same as the positive controls that had been included in the run. Additionally, there were numerous control samples that had altered HRM profiles and for this reason were also selected for Sanger sequencing. The results obtained from Sanger sequencing confirmed the presence of the S301C variant in five of the controls (Appendix VI, Supplementary Figure 6). Therefore, the variant was excluded from further analysis.

3.4.2 Analysis of family ZA92 only using the hypothesis based approach

The difficulties in identifying a single variant across two families that may be accountable for PD in the South African patient cohort resulted in a new approach to data analysis and focus was placed on family ZA92. A comparison of WES data was conducted between the affected siblings and the unaffected family member (ZA_92_unaffected, related control) that was also subjected to WES. Subsequent bioinformatics identified a plausible variant, P1150S (rs34344550) in *EFCAB6* that met all of the necessary filtering criteria as previously established. The *EFCAB6* gene produces the EF-hand calcium-binding domain containing protein 6, a protein that negatively regulates the androgen receptors by recruiting histone deacetylase complex. Moreover, the protein DJ-1 (known to be involved in PD) antagonises this reticence by the abolition of this complex (http://www.ncbi.nlm.nih.gov/gene/64800; Niki et al. 2003). The proline to serine substitution was found to significantly alter the properties of the expressed protein (Figure 3.10). The mutant residue is considerably smaller than the wild type residue and the wild type residue is considered to be more hydrophobic than the mutant residue and it is hypothesised that significant hydrophobic interactions that may occur, either on the surface or in the core of the protein, may be lost due to this alteration in amino acid sequence. Additionally, prolines are recognized as molecules that have an extremely rigid structure, therefore sometimes forcing the backbone into a specific conformation.

Figure 3.10 Diagrammatic representation of the amino acid change inducing the P1150 variant in *EFCAB6*.

The strong evidence that the P1150S variant in *EFCAB6* may alter protein folding and therefore possibly protein functioning led to further investigations as to whether or not this gene (and variant) may play a role in PD. The protein product of *EFCAB6* is more

commonly known as DJ-1 binding protein, and the protein actively interacts with DJ-1 with the interaction spanning amino acid numbers 372 -570 of the DJ-1 protein (Niki et al. 2003). The global frequency of the P1150S variant has been reported to be low in six ethnic groups that have been extensively studied (Table 3.11).

Table 3.11 Global frequency data of P1150S in *EFCAB6.*

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
South Asian	222	16 606	2	0.0134
European (Non Finnish)	474	67 658	3	0.0070
Latino	54	11 602	0	0.0047
African	12	10 556	0	0.0011
East Asian	2	8 766	0	0.0002
European (Finnish)	1	6 732	0	0.0001
Totals	**765**	**121 938**	**5**	**0.00440063**

*Table taken and adapted from ExAC Browser (http://exac.broadinstitute.org/).

Moreover, subsequent communication with our collaborators at the Institut du Cerveau et de la Moelle épinière (ICM) in Paris, France revealed the frequency of the P1150S variant was recorded at 0.0039457 in more than 15 000 control individuals.

Given the insurmountable evidence that the P1150S variant was found at a low frequency in multiple ethnic groups as well as the fact that the gene in which the variant was identified is a confirmed interactor of a known PD gene, namely *DJ-1*, further investigation into the possible pathogenic effects of the variant was deemed necessary.

3.4.2.1 Sanger sequencing validation

Sanger sequencing was performed for validation purposes and P1150S variant was found to be real in the affected sibling pair, but more importantly was found to be absent in the related, unaffected control (Figure 3.11).

P1150S in *EFCAB6* (CCC > TCC)

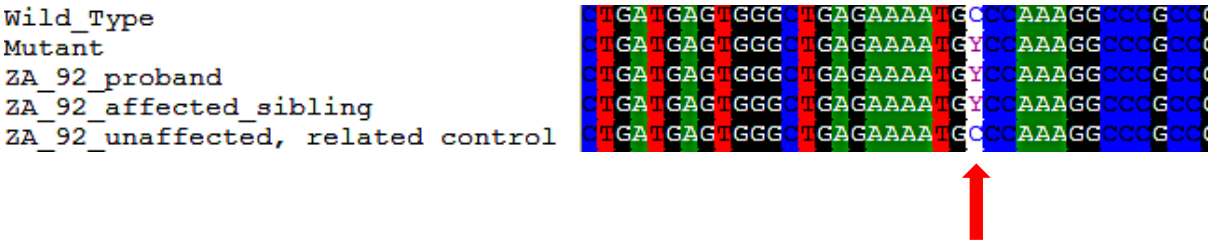


Figure 3.11 Sequence alignments of ZA92 family as well as an unrelated, unaffected control for the P1150S variant in *EFCAB6*. The location of the identified SNP is indicated by the red arrow. The wild type is the reference sample, the mutant is the sample in which a heterozygous change would be present.

3.4.2.2 Genotyping of P1150S in *EFCAB6*

Genotyping was performed using HRM for both variants and two positive controls namely ZA92 and ZA92 (affected sibling) were included in each run. A total of 690 ethnically matched control samples were sequenced. The ethnic breakdown of the control samples were as follows: 184 Afrikaner controls, 160 white controls, 180 mixed ancestry controls and 166 black controls. The P1150S variant could be easily identified using the normalized and difference graphs obtained through HRM analysis (Figure 3.12 and 3.13).

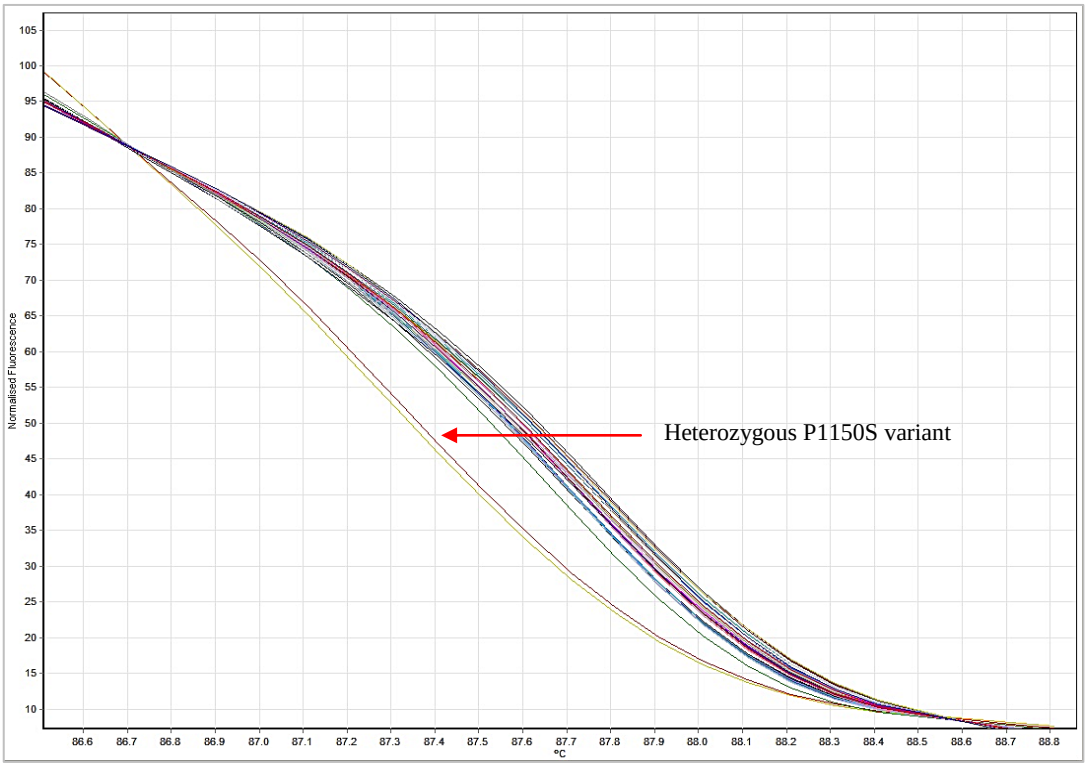


Figure 3.12 HRM normalised graph indicating the heterozygous P1150S variant in the sequence confirmed positive controls.

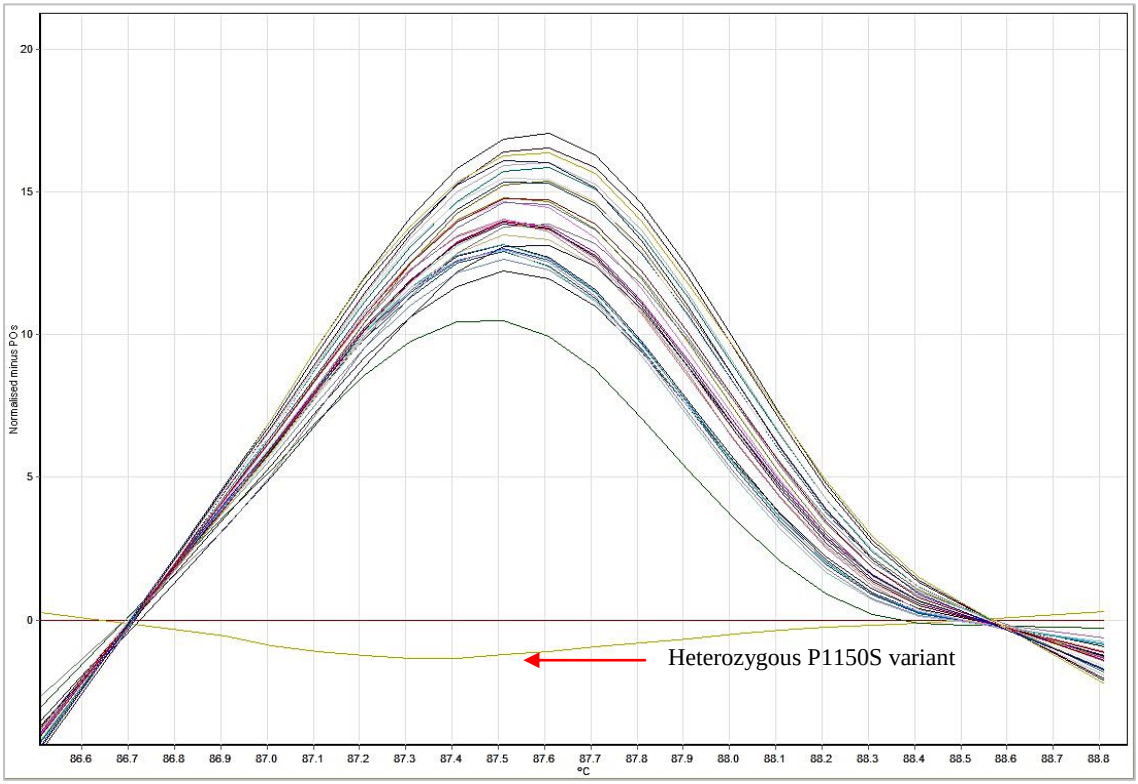


Figure 3.13 HRM difference graph indicating the heterozygous P1150S variant in the sequence confirmed positive controls.

Of the 690 control samples that were screened for the variant, none of the control samples were identified as carriers of the P1150S variant. Subsequently, a total of 458 PD probands were screened for the variant, given its absence in the control population and any samples that exhibited altered HRM profiles were subjected to Sanger sequencing however none of the query samples were confirmed to carry the variant. The frequency of the P1150S variant was concluded to be 0.22% (1/458) in the probands and 0% (0/690) control samples.

Family members of the individual who was originally identified as a carrier of the P1150S variant were sequenced in order to identify additional P1150S–carrying individuals. The variant was identified in both of the children of the proband (Figure 3.14). Currently, the two children are younger than the identified AAO of their mother (55 years). Subsequently, the children of the affected sibling, ZA92_sib were also sequenced and the variant was also identified in these individuals. Both of the children are also younger than the AAO of their father (68 years). Given the fact that the cousins are all younger than the AAO of disease in their respective parents, it is not possible to rule out the fact that P1150S may be pathogenic. However, no symptoms of PD in the children are currently apparent.

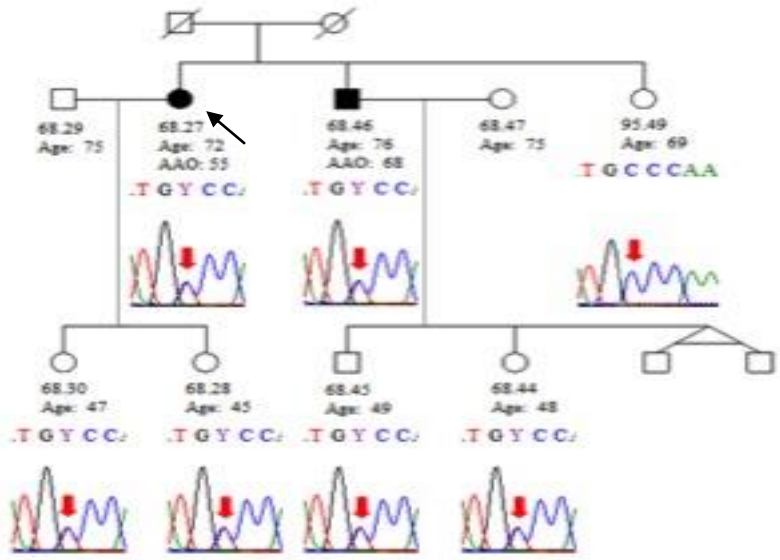


Figure 3.14 Sequencing results from the proband with the P1150S variant and additional family members. The variant is present in both children of the proband, one of the siblings and not in the unaffected sibling.

Following the identification of P1150S in *EFCAB6* in the South African population, it was necessary to determine whether or not the variant had been identified in other PD individuals and simultaneously determine whether or not the variant possibly co-segregates with the disease in this population. It was determined, however that the P1150S variant does not co-segregate with the disease in five Canadian families and it was found in a heterozygous state in five patients and six healthy controls (Prof. Matthew Farrer, personal communication) – for this reason, the variant was excluded as a possible disease-causing variant in the South African PD population.

3.4.3 Analysis of WES data using the hypothesis-free approach

The identification of a single gene and variant that could be attributed to PD in the South African patients was proving to be a challenge and for this reason, we decided to use a hypothesis-free approach to prioritise sequence variants. To this end, we designed a novel in-house bioinformatics pipeline for the filtering of variants, called TAPER™. In order to determine whether or not TAPER™ was an efficient means for data analysis, proof of concept experiments were conducted. This was performed on existing WES data for which a disease-causing gene and variant had been identified. The datasets that were used were sourced from various collaborators (Dr. Suzanne Lesage, Institut du Creveau et de la Moelle

épinière, Hôpital Pitié Salpêtrière, Paris, France; Mr Daniel Evans, Centre for Applied Neurogenetics, University of British Columbia, Vancouver, Canada) as well as files sourced from previously published papers; the datasets used were those that previously identified variants in Parkinson's disease as well as severe intellectual disability and microcephaly and ataxia and myoclonic epilepsy (Pena and Coimbra 2015; Srour et al. 2015). The results for the proof of concept study are shown in Appendix VII Supplementary Table 1. TAPER™ was successfully used to identify the same variants that had previously been implicated in specific diseases. TAPER™ was subsequently applied to the original six family pedigree rather than to the larger 40 family pedigree. The rationale behind TAPER™ and the steps that it uses is comprehensively explained in Section 2.6.2, page 54 and it is unique in that it assumes no phenotypic or clinical distinctions as a means of prioritisation when analysing WES data. The results from this prioritization method are shown in Table 3.12.

Following the identification of variants in each individual proband that was subjected to WES using TAPER™, a new candidate gene list was constructed. The construction of the new candidate list involved the following steps:

- Variants that overlap across the affected sibling pair in ZA92 but are absent from all controls – a total of 45 variants in 39 genes were found to overlap;
- Variants that were found to be in shared segments (as identified in Section 3.2.1.1);
- Variants that are found in at least three of the four WES individuals (ZA92, ZA92 (affected sibling), ZA106 or ZA111);
- Variants that fall into chromosomally shared regions;
- Prioritization of variants that were found to be involved in other movement disorders or any pathways previously associated with PD;
- Variants with high Combined Annotation Dependent Depletion (CADD) scores. CADD is a method for objectively integrating numerous diverse annotations into a single score (C-score) for each variant. A scaled CADD score of 20 means that the variant is amongst the top 1% of deleterious variants; a scaled CADD score of 30 means that the variant is in the top 0.1% of deleterious variants. For this reason, the higher the CADD score, the more likely the variant is to be pathogenic (Kircher et al. 2014).

A list of 20 variants in 20 genes were selected and of these, a total of six of the prioritised variants were selected for further analysis based on the function and possible link to disease (Table 3.13).

Table 3.12 Summary of the total number of variants obtained through each filtration step.

	ZA92 proband	**ZA92 sibling**	**ZA106**	**ZA111**
Total variants in VCF file	74 938	76 117	77 700	79 070
Total number of variants assigned to exonic regions by wANNOVAR	19 596	19 955	20 881	20 966
All synonymous and non-frameshifts removed	9 479	9 646	10 237	10 251
Remove all variants with a frequency >1% in 1KGP	1 086	1 127	1 403	1 357
Remove all variants with a frequency >1% in ESP6500	911	965	1 193	1 156
Remove all variants with positive FATHMM scores	384	416	469	474
Remove all variants with negative GERP+++ scores	246	265	282	293
Remove all variants on X and Y chromosomes	242	261	276	286
Variants linked to relevant diseases	**56**	**54**	**46**	**68**

*VCF - Variant Called Format; 1KGP - 1000 Genomes Project; ESP6500 - Exome Sequencing Project 6500; GERP - Genomic Evolutionary Rate Prediction; FATHMM - functional analysis through hidden Markov Models

Table 3.13 Shortlist of candidate genes prioritised for further analysis.

Gene	SNV	1KGP_freq	ExAC_freq	ESP6500_freq	dbSNP	CADD score	Selected for further analysis
CASP7	E43D	.	8.23E-06	.	.	29.47	Yes
WNK1	S719G	.	0.001785	.	.	31.01	Yes
MAST2	T867R	.	.	.	.	15.99	CADD score too low
DCDC2B	V78A	0.0014	5.28E-03	0.0058	rs144804850	26.80	No neurological disease association
CACNA1E	R2157Q	0.0002	4.65E-04	0.001	rs2480373	35.00	No neurological disease association

MIPEP	V626M	.	4.07E-05	.	.	26.91	Yes
SETD8	L332P	.	.	.	rs61955127	23.40	No neurological disease association
SLC5A2	N654S	0.002	5.27E-03	0.007	rs61742739	17.94	No neurological disease association
SYNJ1	V1405I	.	1.65E-04	.	rs79652470	29.40	Yes
EPB41L5	R28C	0.0016	3.19E-03	0.0042	rs141466977	14.78	No neurological disease association
PKP4	D1127V	0.0016	3.21E-03	0.0036	rs148782148	18.59	No neurological disease association
CYP1B1	R368H	0.0042	5.94E-03	0.0016	rs79204362	24.91	No neurological disease association
CYP2D6	R365H	.	0.12	.	rs1058172	34.00	No; frequency too high
KLHL5	R36S	0.004	7.46E-03	0.0091	rs140053546	7.59	CADD score too low
WDR36	R529Q	0.0002	7.81E-04	0.0013	rs116529882	30.00	No neurological disease association
SYNJ2	M534T	0.0008	5.26E-03	0.0043	rs140670406	18.25	Yes
REEP4	R209W	.	3.25E-05	.	.	12.26	CADD score too low
USP17	C357S	.	7.863e-05	.	.	34.89	Yes
COL27A1	R1688Q	.	2.03E-03	0.0011	rs149629527	None	No neurological disease association
COL6A5	L2591X	.	.	.	.	8.78	CADD score too low

CASP7 – Caspase 7, Apoptosis-Related Cysteine Peptidase; WNK1 - Lysine Deficient Protein Kinase 1; MIPEP - Mitochondrial Intermediate Peptidase; SYNJ1 - synaptojanin1; SYNJ2 – synaptojanin2; USP17 - Ubiquitin Specific Peptidase 17; SNV – single nucleotide variant; CADD – combined annotation dependent depletion

3.4.3.1 Sanger sequencing validation

A total of six variants were shortlisted for further analysis based on the fact that they had previously been associated with other movement disorders or had been directly implicated in pathways that had previously been associated with PD and other movement disorders and were validated using Sanger sequencing. The primer sequences as well as the annealing temperature for each of the SNPs is summarised in Supplementary Table 2, Appendix VII. The results of the genotyping are shown in Table 3.14 and the Sanger Sequencing and genotyping results are shown in Appendix VIII. Three of the candidate variants namely S719G in *WNK1*, E43D in *CASP7* and V626M in *MIPEP* were excluded from the analysis as they were found in the control individuals. The M534T variant in *SYNJ2* was not sequenced or genotyped as it was determined that this variant does not co-segregate with disease (Prof. Matthew Farrer, personal communication). Two of the remaining variants were found to be of interest, namely V1405I in *SYNJ1* and C357S in *USP17*.

Table 3.14 Summary of the genotyping results obtained for the six variants shortlisted for further analysis.

Gene	SNV	Does the variant co-segregate with the disease?	Frequency in Controls				Frequency in Patients				
			Afrikaner (n=184)	White (n=160)	Black (n=166)	Mixed Ancestry (n=180)	Afrikaner (n=148)	White (n=175)	Black (n=26)	Mixed Ancestry (n=104)	Indian (n=5)
WNK1	S719G	Unknown	3 (2.03%)	1 (0.63%)	2 (1.2%)	3 (1.67%)	DNS				
CASP7	E43D	Unknown	2 (1.09%)	1 (0.63%)	1 (0.60%)	0	DNS				
MIPEP	V626M	Unknown	1 (0.54%)	1 (0.63%)	2 (1.2%)	1 (0.56%)	DNS				
SYNJ1	V1405I	Unknown	0	0	0	0	1 (0.68%)	0	0	0	0
SYNJ2	M534T	No	DNS				DNS				
USP17	C357S	Unknown	1 (0.54%)	0	0	0	12 (8.11%)	2 (1.15%)	0	4 (3.85%)	0

CASP7 – Caspase 7, Apoptosis-Related Cysteine Peptidase; WNK1 - Lysine Deficient Protein Kinase 1; MIPEP - Mitochondrial Intermediate Peptidase; SYNJ1 - synaptojanin1; SYNJ2 – synaptojanin2; USP17 - Ubiquitin Specific-processing Peptidase 17; SNV – single nucleotide variant; DNS – did not sequence

3.4.3.2 Frequency in ethnically matched control samples

3.4.3.2.1 Genotyping of V1405I in *SYNJ1*

In silico analysis was performed on the V1405I variant using the methods previously described. In SYNJ1 sequence O43426 (Swiss-Prot) (NP_003886.3), the variation is at residue 1366. In the nucleotide to protein translation of NM_001150306, it is at position 1319. For the purposes of this study, the O43426 Swiss-Prot sequence was used. It should be noted that there are four isoforms for the SYNJ1 protein and the V1366I variant (corresponding to V1405I) is only present in isoform 1. It was determined that there is a catalytic domain that lies from position 500-899 as well as two conserved domains within this protein, namely the SAC domain (from residues 119-449) and an RNA recognition motif (RRM) domain (from residues 902-971). The SAC domain is a region of homology between the N terminus of synaptojanin and the otherwise unrelated yeast protein Sac1p and is approximately 400bp in length (www.ebi.ac.uk/interpro/entry/IPR002013). The RRM motif is on average 90 amino acids that are known to bind to single stranded RNAs (www.ebi.ac.uk/interpro/entry/IPR000504). Additionally, this protein carries a 3 x 3 amino acid (Asparagine, Proline and Phenylalanine) repeats at positions 1396-1398, 1406-1408 and 1417-1419. It was therefore concluded that the residue of interest at position 1366 does not fall into any of the conserved or catalytic domains. In addition to the variant at position 1366 in this isoform, there is an additional natural variant at position 1388 namely the V1388A variant. In terms of the physio-chemical properties, this is a change from a medium sized hydrophobic residue to a small sized hydrophobic residue.

Variant V1366I was analysed and it was determined that the amino acid substitution is a conservative one – both of the proteins are hydrophobic and similar in size (Figure 3.15). However, it was not possible to model the section of the protein containing the variant because of the lack of a suitable template.

Valine **Isoleucine**

Figure 3.15 Diagrammatic representation of the amino acid change inducing the V1366I variation in *SYNJ1*.

Following the examination of the amino acid substitution, a scan for eukaryotic linear motifs (ELM) (Dinkel et al. 2012) was conducted on the sequence. It was determined that there are a number of potential phosphorylation sites in this vicinity:

1. Residues 1368 and 1375: Glycogen Synthase Kinase (GSK) 3 phosphorylation site
2. Residue 1368: Casein Kinase (CK) 1 phosphorylation site
3. Residue 1359 and 1365: PI3 Kinase-related Kinase (PIKK) phosphorylation site

In addition to these phosphorylation sites, it was determined that there is a potential binding motif (residues 1361 and 1365) for the Ubiquitin Specific Protease (USP)7 deubiquitinating enzyme. It is, however very difficult to draw any definitive conclusions as to what the effect of the variant may be on the protein. The so-called natural variant, V1388A is not anticipated to have an effect on the function of the protein and the replacement of the Valine with Isoleucine is conservative. However, due to the proximity of the variant, there is a possibility that the V1366I substitution could interfere with phosphorylation or the binding by the USP7 deubiquitinating enzyme. However, it was not possible to determine the position of the residue in the 3 dimensional structure. There is no solved structure in the protein data bank and therefore we were unable to generate a reliable model due to the lack of a suitable template.

The V1405I variant was found to be homozygous and only in the affected sibling pair. However, this variant was not found in any of the controls that were subsequently selected for genotyping and Sanger sequencing validation in the patients (Figure 3.16 and 3.17). This variant can be easily identified through the use of HRM and was identified in 0.22% (1/458) probands and in none of the controls.

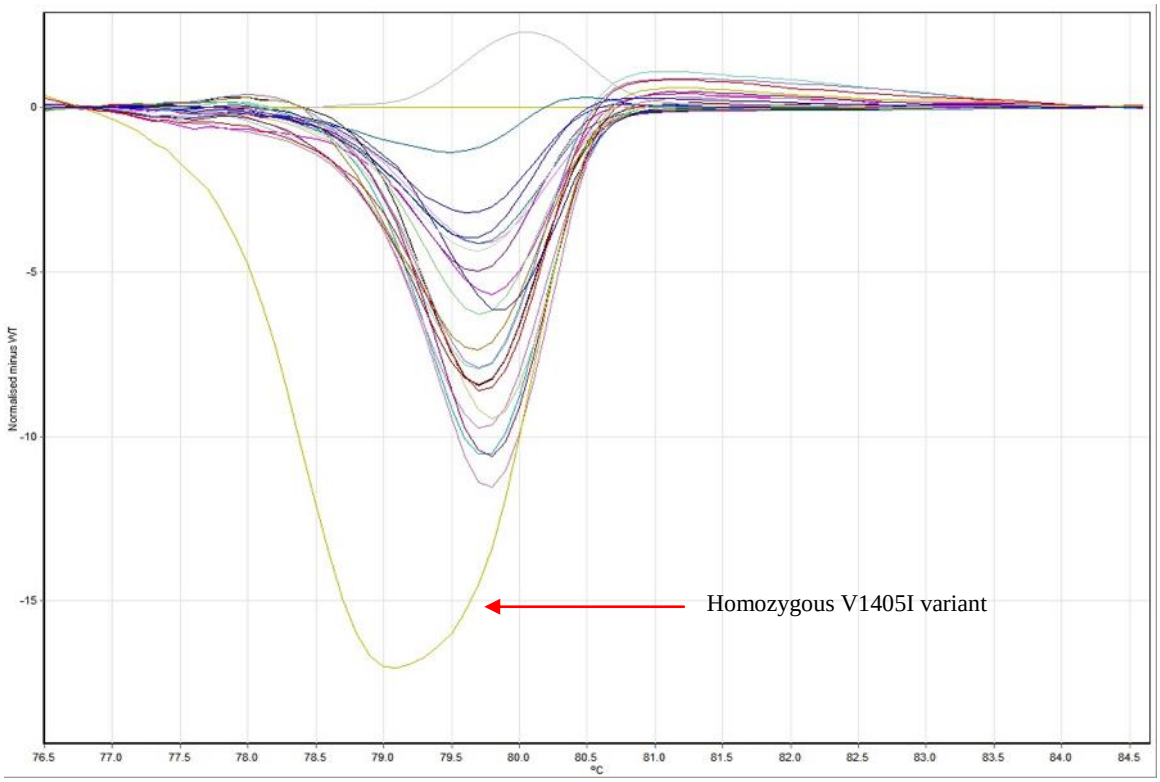


Figure 3.16 HRM difference graph for the V1405I variant in *SYNJ1*. The homozygous variant is indicated by the red arrow and can be easily distinguished from wild type samples.

V1405I in *SYNJ1* (GTT > ATT)

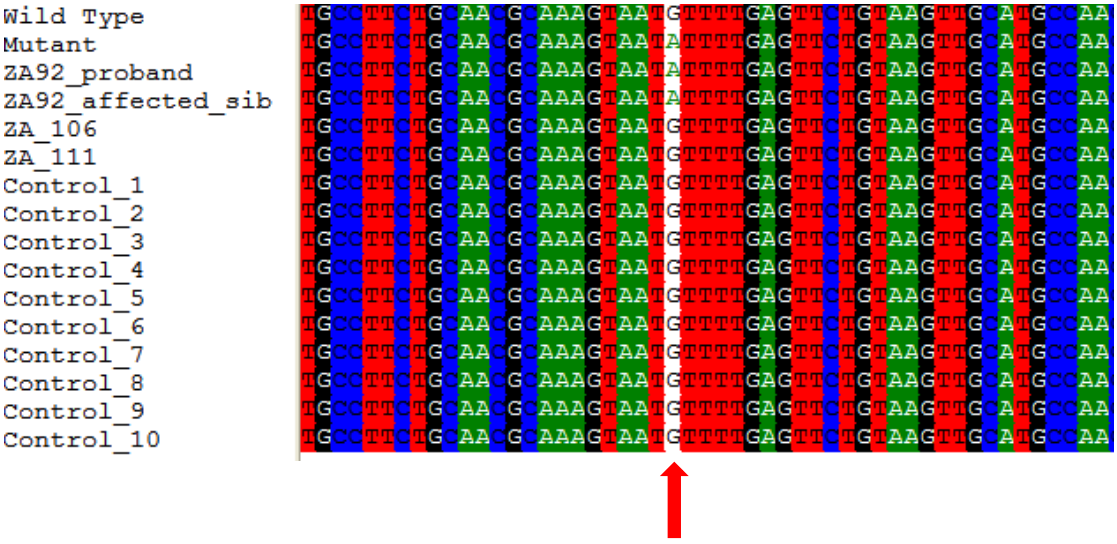


Figure 3.17 Sequence alignments of selected samples for the V1405I variant in *SYNJ1*. The location of the SNP is indicated by the red arrow. The wild type is the reference sample while the mutant sample is a construct indicating the position of the variant in a homozygous state.

3.4.3.2.2 Genotyping of C357S in *USP17*

In silico analysis was performed on the C357S variant using the methods previously described. In USP17 sequence Q6R6M4 (Swiss-Prot) the variation is at residue 357, as previously expected. It was determined that there is an Ubiquitin Specific-processing Protease (USP) domain that spans the position 80-375. The USPs make up the largest family of deubiquitinating enzymes and ubiquitination is a reversible process that affects a number of cellular processes such as protein degradation, trafficking, cell signalling and DNA damage response. USP is composed of a conserved catalytic core that is interspersed at five independent points with insertions; these insertions may be as large as the catalytic domain itself. The insertions are capable of folding into independent domains that are involved in the regulation of deubiquitinase activity (http://www.ebi.ac.uk/interpro/entry/IPR028889). The region is conserved across multiple species and interestingly, the variant falls within this domain, at position 357.

Variant C357S was analysed and it was determined that the amino acid substitution was not a conservative one – the hydrophobic Cysteine is substituted with a hydrophilic Serine amino acid; however the amino acids are similar in size (Figure 3.18). It was not possible to model the section of the protein that contains this amino acid, as there is no suitable template.

Cysteine **Serine**

Figure 3.18 Diagrammatic representation of the amino acid change inducing the C357S variant in *USP17*.

Following the examination of the amino acid substitution, a scan for eukaryotic linear motifs (ELM) (Dinkel et al. 2012) was conducted on the sequence. It was determined that there are a number of potential phosphorylation sites in this vicinity:

1. Residues 357 and 364: Glycogen Synthase Kinase (GSK) 3 phosphorylation site
2. Residue 358-340: Casein Kinase (CK) 1 phosphorylation site

3. Residue 361-367: PI3 Kinase-related Kinase (PIKK) phosphorylation site

In addition to these phosphorylation sites, it was determined that there is a potential binding motif (residues 229-233 and 305-309) also for the USP7 deubiquitinating enzyme. It is, however very difficult to draw any definitive conclusions as to what the effect of the variant may be on the protein. It was not possible to determine the position of the residue in the 3 dimensional structure. There is no solved structure in the protein data bank and therefore we were unable to generate a reliable model due to the lack of a suitable template.

The homozygous C357S was found to be real and in 18 PD probands (12 Afrikaner, 2 White and 4 Mixed Ancestry) – however, one Afrikaner proband was found to carry this variant in a homozygous state. The Sanger sequencing results for this variant is shown in Figure 3.19 and the HRM difference graphs for the genotyping *USP17* is shown in Figure 3.20. This variant can be easily identified through the use of HRM and was identified in 3.93% (18/458) probands and 0.14% (1/690) controls.

C357S in *USP17* (TGT > TCT)

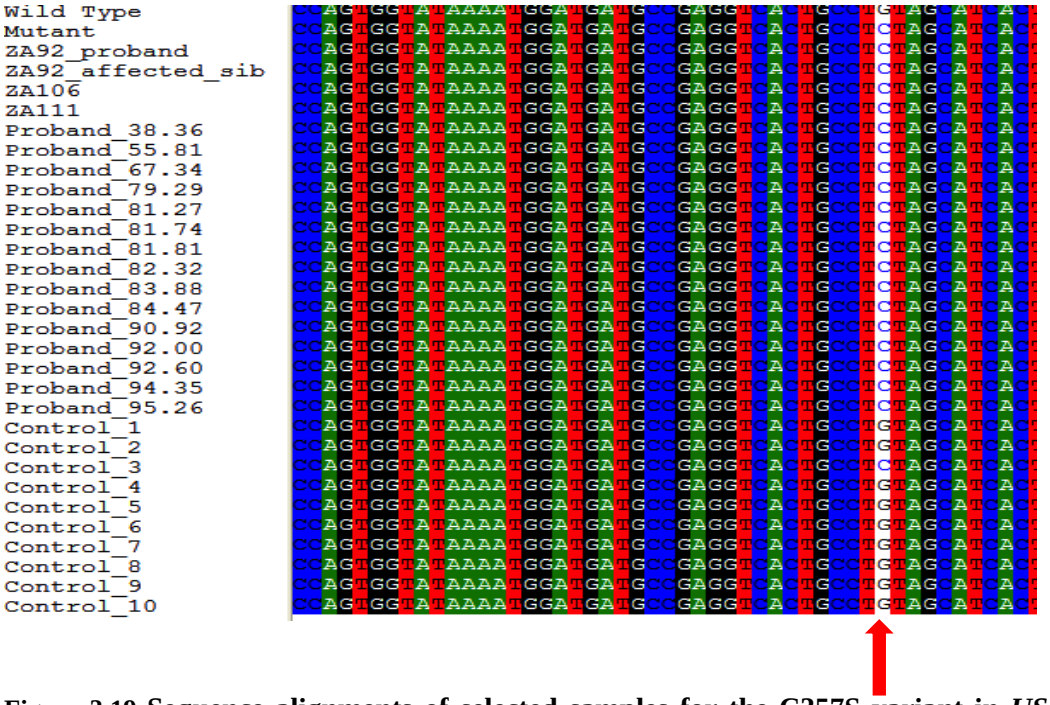


Figure 3.19 Sequence alignments of selected samples for the C357S variant in *USP17*. The location of the SNP is indicated by the red arrow. The wild type is the reference sample while the mutant sample is a construct indicating the position of the variant in a homozygous state.

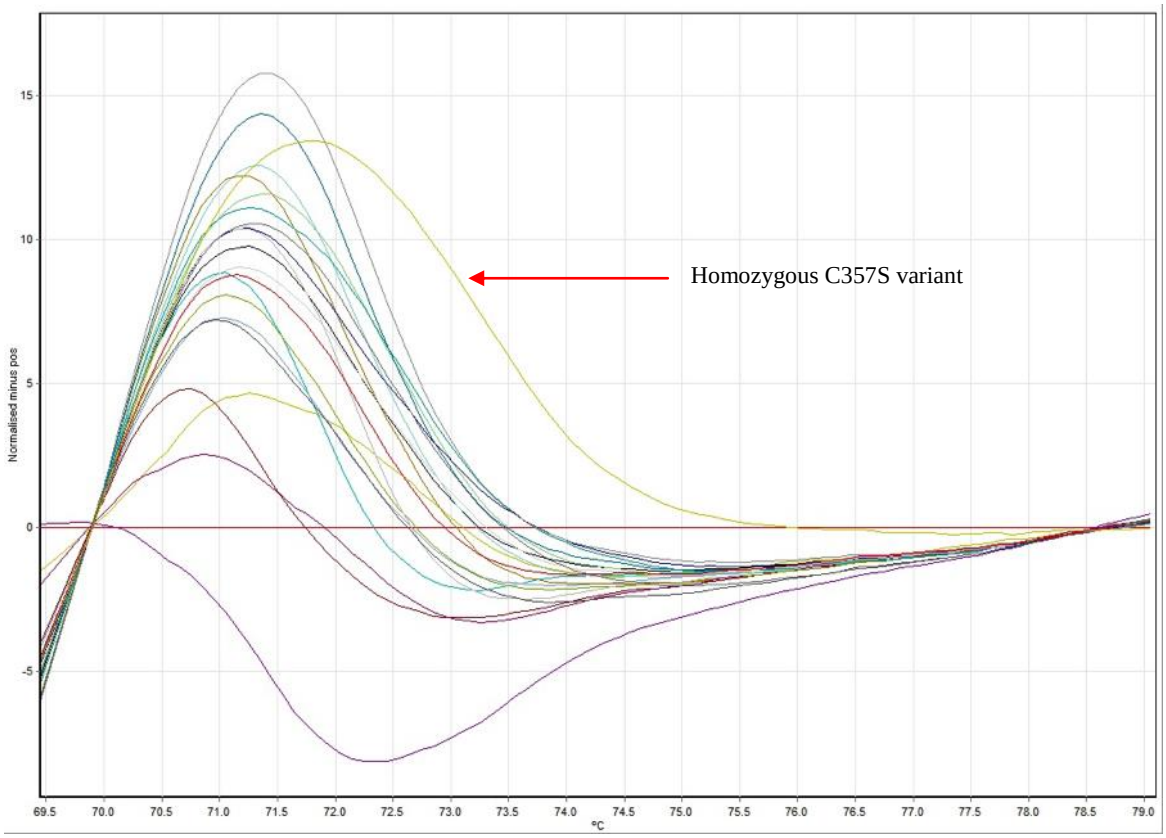


Figure 3.20 HRM difference graph for the C357S variant in *USP17*. The homozygous variant is indicated by the red arrow and can be easily distinguished from wild type samples.

CHAPTER 4: DISCUSSION

INDEX **PAGE**

4.1 Genealogical analysis 107

4.2 Whole genome SNP array 108

4.3 Whole exome sequencing 109

4.3.1 Analysis of P1150S in *EFCAB6* identified in family ZA92 111

4.3.2 Analysis of WES using a hypothesis-free approach 112

4.3.2.1 V1405I in *SYNJ1* 113

4.3.2.2 C357S in *USP17* 118

4.4 Limitations of the study 120

4.5 Future work 122

4.6 Concluding remarks 124

CHAPTER 4: Discussion

The present study undertook an investigation into whether a founder effect for PD might exist in the South African Afrikaner patients and this was analysed using genealogical tracking and WES. Moreover, the study aimed to determine whether this putative disease-causing mutation could be attributed to the development of PD in other South African ethnic groups. It has been determined that the known PD genes do not play a significant role in the South African PD cohort and it was therefore sought to identify a novel variant. WES was performed on three PD probands and an affected sibling and a number of candidate genes were identified using a hypothesis-based approach, however, these were identified as either sequencing artefacts or as high frequency variants in ethnically matched controls. A novel variant prioritisation tool TAPER™ was designed and applied to the WES data and a number of plausible candidate genes were identified. Of these, two genes (*SYNJ1* and *USP17*) were prioritised for verification and further study. *SYNJ1* had previously been identified, as a role player in early-onset PD while the protein product of *USP17* is involved in the deubiquitination of tagged proteins so as to regulate cellular processes. *In silico* analysis was employed to determine the possible effects of each of the variants on the protein product and how this may contribute to the pathobiology of disease.

4.1 Genealogical analysis

In a previous study conducted by our group, a total of 262 South African PD patients were screened for the known genes that cause PD (B Glanzmann, MSc thesis, March 2013). Of these, 76 (29.0%) were self-identified as Afrikaner and none of the known PD genes played any significant role in disease in these individuals. Genealogical tracking was subsequently performed on 12 of the Afrikaner probands and complete pedigrees were constructed for each of the families. It was then determined that six of these Afrikaner individuals could be traced back to a common founder couple who had immigrated to the country in the 1600s.

Following the discovery that six of the apparently unrelated PD probands trace back to a common founder couple, genealogical analysis was conducted on additional selected Afrikaner patients. Since the advent of the present study, a total of 48 Afrikaner probands were selected for genealogical analysis based on the age at onset of the disease as well as the positive family history of PD. Of these, a total of 40 probands traced back to the same common founder couple. Given that some progenitors of Afrikaners have large numbers of

present day descendants, it has been suggested that any two present day Afrikaners may share these progenitors as ancestors to some degree (Greeff 2007). However, this does not seem to be the case and this is illustrated with genealogical data that is available for long QT syndrome (Geldenhuys et al. 2014). Here, 22 families were shown to share a common founder couple through the construction of 12 complete and ten partial ancestral charts. In the 12 complete ancestral charts that had been constructed, only four were found to have direct ancestral lines to the founder couple. Given that the total number of both male and female progenitors for the Afrikaner population between the period of 1657 and 1806 is estimated at 4000, the likelihood that 40 randomly selected Afrikaner individuals would share a common founder couple is rather unlikely (Geldenhuys et al. 2014).

4.2 Whole genome SNP array

SNP genotyping is used to outline the nature and the extent of chromosomal variation, analyse population genetic structure and to find specific loci that may contribute to disease. SNPs are used as proxies for the unobserved sequence variants in the surrounding DNA, thereby allowing for the measurement of the flow of genetic material through populations (Manolio, Brooks, and Collins 2008).

IBD is the foundation for many of the significant problems in genetics, some of which include haplotype phase, an understanding of familial diseases and the detection of population structure (Browning and Browning 2013). For the majority of applications, it is useful to know not just whether or not two alleles are identical at a specific locus but whether IBD extends on either side of the locus, thereby giving an indication of segmental sharing (Browning and Browning 2013). This is important as it allows for the identification of a chromosomal region with a specific length that is transmitted from a common ancestor without recombination (Browning and Browning 2013; Glodzik et al. 2013). In general, the longer the shared segment between two individuals, the more recent the ancestor (Speed and Balding 2015).

We performed the whole genome SNP array on the 40 Afrikaner probands; in addition to this, we included one affected sibling, one unaffected sibling, two QC samples and four additional, unrelated, unaffected control individuals all of whom were at an age that was older than the AAO of the probands when they were recruited for the study. Due to costs and

logistical reasons (24 samples could be genotyped per array), we were only able to include four controls. The inclusion of the affected and unaffected siblings was to determine whether or not our IBD calculations were correct as siblings are expected to have an IBD of 0.5. Although only the data for the affected sibling is shown in the results, when the same analysis was conducted using the unaffected sibling, an IBD of 0.5 was obtained for all three siblings, suggesting that the IBD calculations were correct. The whole genome SNP array was to determine whether or not the 40 probands were, in fact, related to one another through the calculation of IBD and the identification of segmental sharing. In addition to this, we aimed to determine whether randomly selected Afrikaner control individuals were related to any of the 40 probands. We successfully determined that all 40 of the Afrikaner probands were related to one another with varying degrees of relatedness (ranging between 8-12 generations back), which provided the necessary genetic support for the genealogical data. Furthermore, a total of three of the four control individuals were shown to share no relatedness between any of the probands or any of the controls. However, one of the control individuals, namely 11.937 (Control_4) shared a significant degree of IBD with the 21 of the probands. This could be due to a number of factors one of which may be that although this individual stated that they had no family history of PD, this may not actually be the case. Moreover, our control population is a relatively small one – for the purposes of this study, only 184 Afrikaner controls were available for the study and for this reason, more controls need to be included in future work so as to see whether the phenomenon seen in Control_4 is a common occurrence in the Afrikaner population.

4.3 Whole exome sequencing

Given the strong supporting evidence that the 40 Afrikaner probands may have PD for the same genetic reason due to their degree of relatedness, a NGS approach, namely WES was employed on select individuals for novel gene discovery. Three of the Afrikaner probands and one affected sibling were selected for WES in order to identify common variants that may be attributed to disease. Approximately 78 000 variants were identified per individual and following sample QC, VCF file generation and variant calling, there were approximately 20 000 variants for analysis.

High throughput sequencing technologies such as WES have shown a rapid development and have made a significant impact on how genetic research is being conducted especially in cases where the genetic cause for a disease is unknown (Pabinger et al. 2013). WES has

become technically feasible and cost-effective. It can typically yield hundreds of thousands of variants per sequenced individual; however, advances in WES technologies coupled to a lower cost of sequencing have resulted in a data deluge that poses numerous challenges and threatens to overwhelm the analytical capacity of many laboratories and possibly generate an analysis bottleneck (Pabinger et al. 2013).

There are currently more than 600 bioinformatics tools available for data analysis and interpretation (van Dijk et al. 2014). These tools include those that assess the quality of short reads and those that can be used for sequence alignment. Also, there are relatively standard means for obtaining a processed file that can be further scrutinized for variant identification. Initial bioinformatics analysis conducted on the four WES individuals made use of a combination of an in-house or custom method, whereby variants across all four affected individuals were compared and analysed using open source software and basic scripting methods coupled to variants called by the commercially available software program, IVA (www.ingenuity.com). Filtering of these variants left a total of 24 variants in 22 genes of which 12 were selected for further analysis. Of these, a total of 75% (9/12) were sequencing artefacts while two (D172G in *IL32* and S301C in *KATNAL2*) were found in control individuals and therefore excluded as possibly disease-causing. The remaining G23E variant in *TIMM23* was found to be real and therefore selected for further analysis. *TIMM23* was considered to be a good candidate as the protein encoded by this gene forms part of a complex that is located in the inner mitochondrial membrane. This complex is significant as it is responsible for the mediation of transport of transit peptide-containing proteins across the mitochondrial membrane (Zhang et al. 2012).

Genotyping was conducted on 1148 individuals (458 PD probands and 690 ethnically matched control individuals) using the TaqMan® SNP genotyping assay. No genotype calls could be made and Sanger sequencing was then performed on 10 randomly selected control individuals. The G23E variant was present in all of these individuals and it was determined that there was an error in the sequence that was obtained from dbSNP – a G nucleotide was missing from the sequence that had been used for the design of the TaqMan® probes. The errors in dbSNP are not uncommon – in a study which investigated the dbSNP Build 129 for contamination with what was termed as "Single Nucleotide Differences" or SNDs as a parallel to SNPs, it was determined that the frequency of the SNDs was 8.32%. Although the dbSNP build used for this project was 138, it remains reasonable to assume that not all of the

errors have necessarily been corrected and that new errors may still be introduced into the database (Kitts et al. 2014).

4.3.1 Analysis of P1150S in *EFCAB6* identified in family ZA92

The lack of success in the identification of a single variant that could be attributed to PD in the three probands led to the focus on only the affected siblings of family ZA92. Data analysis resulted in the identification of P1150S in *EFCAB6* as a plausible disease-causing variant. The proline to serine amino acid change at position 1150 was found to significantly change the properties of the expressed protein and may therefore affect protein folding and possibly function. EFCAB6 binds to a known PD causing gene, namely DJ-1 as well as to an androgen receptor to form a ternary complex in the cells (Niki et al. 2003). This binding protein subsequently recruits histone-deacetylase complexes so as to repress transcription activity of an androgen receptor (Niki et al. 2003).

Sanger sequencing confirmed that the P1150S variant was found in a heterozygous state in both affected siblings, but was absent from the unaffected sibling. However, this variant was not identified in the unaffected sibling who is currently 69 years of age. Subsequent genotyping using HRM did not identify the variant in any of the 184 ethnically matched controls that were examined and following this, the 458 South African PD probands were then genotyped. No additional probands were found to carry this variant and the frequency of the P1150S variant was concluded to be 0.22% (1/458) in the PD patients.

The available family members of the proband and the affected sibling of ZA92 were screened and it was determined that the P1150S variant was found in both children of the proband (aged 45 and 47 years of age respectively) and in both children of the affected sibling (aged 48 and 49 years of age respectively). However, this variant was not identified in the unaffected sibling who is currently 69 years of age. Genetic material for both of the parents of the affected siblings was not available as these individuals were deceased and it was therefore not possible to investigate whether or not this variant co-segregates with the disease in the family. However, it was determined that this variant does not co-segregate with disease in five Canadian families (Prof. Matthew Farrer; personal communication) and for this reason; the P1150S variant was excluded as a possible disease-causing mutation in the South African cohort.

4.3.2 Analysis of WES using a hypothesis-free approach

WES has provided a means for researchers to gain access to a highly enriched subset of the human genome in which to search for variants and possibly provide insights into a specific disease. Following our own shortcomings in the identification of a novel variant that could be associated with PD, we developed a novel toolkit for the filtration of WES data, namely TAPER™. TAPER™ is significant as it is considered to be a hypothesis-free approach to data analysis. By this it is meant that no extensive phenotypic information about the disease of interest is necessary and factors such as inheritance patterns or knowledge of disease pathways are not required for variant prioritization. This is a unique approach to data analysis and is of particular relevance in resource-constrained research environments such as those in South Africa because in many cases, detailed information on inheritance patterns or family history of the disorder is not available thereby making some of the conventional analytical approaches difficult to apply.

The use of TAPER™ identified a total of 20 variants in 20 genes that met all of the prioritisation criteria. Following prior bioinformatics analysis, this is the shortest list that had been obtained for this project. Variants that were found in both affected siblings and at least one other proband that had undergone WES were prioritised for further analysis. Of these, six genes and six variants were selected based on the fact that they had previously been associated with other movement disorders or had been directly implicated in pathways that had previously been associated with PD and related movement disorders.

Of the six prioritised genes and variants, three were excluded from further analysis due to their frequency in ethnically matched control individuals – the frequency cut-off for variants was 0.50% of the total number of control individuals genotyped. This was determined by halving the frequency cut-off previously established for the probands (namely 1.0%). Given the uniqueness of the South African population, it is expected that a pathogenic variant will be at a much lower frequency in control individuals than in the affected patients. The S719G variant in *WNK1* was found in 1.30% (9/690) of control individuals; the E43D variant in *CASP7* was found in 0.58% (4/690) of control individuals while the V626M variant in *MIPEP* was found in 0.72% (5/690) of control individuals. Further genotyping in the patients was therefore not conducted on any of these variants. The M534T variant in *SYNJ2* was neither sequenced nor genotyped as the variant does not co-segregate with the disease in

the Canadian population. The remaining two variants, V1405I in *SYNJ* and C357S in *USP17* were found to be plausible PD-causing variants.

4.3.2.1 V1405I in *SYNJ1*

V1405I was only found to be a real variant in the affected sibling pair and was identified in a homozygous state. A total of 690 control individuals were genotyped and this variant was not identified in any controls. Moreover, this variant was absent from the additional 457 PD probands screened.

Synaptojanin (SYNJ1) is a 145kDa protein that is located on chromosome 21q22 and interacts with growth factor receptor-bound protein 2 (Grb2) as well as a phosphoprotein that is involved in synaptic vesicle recycling and endocytosis (Krebs et al. 2013). There are four known isoforms of SYNJ1; two isoforms of 170kDa (isoform A: NP_003886.3, 1612 amino acids) and 145kDa (isoform B: NP_982271.2, 1350 amino acids) have been extensively studied. Both isoforms are generated from two open reading frames (ORFs) that are separated by an in-frame TAA stop codon (McPherson et al. 1996). Interestingly, both isoforms A and B are ubiquitously expressed but the 145kDa isoform is expressed at significantly higher levels in the brain where it is localized on coated endocytic intermediates in the nerve terminals (Ramjaun and McPherson 1996; McPherson et al. 1996). Both isoforms harbour numerous functional domains: a C-terminal proline-rich domain (PRD), a 5′-phosphatase domain in the centre and a suppressor of actin1 Sac1-like domain on the N-terminal. The longer 170kDa isoform carries an extra PRD translated from the second ORF (Figure 4.1). There are two additional SYNJ1 isoforms listed in RefSeq (isoform C: NP_001153774.1, 1295 amino acids and isoform D: NP_001153778.1, 1526 amino acids) that are of unknown functional relevance. Isoforms C and D have significantly shorter N-terminus and a distinct C-terminus. Although these isoforms are much shorter than isoform A, the functional domains of isoform C and D are the same as in isoforms A and B (Krebs et al. 2013; Drouet and Lesage 2014).

Mutations in *SYNJ1* have previously been associated with PD. A homozygous mutation, Arg258Gln (R258Q) has previously been identified by two independent research teams in two consanguineous families, one from Sicily in Italy and one from Iran (Krebs et al. 2013; Quadri et al. 2013). In both cases, homozygosity mapping coupled to WES were used to identify this variant in these individuals. The R258Q mutation is found in exon 5, and is

found in the Sac1 domain of the protein (Figure 4.1). This mutation is predicted to be damaging across multiple programs and the arginine at position 258 has been shown to be conserved in thirteen SYNJ1 orthologs and five Sac1-like domains containing proteins (Krebs et al. 2013; Quadri et al. 2013). Moreover, this mutation damages the Sac1 phosphatase activity targeting phosphatidylinositol monophosphate, suggesting that impaired synaptic vesicle recycling could be involved in PD pathology (Krebs et al. 2013; Quadri et al. 2013).

Mutations in this gene are extremely rare; to date there are only six early-onset (AAO younger than 30 years of age) PD patients (from three families with two affected siblings respectively) who carry the homozygous R258Q mutation. Screening of the parents of the affected sibling pairs shows that are all parents are heterozygous for this variant while unaffected siblings are homozygous carriers for the wild type allele or heterozygous mutation carriers (Krebs et al. 2013; Quadri et al. 2013; Olgiati et al. 2014).

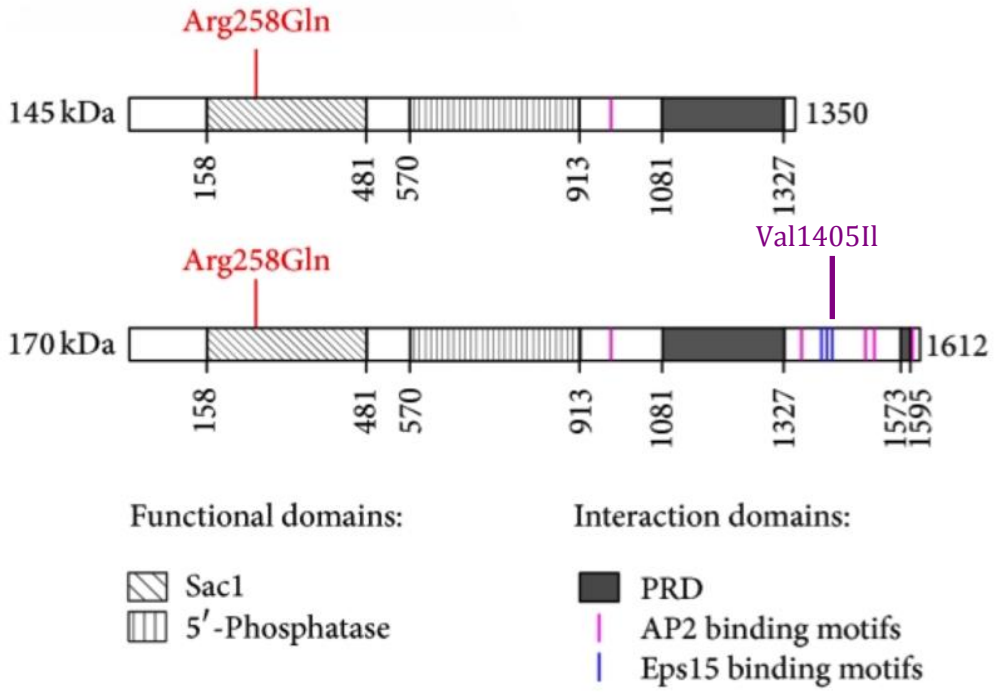


Figure 4.1 Functional and interaction domains of isoforms A and B of SYNJ1. The 145kDa (top) and 170kDa (bottom) isoforms contain an N-terminal Sac1 domain, a central 5′-phosphatase domain and two functional inositol phosphatase domains. Numerous protein-protein interaction domains are found in the C-terminal region: one or more PRD domains, AP2 binding motifs (WxxF, FxDxF, and DxF, indicated in pink), and Eps15 binding motifs (NPF: asparagine-proline-phenylalanine, indicated in blue). The homozygous mutation Arg258Gln, found in Parkinson's disease patients, is indicated in red and the V1405I variant identified in the present study is shown in purple. Numbers indicate the amino acid positions along the proteins. Sac1 - suppressor of actin1; PRD - proline-rich domain; AP2 - adaptor protein complex 2; Eps15 - epidermal growth factor receptor pathway substrate 15 (Taken from Drouet and Lesage 2014).

In the present study, the V1405I variant identified in the affected sibling pair was found to be absent from their unaffected sibling. Moreover, this variant was not identified in any of the other PD patients or controls screened. Although this variant is not a novel variant (rs79652470), it has not been identified in any population group in a homozygous form. In addition to this, the global frequency of the variant in a heterozygous state is extremely low, at 0.01104%. This frequency is as low as some of the well-established PD-causing mutations such as G2019S in *LRRK2*, R1441C in *LRRK2*, R275W in *Parkin* and G430D in *Parkin* (Table 4.1). To date, this variant has not been found in a homozygous form in any of the PD patients available to us from the laboratories of our collaborators (Prof. Matthew Farrer, personal communication; Dr. Suzanne Lesage, personal communication).

The R258Q mutation is, to date, the only mutation in *SYNJ1* that has been associated with PD. It is plausible that the V1405I (Valine to Isoleucine) homozygous variant could be attributed to the disease. Following *in silico* analysis, it was determined that the amino acid substitution is a conservative one; both amino acids are similar in size and are hydrophobic. However, the effect of the amino acid substitution could not be determined because of the fact that there is no suitable crystal structure available for analysis. The V1405I variant is predicted to be deleterious by SIFT, damaging by PolyPhen2, deleterious by FATHMM and the CADD score is 29.40 (this figure means that the variant is amongst the top 1% of deleterious variants). Moreover, GERP+++ and PhyloP predict the variant to be in a highly conserved region, with both programs predicting high conservation scores, thus indicating that the variant may be damaging potentially affecting protein structure and function.

The functions of SYNJ1 in actin dynamics and synaptic vesicle recycling in both pre- and postsynaptic compartments are of relevance to aid in the understanding of the physiopathology of PD (Drouet and Lesage 2014). Research into synaptic vesicle trafficking pathways has provided strong evidence that these pathways may be implicated in PD mechanisms. Most of the proteins that have been implicated in autosomal dominant PD, as well as those responsible for autosomal recessive forms of Parkinsonism, have been implicated, directly or indirectly, in synaptic vesicle turnover (Figure 4.2). SYNJ1 is a phosphoinositide phosphatase protein that is required for proper synaptic activity. The identification of the homozygous V1405I in an affected sibling pair and its absence in both the ethnically matched controls as well as the low frequency in the global population according to the ExAC database is an indication that this variant may be pathogenic and

should be investigated further.

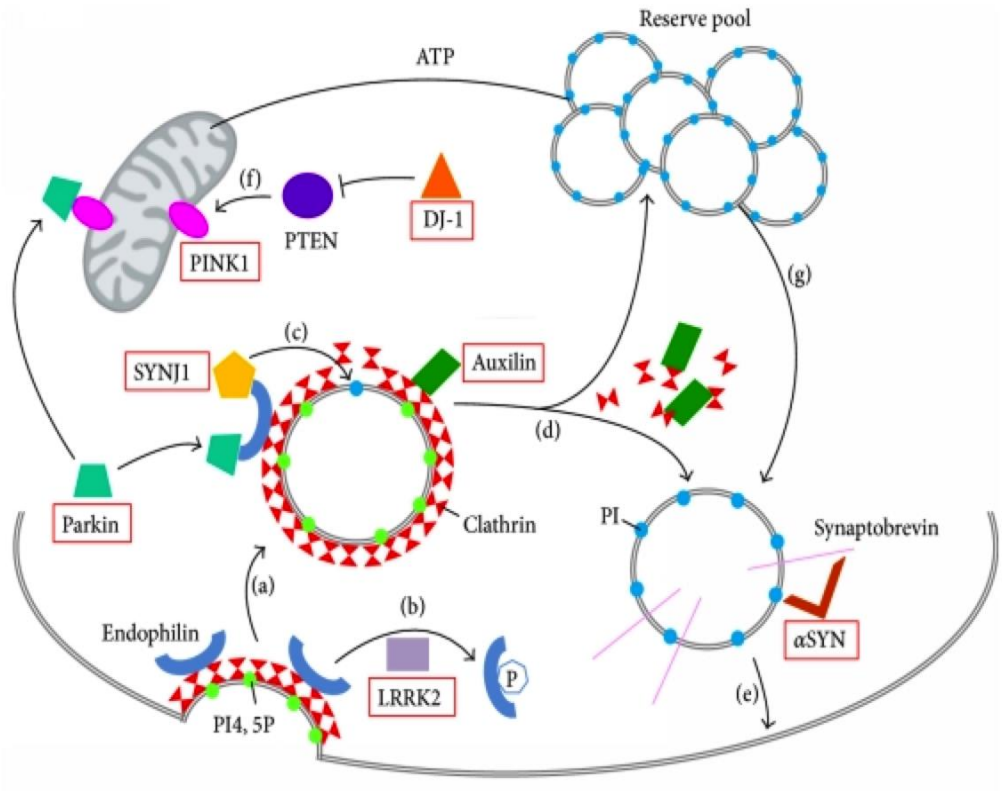


Figure 4.2 Synaptic recycling and PD genes. Diagrammatic representation of a presynaptic terminal showing the PD genes (shown in red boxes) and their respective role in synaptic vesicle recycling. (a) During endocytosis, invagination of the clathrin-coated membrane requires endophilin. Endophilin harbours numerous SH3 domains that interact with the SYNJ1 PRD domain and/or Parkin. (b) LRRK2 is responsible for the phosphorylation of endophilin, which leads to the dissociation of the latter from clathin-coated vesicles. (c) SYNJ1 is recruited to the coated vesicles through endophilin and will dephosphorylate PI4,5P into PI, thereby shedding clathrin and its adaptor from the bilayer. (d) Uncoating of the vesicles also requires auxilin intervention and subsequent chaperoning of the clathrin molecules. The postendocytic vesicles are then able to return to the reserve pool where they undergo clustering, or return directly to the release site and begin an exocytosis step. (e) Synaptic vesicles are docked and then fused to the membrane by means of a multi-protein complex that includes synaptobrevin and αSYN. (f) PTEN is a lipid phosphatase that is inhibited by DJ-1, and can increase levels of the mitochondrial PINK1 protein. This pathway is involved in NMDA receptor signalling. (g) Proper mitochondrial functioning leads to ATP synthesis, necessary to mobilize the reserve pool of vesicles during synapse stimulation. Abbreviations: PI4,5P - phosphatidylinositol 4,5-bisphosphates; PI - phosphatidylinositol; ATP - adenosine triphosphate; SYNJ1 - synaptojanin 1; LRRK2 - leucine-rich repeat serine/threonine-protein kinase 2; PTEN - phosphatase and tensin homologue; PINK1 - PTEN induced putative kinase 1; DJ-1 - Parkinson's disease protein 7; αSYN - alpha-synuclein; NMDR - N-methyl-D-aspartate receptor (Taken from Drouet and Lesage 2014).

Table 4.1 Global population frequencies of V1405I in *SYNJ1* and C357S *USP17* as compared to other PD causing genes.

	Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
***SYNJ1* (V1405I)**	African	2	9 832	0	0.0002034
	European (Non-Finnish)	11	64 422	0	0.0001707
	East Asian	0	8 466	0	0
	European (Finnish)	0	6 478	0	0
	Latino	0	11 386	0	0
	Other	0	888	0	0
	South Asian	0	16 280	0	0
	TOTAL	**13**	**117 752**	**0**	**0.0001104**
***USP17* (C357S)**	African	0	5 744	0	0
	European (Non-Finnish)	1	2 980	0	0.0003356
	East Asian	0	1 022	0	0
	European (Finnish)	0	82	0	0
	Latino	0	844	0	0
	Other	0	56	0	0
	South Asian	0	1990	0	0
	TOTAL	**1**	**12 718**	**0**	**7.863e-05**
***LRRK2* (G2019S)**	African	3	10 396	0	0.0002886
	European (Non-Finnish)	42	66 730	0	0.0006294
	East Asian	0	8 654	0	0
	European (Finnish)	0	6 641	0	0
	Latino	2	11 536	0	0.0001734
	Other	0	908	0	0
	South Asian	0	16 512	0	0
	TOTAL	**47**	**121 350**	**0**	**0.0003873**
***LRRK2* (R1441C)**	African	0	10 389	0	0
	European (Non-Finnish)	1	66 692	0	1.499e-05
	East Asian	0	8 652	0	0

	European (Finnish)	0	6 608	0	0
	Latino	0	11 526	0	0
	Other	0	906	0	0
	South Asian	0	16 512	0	0
	TOTAL	1	121 294	0	8.244e-06
***Parkin* (G430D)**	African	1	8 598	0	0.0001772
	European (Non-Finnish)	10	56 440	0	0.0001163
	East Asian	0	7 228	0	0
	European (Finnish)	0	5 052	0	0
	Latino	0	9 674	0	0
	Other	0	760	0	0
	South Asian	0	14 510	0	0
	TOTAL	11	102 262	0	0.0001076
***Parkin* (R275W)**	African	3	10 206	0	0.0002939
	European (Non-Finnish)	204	65 874	0	0.003097
	East Asian	0	8 564	0	0
	European (Finnish)	13	6 558	0	0.001982
	Latino	15	11 324	0	0.001325
	Other	2	884	0	0.00262
	South Asian	9	16 046	0	0.005609
	TOTAL	246	119 456	0	0.002059

Taken from ExAC Genome Browser (http://exac.broadinstitute.org). Date accessed: 28 August 2015.

4.3.2.2 C357S in *USP17*

The C357S variant in *USP17* was identified in homozygous form in all four individuals who had undergone WES. Subsequent genotyping revealed that the variant was in a total of 18/458 probands of which twelve were Afrikaner, four were Mixed Ancestry and two were White (Table 3.14, page 99). All probands found to carry this variation were homozygous. Interestingly, of the twelve Afrikaner probands that were found to carry this variant, 7/12 (58.3%) form part of the large, 40 family pedigree. The same variant, was however, identified in a single Afrikaner control individual. This control is 80 years of age and was sourced from the Western Province Blood Transfusion Services (WPBTS) in Cape Town.

This individual has been de-identified and it is not possible to assess whether or not this individual may have developed PD subsequent to his participation in the study.

Although the C357S variant is not a novel variant, and is present in the ExAC database, it has not been recorded in either dbSNP or Ensembl. Moreover, it has not been identified in any population group in a homozygous form (Table 4.1). In addition to this, the global frequency of the variant in a heterozygous state is extremely low, at 0.007863%. This frequency is as low as some of the well-established PD-causing mutations such as G2019S in *LRRK2*, R1441C in *LRRK2*, R275W in *Parkin* and G430D in *Parkin* (Table 4.1). To date, this variant has not been found in a homozygous form in any of the PD patients available to us from the laboratories of our collaborators (Prof. Matthew Farrer, personal communication; Dr. Suzanne Lesage, personal communication).

The effect of the amino acid change (Cysteine to Serine) could not be determined due to the fact that there is no crystal structure available for analysis. However, the C357S variant is predicted to be deleterious by SIFT, damaging by PolyPhen2, deleterious by FATHMM and the CADD score is 34.89 (this figure means that the variant is amongst the top 0.1% of deleterious variants). Moreover, GERP+++ and PhyloP predict the variant to be in a highly conserved region, with both programs predicting high conservation scores, thus indicating that the variant may be detrimental should it be present in an individual.

Ubiquitin Specific-processing Protease 17 (USP17) is a 59.6kDa protein that is located on chromosome 8p23.1. USP17 interacts with SET nuclear proto-oncogene, which inhibits the acetylation of nucleosomes by histone acetylases (Cunha et al. 2014) and CBX1 (chromobox homolog 1), which is a highly conserved nonhistone protein (Bian et al. 2014). USP17 is a deubiquitinating enzyme that removes conjugated ubiquitin from specific proteins so as to regulate multiple cellular processes.

USP17 is an immediate early gene that belongs to a subfamily of cytokine inducible deubiquitinating enzymes (DUBs). DUBs are important as they are involved in the removal of ubiquitin from post-translationally modified proteins which is important for numerous cellular functions such as transcription, cell cycle progression, DNA repair and apoptosis (de la Vega et al. 2011). USP17 is induced by interleukin (IL)-4 and IL-6, which regulate the growth and differentiation of leukocytes. However, more recently, it was demonstrated that

USP17 controls the functioning of the small GTPase Ras through posttranslational processing and membrane localization (Burrows et al. 2004; Burrows et al. 2009). Mutations in this gene do not appear to be common but very little work has been done on the gene itself. There is a known variant, C89S that abolishes both enzymatic activity and the effects on cell proliferation (Burrows et al. 2004). The major domain of this protein is the USP domain, which spans the amino acids 80 – 375. The novel variant that was identified in the South African Afrikaner is found in this domain.

The post-translational modification of proteins through the covalent attachment of ubiquitin targets these proteins for degradation by the proteasome. Ubiquitination and deubiquitination are therefore thought to work in combination with each other and the selectivity of proteolysis will be determined by the combination of ubiquitination enzymes and DUBs that are present at a specific time point (de la Vega et al. 2011). Moreover, DUBs play numerous roles in the UPS. Mutations in genes that code for these DUBs could therefore significantly alter the functioning and processing of proteins within a cell and pathway. DUBs are involved in the activation of so-called ubiquitin pro-proteins and this is done co-translationally (Reyes-Turcu, Ventii, and Wilkinson 2009). Ubiquitin is not expressed independently, but is rather, expressed as linear polyubiquitination that consists of multiple mono-ubiquitin that must undergo processing to yield a mature ubiquitin monomer or as a pro-protein that is fused to other ribosomal proteins (Reyes-Turcu, Ventii, and Wilkinson 2009; Dikic and Bremm 2014). DUBs also recycle ubiquitin and importantly, DUBs reverse the ubiquitination or ubiquitin-like modification of specific target proteins (Nijman et al. 2005; Kumar et al. 2015). This role of DUBs is of particular significance as it antagonises the ubiquitination of proteins thereby playing a role that is similar to that of the phosphatases in a phosphatase/kinase pathway (Reyes-Turcu, Ventii, and Wilkinson 2009; Dikic and Bremm 2014). Lastly, DUBs are responsible for the regeneration of monoubiquitin from unanchored polyubiquitin. Given the significant role that DUBs such as USP17 play in the UPS, disruptions in this pathway due to mutations in genes coding for DUBs may contribute to the pathobiology of PD.

4.4 Limitations of the study

The most successful studies that have identified a novel disease-causing gene using WES have mainly relied on discrete filtration of data and in most cases, this has been coupled to

linkage data (Trinh and Farrer 2013). This is the first WES project to be conducted on South African PD patients and for this reason, there were numerous limitations. Limitations of employing WES as a method for novel mutation detection include the fact that a significant portion of the human genome is not examined (98.8%). Moreover, structural variations such as CNVs are difficult to detect using this method. Additional factors such as analytical and technical limitations that could contribute to difficulties in novel variant identification were also identified.

Technical limitations could account for the lack of a specific variant being identified in the South African Afrikaner. One such limitation could be that the causal variant was covered but was not accurately called. This was illustrated by Dewey et al. where challenges to the interpretation of NGS data such as assembly and variant calling against the human reference genome were highlighted (Dewey et al. 2011). Variants that were identified as heterozygous by the variant calling software were identifying the major allele as the minor allele; essentially an allele swop. This could lead to a large number of false positive calls, the identification of numerous sequencing artefacts or the putative variant being missed. Another technical limitation of WES is that part or all of the gene of interest may not be in the target definition of a specific exon - WES technologies are continually being improved and probes in specific sequence capture methods are designed based on the current sequence information that is available from databases such as consensus coding sequence (CCDS) database and Refseq database. It is for this reason then, that unknown exons or yet to be annotated exons cannot be captured. In addition to this, the role of variations in the non-coding regions of the genome in Mendelian disease has yet to be determined. One such example is the intronic hexanucleotide repeat in C9orf72, associated with ALS and frontotemporal dementia (McMillan et al. 2014). This variant was missed using WES alone as the variant did not form part of the target exome definition and was only identified through Sanger sequencing.

One of the most prominent analytical limitations in developing countries that are resource constrained pose a significant challenge when employing this technology to specific diseases in these individuals. A limited number of bioinformaticists and the lack of adequate computational infrastructure further limit the successful application and implementation of NGS technologies such as WES to various diseases. The scarcity of trained bioinformaticists means that laboratory scientists with limited bioinformatics knowledge are left with the daunting task of prioritising candidate disease-causing variants. Another analytical limitation

of the present study is the fact that there are no trios available for WES analysis. The parents of the probands were not available for WES as a means to exclude or include specific variants as plausible disease-causing candidates. Lack of trios and additional first and second degree family members meant that comprehensive co-segregation studies could not be conducted. Rare recessive variants as well as cases where two mutations come from the same parent compound the analytical limitations of WES. The lack of a universal analysis pipeline is an additional facet that must be taken into account. Depending on the filtration criteria that are used, a pathogenic variant may be present in an individual but may not be identified as this is dependent on the filtration parameters that are employed. Finally, a significant analytical limitation is the fact that there are no population specific databases – pathogenic mutations may be successfully identified in the affected individuals through an appropriate filtration method, but these mutations may be present at low frequencies in the background population. There are currently more than 17 million SNPs that have been identified in the human genome but there is an error rate of approximately 15-17% (Day 2010; Kitts et al. 2014). Using an appropriate MAF for the variants and the use of additional databases such as 1KGP and ESP6500 will help to reduce this type of error. Over and above the population specific variants, there is no way to account for pseudogenes or low penetrance alleles (variants may be excluded as potential candidates based on the fact that they are found in healthy control individuals who never develop the disease). Moreover, phenocopies are another significant problem and may be present at a relatively high frequency of 18% in PD patients (Prof Matt Farrer, personal communication), thereby further compounding the problems with data filtration pipelines. Finally, the whole genome SNP array did not identify runs of homozygosity (ROH) and this could be attributed to the small number of SNPs that were genotyped. Expanding the number of SNPs (SNP density) may have identified ROH and provided us with a better means to identify regions of interest in the four samples subjected to WES. Improvements to NGS technologies such as WES as well as an improved understanding of variation of the human exome in diverse populations may allow some of these limitations to be overcome in the near future.

4.5 Future work

A relatively small number of plausible disease-causing variants were identified during the course of the current study. Two of these, namely V1405I in *SYNJ1* and C357S in *USP17* were identified as potential disease-causing candidates and should be analysed further. Given

the low frequency of the V1405I variant in *SYNJ1*, it is possible that this is a family specific rare polymorphism. However, in future studies the frequency of this variant should be assessed in the approximately 41 000 PD patients and 41 000 controls potentially available from the GEOPD consortium of which we are a member (http://www.geopd.org/about/). The same should be conducted for the C357S variant as it is found in 58.3% of the Afrikaner probands that belong to the large 40 family pedigree. In the case of *USP17*, the effect of the C357S variant on deubiquitination and protein aggregation could be investigatedthrough the use of high performance liquid chromatography (HPLC) and fluorescence assays using ubiquitin-amidomethylcoumarin assays (Russel and Wilkinson 2005). HPLC can be used to monitor the enzymatic activity of DUBs while fluorescence assays such as Ub-amidomethylcoumarin (AMC) assays have been used to identify substrate specificity. The quantification of the rate of release of the fluorescent tag from a substrate allows for the calculation of the amount of DUB enzymatic activity. In the case of the V1405I variant, the protein product SYNJ1 is a phosphoinositide phosphatase protein that is required for proper synaptic activity. The *in vitro* functional effect of the variant could be analysed using phosphoinositide phosphatase assay to assess the effect of the variant on the synaptic activity. This is a colorimetric method for the determination of inorganic phosphate that utilizes malachite green, and is used for the quantification of protein phosphatase activity (Mavrantoni et al. 2015). The assay is based on the change of absorbance at 620mn due to the formation of malachite green complexes and is able to detect phosphate release from protein phosphate substrates.

Future work on the South African PD cohort should focus on the identification of families for which multiple affected individuals are available for study. Also, further analysis using WES should involve the inclusion of parents of affected individuals where possible. Most of the successes of WES have been achieved on rare Mendelian disorders rather than complex diseases such as PD. The variants that have been identified using WES are typically high penetrance alleles that co-segregate with the disease. It is hypothesised that WES coupled to linkage mapping strategies has the potential to provide important insights into complex diseases such as PD (Gustavsson et al. 2015). The challenges experienced with population specific variants are highlighted throughout the course of the current study. The challenge going forwards is the need to expand control or background population databases to include data from various South African ethnic groups so as to aid novel gene discovery.

As the cost of NGS technologies continue to decrease, it is likely that whole genome sequencing will take precedence over WES. For this approach, it should be noted that taking advantage of the more complete and comprehensive dataset for putative disease-causing variants in patients is reliant on the development of more universal analytical strategies, especially those that make use of non-coding variation. It is therefore necessary for more detailed phenotypic curation as well as the need for improved statistical, technical and bioinformatics strategies to aid in the reduction of false positive and negative variant calls, for the selection and prioritization of indels and candidate variants as well as for the prediction and annotation of potential functional impact of the variant(s) of interest. The identification of candidate genes for complex diseases such as PD is a realistic goal, however narrowing down candidate gene lists is likely to require unprecedented collaborative efforts in the field of neurogenetics. This should involve the development of large consortia groups for data sharing and with streamlined and high-throughput approaches to conduct candidate prioritization and screening in replication cohorts of well-characterized patients.

4.6 Concluding remarks

Neurodegenerative diseases such as PD present a significant health burden and affect the quality of life of both patients as well as their caregivers. However, this is not isolated to individuals alone, but also has a significant impact on the wider society and on the economy (WHO 2014). As the life expectancy of global populations increases it is hypothesised that this burden will increase even further (Collins et al. 2011, WHO 2014). Moreover, the chronic nature of the diseases, especially those who present with juvenile or early onset PD, adds to the challenges faced by the global healthcare system. Over the last 20 years, several risk factors and a large number of genes have been identified and this knowledge has helped ease the burden on both global disease and affected patients (Singleton 2015). It remains important to pinpoint additional genes that may in some way be associated with this disease so as to gain key biological insights that underlie this debilitating disorder (Novarino et al. 2014). As more PD genes are identified, an understanding into the relationships between genes, their protein products and the pathways in which they are found, is generating significant insight into the pathobiological network of processes and interactions that may lead to disease onset (MacLeod et al. 2013; Beilina et al. 2014). Knowledge gained through the identification of variants in genes that may be associated with PD may show coordinated

disease networks as well as common risk alleles and rare mutations that may be found in the same biological pathway (Singleton 2015).

The advent of NGS technologies such as WES has enabled researchers to analyse biological systems and to make biological discoveries at a level that was never before possible. As the sequencing technologies have improved, data analysis technologies have evolved but it was determined that none of those that were implemented were successful in the identification of a plausible novel PD causing variant in the South African cohort. For this reason, a novel filtration method, namely TAPER™ was designed to facilitate novel gene discovery in complex disorders using a hypothesis-free approach in resource-limited environments. Through the use of TAPER™, we successfully identified and validated two potentially disease-causing variants in the South African PD cohort namely V1405I in *SYNJ1I* and C357S in *USP17*. However, the results from this study show that although the genealogical and whole genome SNP array data supported the possibility of a founder effect for PD in the South African Afrikaner, the WES results did not. It was determined that the Afrikaner probands linked to a common founder couple are related but are unlikely to have PD for the same genetic reason and therefore from our work, it is concluded that a founder effect for the disease in these patients does notappear to be present.

There are numerous genes that have been linked to PD but the pathogenic processes that are driven by these genes remain poorly understood. However, it is plausible that the identification of new PD genes have the potential to provide key insights into the underlying pathogenic processes that predispose an individual to develop PD. The identification of novel disease-causing variants is likely to increase and improve the insights into the pathobiology of this disease in South African patients. In conclusion, the present study represents an important first step in the application of WES and bioinformatics to the identification of new PD-causing genes in South African patients. Future studies will build on the lessons learnt and will be more focussed and better designed to ensure a higher chance of success.

References

Abbas, N., C. B. Luecking, S. Ricard., et al. 1999. "A Wide Variety of Mutations in the Parkin Gene Are Responsible for Autosomal Recessive Parkinsonism in Europe." *Human Molecular Genetics* 8 (4): 567–74.

Anand, V. S., and S. P. Braithwaite. 2009. "LRRK2 in Parkinson's Disease: Biochemical Functions." *FEBS Journal* 276 (22): 6428–35.

Annesi, G., G. Savettieri, P. Pugliese., et al. 2005. "DJ-1 Mutations and Parkinsonism-Dementia-Amyotrophic Lateral Sclerosis Complex." *Annals of Neurology* 58 (5): 803–7.

Bamshad, M. J., S. B. Ng, A. W. Bigham., et al. 2011. "Exome Sequencing as a Tool for Mendelian Disease Gene Discovery." *Nature Reviews Genetics* 12 (11): 745–55.

Bardien, S., R. Keyser, Y. Yako., et al. 2009. "Molecular Analysis of the Parkin Gene in South African Patients Diagnosed with Parkinson's Disease." *Parkinsonism & Related Disorders* 15 (2): 116–21.

Beasley, S. A., V.A. Hristova, and G. S. Shaw. 2007. "Structure of the Parkin In-between-Ring Domain Provides Insights for E3-Ligase Dysfunction in Autosomal Recessive Parkinson's Disease." *Proceedings of the National Academy of Sciences of the United States of America* 104 (9): 3095–3100.

Beilina, A., I. N. Rudenko, A. Kaganovich., et al. 2014. "Unbiased Screen for Interactors of Leucine-Rich Repeat Kinase 2 Supports a Common Pathway for Sporadic and Familial Parkinson Disease." *Proceedings of the National Academy of Sciences of the United States of America* 111 (7): 2626–31.

Benkert, P, M. Biasini, and T. Schwede. 2011. "Toward the Estimation of the Absolute Quality of Individual Protein Structure Models." *Bioinformatics* 27 (3): 343–50.

Bezard, E., and P-O. Fernagut. 2014. "Premotor Parkinsonism Models." *Parkinsonism & Related Disorders* 20, Supplement 1 (January): S17–19.

Bian, Y., C. Song, K. Cheng., et al. 2014. "An Enzyme Assisted RP-RPLC Approach for in-Depth Analysis of Human Liver Phosphoproteome." *Journal of Proteomics* 96 (January): 253–62.

Bisaglia, M., I. Tessari, S. Mammi., et al. 2009. "Interaction between Alpha-Synuclein and Metal Ions, Still Looking for a Role in the Pathogenesis of Parkinson's Disease." *Neuromolecular Medicine* 11 (4): 239–51.

Blackinton, J., M. Lakshminarasimhan, K. J. Thomas., et al. 2009. "Formation of a Stabilized Cysteine Sulfinic Acid Is Critical for the Mitochondrial Function of the Parkinsonism Protein DJ-1." *Journal of Biological Chemistry* 284 (10): 6476–85.

Blanckenberg, J., S. Bardien, B. Glanzmann., et al. 2013. "The Prevalence and Genetics of Parkinson's Disease in Sub-Saharan Africans." *Journal of the Neurological Sciences* 335 (1–2): 22–25.

Blanckenberg, J., C. Ntsapi, J. A. Carr., et al. 2014. "EIF4G1 R1205H and VPS35 D620N Mutations Are Rare in Parkinson's Disease from South Africa." *Neurobiology of Aging* 35 (2): 445.e1–445.e3.

Bonifati, V. Rizzu P, Squitieri F., et al. 2003. "DJ-1 (PARK7), a Novel Gene for Autosomal Recessive, Early Onset Parkinsonism." *Neurological Sciences* 24 (3): 159–60.

Bonifati, V. 2014. "Genetics of Parkinson's Disease – State of the Art, 2013." *Parkinsonism & Related Disorders* 20, Supplement 1 (January): S23–28.

Botha, M. C., and P. Beighton. 1983a. "Inherited Disorders in the Afrikaner Population of Southern Africa. Part I. Historical and Demographic Background, Cardiovascular, Neurological, Metabolic and Intestinal Conditions." *South African Medical Journal* 64 (16): 609–12.

Botha, M. C., and P. Beighton 1983b. "Inherited Disorders in the Afrikaner Population of Southern Africa. Part II. Skeletal, Dermal and Haematological Conditions; the Afrikaners of Gamkaskloof; Demographic Considerations." *South African Medical Journal* 64 (17): 664–67.

Botstein, D., and N. Risch. 2003. "Discovering Genotypes Underlying Human Phenotypes: Past Successes for Mendelian Disease, Future Approaches for Complex Disease." *Nature Genetics* 33 Suppl (March): 228–37.

Bras, J.M., and A.B Singleton. 2011. "Exome Sequencing in Parkinson's Disease." *Clinical Genetics* 80 (2): 104–9.

Bras, J., R. Guerreiro, and J. Hardy. 2012. "Use of next-Generation Sequencing and Other Whole-Genome Strategies to Dissect Neurological Disease." *Nature Reviews Neuroscience* 13 (7): 453–64.

Bras, J., A. Verloes, S. A. Schneider, S. E. Mole, and R. J. Guerreiro. 2012. "Mutation of the Parkinsonism Gene ATP13A2 Causes Neuronal Ceroid-Lipofuscinosis." *Human Molecular Genetics* 21 (12): 2646–50. doi:10.1093/hmg/dds089.

Brink, A. J., and M. Torrington. 1977. "Progressive Familial Heart Block--Two Types." *South African Medical Journal = Suid-Afrikaanse Tydskrif Vir Geneeskunde* 52 (2): 53–59.

Brink, P.A., L.T. Steyn, G.A. Coetzee., et al. 1987. "Familial Hypercholesterolemia in South African Afrikaners - Pvu II and Stu I DNA Polymorphisms in the LDL-Receptor Gene Consistent with a Predominating Founder Gene Effect." *Human Genetics* 77 (1): 32–35.

Brink, P. A., L. Crotti, V. Corfield., et al. 2005. "Phenotypic Variability and Unusual Clinical Severity of Congenital Long-QT Syndrome in a Founder Population." *Circulation* 112 (17): 2602–10.

Browning, B. L., and S. R. Browning. 2013. "Detecting Identity by Descent and Estimating Genotype Error Rates in Sequence Data." *American Journal of Human Genetics* 93 (5): 840–51.

Burrows, J. F., A. A. Kelvin, C. McFarlane., et al. 2009. "USP17 Regulates Ras Activation and Cell Proliferation by Blocking RCE1 Activity." *The Journal of Biological Chemistry* 284 (14): 9587–95.

Burrows, J. F., M. J. McGrattan, A. Rascle., et al. 2004. "DUB-3, a Cytokine-Inducible Deubiquitinating Enzyme That Blocks Proliferation." *The Journal of Biological Chemistry* 279 (14): 13993–0.

Canet-Avilés, R M., M. A. Wilson, D. W. Miller., et al. 2004. "The Parkinson's Disease Protein DJ-1 Is Neuroprotective due to Cysteine-Sulfinic Acid-Driven Mitochondrial Localization." *Proceedings of the National Academy of Sciences of the United States of America* 101 (24): 9103–8.

Casals, J., T. S. Elizan, and M. D. Yahr. 1998. "Postencephalitic Parkinsonism – a Review." *Journal of Neural Transmission* 105 (6-7): 645–76.

Cavalli-Sforza, L. L., and M. W. Feldman. 2003. "The Application of Molecular Genetic Approaches to the Study of Human Evolution." *Nature Genetics* 33 (March): 266–75.

Caviness, J. N. 2014. "Pathophysiology of Parkinson's Disease Behavior – a View from the Network." *Parkinsonism & Related Disorders* 20, Supplement 1 (January): S39–43.

Cawley S. 2005. "Sanger's Method." http://stat-www.berkeley.edu/users/terry/Classes/s260.1998/Week8b/week8b/node9.html.

Chahrour, M. H., T. W. Yu, E. T. Lim., et al. 2012. "Whole-Exome Sequencing and Homozygosity Analysis Implicate Depolarization-Regulated Neuronal Genes in Autism." *PLoS Genet* 8 (4): e1002635.

Chartier-Harlin, M-C., J. C. Dachsel, C. Vilariño-Güell., et al. 2011. "Translation Initiator EIF4G1 Mutations in Familial Parkinson Disease." *The American Journal of Human Genetics* 89 (3): 398–406.

Chartier-Harlin, M-C., J. Kachergus, C. Roumier., et al. 2004. "A-Synuclein Locus Duplication as a Cause of Familial Parkinson's Disease." *The Lancet* 364 (9440): 1167–69.

Clark, M. J., R. Chen, H. Y. K. Lam., et.al. 2011. "Performance Comparison of Exome DNA Sequencing Technologies." *Nature Biotechnology* 29 (10): 908–14.

Collins, P. Y., V.Patel, S. S. Joestl., et al. 2011. "Grand Challenges in Global Mental Health." *Nature* 475 (7354): 27–30.

Consortium, The UniProt. 2015. "UniProt: A Hub for Protein Information." *Nucleic Acids Research* 43 (D1): D204–12.

Cookson, M. R. 2012. "Cellular Effects of LRRK2 Mutations." *Biochemical Society Transactions* 40 (5): 1070–73.

Cookson, M. R., and O. Bandmann. 2010. "Parkinson's Disease: Insights from Pathways." *Human Molecular Genetics* 19 (R1): R21–27.

Cooper, J. M., S. E. Daniel, C. D. Marsden., et al. 1995. "L-Dihydroxyphenylalanine and Complex I Deficiency in Parkinson's Disease Brain." *Movement Disorders: Official Journal of the Movement Disorder Society* 10 (3): 295–97.

Cuervo, A. M., E. Bergamini, U. T. Brunk., et al. 2005. "Autophagy and Aging: The Importance of Maintaining 'Clean' Cells." *Autophagy* 1 (3): 131–40.

Cunha, S., Y-C. Lin, E. A. Goossen., et al.. 2014. "The RON Receptor Tyrosine Kinase Promotes Metastasis by Triggering MBD4-Dependent DNA Methylation Reprogramming." *Cell Reports* 6 (1): 141–54.

Da Costa, C. A., C. Sunyach, E. Giaime., et al. 2009. "Transcriptional Repression of p53 by Parkin and Impairment by Mutations Associated with Autosomal Recessive Juvenile Parkinson's Disease." *Nature Cell Biology* 11 (11): 1370–75.

D'Antonio, M., P. D'Onorio De Meo, D. Paoletti., et al. 2013. "WEP: A High-Performance Analysis Pipeline for Whole-Exome Data." *BMC Bioinformatics* 14 (Suppl 7): S11.

Darios, F., O. Corti, C. B. Lücking., et al. 2003. "Parkin Prevents Mitochondrial Swelling and Cytochrome c Release in Mitochondria-Dependent Cell Death." *Human Molecular Genetics* 12 (5): 517–26.

Dauer, W., and S. Przedborski. 2003. "Parkinson's Disease: Mechanisms and Models." *Neuron* 39 (6): 889–909.

Dawson, T. M., and V. L. Dawson. 2003. "Molecular Pathways of Neurodegeneration in Parkinson's Disease." *Science* 302 (5646): 819–22.

Day, I. N.M. 2010. "dbSNP in the Detail and Copy Number Complexities." *Human Mutation* 31 (1): 2–4.

Defesche, J.C., D. E. Van Diermen, M.R.Hayden., et al. 1996. "Origin and Migration of an Afrikaner Founder Mutation FH(Afrikaner-2) (V408M) Causing Familial Hypercholesterolemia." *Gene Geography* 10 (1): 1–10.

Defesche, J. C., D. E. Van Diermen, P. J. Lansberg., et al. 1993. "South African Founder Mutations in the Low-Density Lipoprotein Receptor Gene Causing Familial Hypercholesterolemia in the Dutch Population." *Human Genetics* 92 (6): 567–70.

De Lau, L.M., and M. B. Breteler. 2006. "Epidemiology of Parkinson's Disease." *The Lancet Neurology* 5 (6): 525–35.

De la Vega, M., A. A. Kelvin, D. J. Dunican., et al. 2011. "The Deubiquitinating Enzyme USP17 Is Essential for GTPase Subcellular Localization and Cell Motility." *Nature Communications* 2 (March): 259.

De Wit, E., W. Delport, C. E. Rugamika., et al. 2010. "Genome-Wide Analysis of the Structure of the South African Coloured Population in the Western Cape." *Human Genetics* 128 (2): 145–53.

Devine, M. J., K. Gwinn, A. Singleton., et al. 2011. "Parkinson's Disease and A-Synuclein Expression." *Movement Disorders* 26 (12): 2160–68.

Dewey, F. E., R. Chen, S. P. Cordero., et al. 2011. "Phased Whole-Genome Genetic Risk in a Family Quartet Using a Major Allele Reference Sequence." *PLoS Genet* 7 (9): e1002280.

Dikic, I., and A. Bremm. 2014. "DUBs Counteract Parkin for Efficient Mitophagy." *The EMBO Journal* 33 (21): 2442–43.

Dinkel, H., S. Michael, R. J. Weatheritt., et al. 2012. "ELM—the Database of Eukaryotic Linear Motifs." *Nucleic Acids Research* 40 (D1): D242–51.

Dotchin., C., and R. Walker . 2012. "The Management of Parkinson's Disease in Sub-Saharan Africa." *Expert Review of Neurotherapeutics* 12 (6): 661–66.

Drouet, V., and S. Lesage. 2014. "Synaptojanin 1 Mutation in Parkinson’s Disease Brings Further Insight into the Neuropathological Mechanisms." *BioMed Research International* 2014 (September): e289728.

Edvardson, S., Y. Cinnamon, A. Ta-Shma., 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.

Fajardo, K. V. F., D. Adams, C. E Mason., et al. 2012. "Detecting False Positive Signals in Exome Sequencing." *Human Mutation* 33 (4): 609–13.

Fan, H-C., S-J. Chen, H-J. Harn., et al. 2013. "Parkinson's Disease: From Genetics to Treatments." *Cell Transplantation* 22 (4): 639–52.

Farrer, M. J., 2006. "Genetics of Parkinson Disease: Paradigm Shifts and Future Prospects." *Nature Reviews Genetics* 7 (4): 306–18.

Fortin, D. L., M. D. Troyer, K. Nakamura., et al. 2004. "Lipid Rafts Mediate the Synaptic Localization of Alpha-Synuclein." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 24 (30): 6715–23.

Funayama, M., K. Ohe, T. Amo., et al. 2015. "CHCHD2 mutations in autosomal dominant late-onset Parkinson's disease - a genome-wide linkage and sequencing study. " *Lancet Neurology.* 14 (3): 274-82.

Gasser, T. 2001. "Genetics of Parkinson's Disease." *Journal of Neurology* 248 (3): 833–40.

Gasser, T. 2010. "Chapter 1 - Identifying PD-Causing Genes and Genetic Susceptibility Factors: Current Approaches and Future Prospects." In *Recent Advances in Parkinson's Disease: Basic Research*, (183):2–20.

Geldenhuys, G., B. Glanzmann, D. Lombard., et al. 2014. "Identification of a Common Founder Couple for 40 South African Afrikaner Families with Parkinson's Disease." *South African Medical Journal* 104 (6): 413.

Gibbs, R. A., J. W. Belmont, P. Hardenbol., et al. 2003. "The International HapMap Project." *Nature* 426 (6968): 789–96.

Gibb, W R, and A J Lees. 1988. "A Comparison of Clinical and Pathological Features of Young- and Old-Onset Parkinson's Disease." *Neurology* 38 (9): 1402–6.

Glazov, E. A., A. Zankl, M. Donskoi., et al. 2011. "Whole-Exome Re-Sequencing in a Family Quartet Identifies POP1 Mutations As the Cause of a Novel Skeletal Dysplasia." *PLoS Genet* 7 (3): e1002027.

Glickman, M. H., and A. Ciechanover. 2002. "The Ubiquitin-Proteasome Proteolytic Pathway: Destruction for the Sake of Construction." *Physiological Reviews*, 82 (2): 372-428.

Glodzik, D., P. Navarro, V. Vitart., et al. 2013. "Inference of Identity by Descent in Population Isolates and Optimal Sequencing Studies." *European Journal of Human Genetics* 21 (10): 1140–45.

Goetz, C. G. 2011. "The History of Parkinson's Disease: Early Clinical Descriptions and Neurological Therapies." *Cold Spring Harbor Perspectives in Medicine* 1 (1): a008862.

Grada, A., and K. Weinbrecht. 2013. "Next-Generation Sequencing: Methodology and Application." *Journal of Investigative Dermatology* 133 (8): e11.

Greeff, J. M. 2007. "Deconstructing Jaco: Genetic Heritage of an Afrikaner." *Annals of Human Genetics* 71 (5): 674–88.

Greeff, J. M., and J. C. Erasmus. 2015. "Three Hundred Years of Low Non-Paternity in a Human Population." *Heredity*, Online Publication.

Gustavsson, E. K., J. Trinh, I. Guella., et al. 2015. "DNAJC13 Genetic Variants in Parkinsonism." *Movement Disorders: Official Journal of the Movement Disorder Society* 30 (2): 273–78.

Haas, R. H., F. Nasirian, K. Nakano,. et al. 1995. "Low Platelet Mitochondrial Complex I and Complex II/III Activity in Early Untreated Parkinson's Disease." *Annals of Neurology* 37 (6): 714–22.

Hall, D., E. M. Wijsman, J. L. Roos, et al. 2002. "Extended Intermarker Linkage Disequilibrium in the Afrikaners." *Genome Research* 12 (6): 956–61.

Hayden, M. R. 1980. "The Prevalence of Huntington's Chorea in South Africa." *The South African Medical Journal* 58 (5): 193–96.

Hayden, M. R., H. C. Hopkins, M. Macreaet., al. 1980. "The Origin of Huntington's Chorea in the Afrikaner Population of South Africa." *South African Medical Journal* 58 (5): 197–200.

Haylett, W. L., R. J. Keyser, M. C. du Plessis., et al. 2012. "Mutations in the Parkin Gene Are a Minor Cause of Parkinson's Disease in the South African Population." *Parkinsonism & Related Disorders* 18 (1): 89–92.

Hedges, D. J, D. Burges, E. Powell., et al. 2009. "Exome Sequencing of a Multigenerational Human Pedigree." *PloS One* 4 (12): e8232.

Hedrich, K., K. Marder, J. Harris., et al. 2002. "Evaluation of 50 Probands with Early-Onset Parkinson's Disease for Parkin Mutations." *Neurology* 58 (8): 1239–46.

Heese, J.A. 1971. *Die Herkoms van Die Afrikaner*. A.A Balkema.

Hershko, A., and A. Ciechanover. 1998. "The Ubiquitin System." *Annual Review of Biochemistry* 67 (1): 425.

Horner, D. S., G. Pavesi, T. Castrignano., et al. 2009. "Bioinformatics Approaches for Genomics and Post Genomics Applications of next-Generation Sequencing." *Briefings in Bioinformatics* 11 (2): 181–97.

Irrcher, I., H. Aleyasin, E. L. Seifert., et al. 2010. "Loss of the Parkinson's Disease-Linked Gene DJ-1 Perturbs Mitochondrial Dynamics." *Human Molecular Genetics* 19 (19): 3734–46.

Jankovic, Joseph. 2012. "Current Concepts in Parkinson's Disease and Other Movement Disorders." *Current Opinion in Neurology* 25 (4): 429–32.

Jasny, B. R., and L. Roberts. 2003. "Building on the DNA Revolution." *Science* 300 (5617): 277–277.

Kahle, P. J., and C. Haass. 2004. "How Does Parkin Ligate Ubiquitin to Parkinson's Disease?" *EMBO Reports* 5 (7): 681–85.

Kahle, P. J., J. Waak, and T. Gasser. 2009. "DJ-1 and Prevention of Oxidative Stress in Parkinson's Disease and Other Age-Related Disorders." *Free Radical Biology and Medicine* 47 (10): 1354–61.

Kaplin, A. I., M. Williams, D. G. Hirtz., et al. 2007. "How common are the 'common' neurlogic disorders?" *Neurology* 69 (4): 410–11.

Karayiorgou, M., M. Torrington, G. R. Abecasis., et al. 2004. "Phenotypic Characterization and Genealogical Tracing in an Afrikaner Schizophrenia Database." *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 124B (1): 20–28.

Keyser, R. J., S. Lesage, A. Brice., et al. 2010. "Assessing the Prevalence of PINK1 Genetic Variants in South African Patients Diagnosed with Early- and Late-Onset Parkinson's Disease." *Biochemical and Biophysical Research Communications* 398 (1): 125–29.

Kim, S. W., H. S. Ko, V. L. Dawson., et al. 2005. "Recent Advances in Our Understanding of Parkinson's Disease." *Drug Discovery Today: Disease Mechanisms* 2 (4): 427–33.

Kircher, M., D. M. Witten, P. Jain., et al.. 2014. "A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants." *Nature Genetics* 46 (3): 310–15.

Kitada, T., S. Asakawa, N. Hattori.,et al. 1998. "Mutations in the Parkin Gene Cause Autosomal Recessive Juvenile Parkinsonism." *Nature* 392 (6676): 605–8.

Kitts, A., L. Phan, M. Warden., et al.. 2014. "The Database of Short Genetic Variation (dbSNP)," April. http://www.ncbi.nlm.nih.gov/books/NBK174586/.

Klein, C., S. A. Schneider, and A. E. Lang. 2009. "Hereditary Parkinsonism: Parkinson Disease Look-alikes—An Algorithm for Clinicians to 'PARK' Genes and beyond." *Movement Disorders* 24 (14): 2042–58.

Krebs, C. E., S. Karkheiran, J. C. Powell., et al. 2013a. "The Sac1 Domain of SYNJ1 Identified Mutated in a Family with Early-Onset Progressive Parkinsonism with Generalized Seizures." *Human Mutation* 34 (9): 1200–1207.

Ku, C-S., N. Naidoo, and Y. Pawitan. 2011. "Revisiting Mendelian Disorders through Exome Sequencing." *Human Genetics* 129 (4): 351–70.

Kumar, A., J. D. Aguirre, T. E. Condos., et al. 2015. "Disruption of the Autoinhibited State Primes the E3 Ligase Parkin for Activation and Catalysis." *The EMBO Journal*, Online Publication.

Lander, E. S., L. M. Linton, B. Birren., et al. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822): 860–921.

Langston, J. W. 1983. "Parkinson's Disease in a Chemist Working with 1-Methyl-4-Phenyl-1,2,5,6-Tetrahydropyridine." *The New England Journal of Medicine* 309 (5): 310.

Lesage, S., C. Condroyer, S. Klebe., et al. 2012. "EIF4G1 in Familial Parkinson's Disease: Pathogenic Mutations or Rare Benign Variants?" *Neurobiology of Aging* 33 (9): 2233.e1–2233.e5.

Levine, B. 2005. "Eating Oneself and Uninvited Guests: Autophagy-Related Pathways in Cellular Defense." *Cell* 120 (2): 159–62.

Levine, Beth, N. Mizushima, and H. W. Virgin. 2011. "Autophagy in Immunity and Inflammation." *Nature* 469 (7330): 323–35.

Li, M-X., J. S. H. Kwan, S-Y. Bao., et al. 2013. "Predicting Mendelian Disease-Causing Non-Synonymous Single Nucleotide Variants in Exome Sequencing Studies." *PLoS Genet* 9 (1): e1003143. doi:10.1371/journal.pgen.1003143.

Luecking, C. B., A. Duerr, V. Bonifati., et al. 2000. "Association between Early-Onset Parkinson's Disease and Mutations in the Parkin Gene." *New England Journal of Medicine* 342 (21): 1560–67.

MacLeod, D. A., H. Rhinn, T. Kuwahara., et al. 2013. "RAB7L1 Interacts with LRRK2 to Modify Intraneuronal Protein Sorting and Parkinson's Disease Risk." *Neuron* 77 (3): 425–39.

Mann, V. M., J. M. Cooper, S. E. Daniel., et al. 1994. "Complex I, Iron, and Ferritin in Parkinson's Disease Substantia Nigra." *Annals of Neurology* 36 (6): 876–81.

Manolio, T. A., L. D. Brooks, and F. S. Collins. 2008. "A HapMap Harvest of Insights into the Genetics of Common Disease." *Journal of Clinical Investigation* 118 (5): 1590–1605.

Manzoni, C., and P. A. Lewis. 2013. "Dysfunction of the Autophagy/lysosomal Degradation Pathway Is a Shared Feature of the Genetic Synucleinopathies." *FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology* 27 (9): 3424–29.

Martinat, C., S. Shendeiman, A. Ionason., et al. 2004. "Sensitivity to Oxidative Stress in DJ-1-Deficient Dopamine Neurons: An ES-Derived Cell Model of Primary Parkinsonism." *PLoS Biology* 2 (11): 1754–63.

Matsuda, S., Y. Kitagishi, and M. Kobayashi. 2013. "Function and Characteristics of PINK1 in Mitochondria." *Oxidative Medicine and Cellular Longevity* 2013: 601587.

Matsumine, H., M. Saito, S. Shimoda-Matsubayashi., et al. 1997. "Localization of a Gene for an Autosomal Recessive Form of Juvenile Parkinsonism to Chromosome 6q25.2-27." *American Journal of Human Genetics* 60 (3): 588.

McMillan, C. T., J. B. Toledo, B. B. Avants., et al. 2014. "Genetic and Neuroanatomic Associations in Sporadic Frontotemporal Lobar Degeneration." *Neurobiology of Aging*. 35 (6): 1473-82.

McPherson, P. S., E. P. Garcia, V. I. Slepnev., et al. 1996. "A Presynaptic Inositol-5-Phosphatase." *Nature* 379 (6563): 353–57.

Medeiros-Domingo, A., T. Kaku, D. J. Tester., et al. 2007. "SCN4B-Encoded Sodium Channel β4 Subunit in Congenital Long-QT Syndrome." *Circulation* 116 (2): 134–42.

Melcon, M. O., D. W. Anderson, R. H. Vergara., et al. 1997. "Prevalence of Parkinson's Disease in Junín, Buenos Aires Province, Argentina." *Movement Disorders* 12 (2): 197–205.

Metzenberg S. 2008. "Sanger Method." http://www.csun.edu/%7Ehcbio027/biotechnology/lec3/sanger.html.

Moore, D. J., A. B. West, V. L. Dawson., et al. 2005. "Molecular Pathophysiology of Parkinson's Disease." *Annual Review of Neuroscience* 28 (1): 57–87.

Muangpaisan, W., H. Hori, and C. Brayne. 2009. "Systematic Review of the Prevalence and Incidence of Parkinson's Disease in Asia." *Journal of Epidemiology* 19 (6): 281–93.

Nagakubo, D., T. Taira, H. Kitaura., et al. 1997. "DJ-1, a Novel Oncogene Which Transforms Mouse NIH3T3 Cells in Cooperation Withras." *Biochemical and Biophysical Research Communications* 231 (2): 509–13.

Nalls, M. A., N. Pankratz, C. M. Lill., et al. 2014. "Large-Scale Meta-Analysis of Genome-Wide Association Data Identifies Six New Risk Loci for Parkinson's Disease." *Nature Genetics* 46 (9): 989–93.

Narendra, D., A. Tanaka, D-F. Suen., et al. 2008. "Parkin Is Recruited Selectively to Impaired Mitochondria and Promotes Their Autophagy." *Journal of Cell Biology* 183 (5): 795–803.

Ng, S B., K. J. Buckingham, C. Lee., et al. 2010. "Exome Sequencing Identifies the Cause of a Mendelian Disorder." *Nature Genetics* 42 (1): 30–35.

Ng, S. B., E. H. Turner, P. D. Robertson., et al. 2009. "Targeted Capture and Massively Parallel Sequencing of 12 Human Exomes." *Nature* 461 (7261): 272–76.

Nichols, N., J. M. Bras, D. G. Hernandez., et al. 2015. "EIF4G1 Mutations Do Not Cause Parkinson's Disease." *Neurobiology of Aging* 36 (8): 2444.e1–2444.e4.

Nicklas, W. J. 1987. "MPTP, MPP and Mitochondrial Function." *Life Sciences* 40 (8): 721–29.

Nijman, S.M.B., M. P. Luna-Vargas, A. Velds., et al. 2005. "A Genomic and Functional Inventory of Deubiquitinating Enzymes." *Cell* 123 (5): 773–86.

Niki, T., K. Takahashi-Niki, T. Taira., et al. 2003. "DJBP: A Novel DJ-1-Binding Protein, Negatively Regulates the Androgen Receptor by Recruiting Histone Deacetylase Complex, and DJ-1 Antagonizes This Inhibition by Abrogation of This Complex1. " *Molecular Cancer Research* 1 (4): 247–61.

Nishioka, K., M. Funayama, C. Vilariño-Güell., et al. 2014. "EIF4G1 Gene Mutations Are Not a Common Cause of Parkinson's Disease in the Japanese Population." *Parkinsonism & Related Disorders* 20 (6): 659–61.

Novarino, G., A. G. Fenstermaker, M. S. Zaki., et al. 2014. "Exome Sequencing Links Corticospinal Motor Neuron Disease to Common Neurodegenerative Disorders." *Science* 343 (6170): 506–11.

Okubadejo, N. U. 2008. "An Analysis of Genetic Studies of Parkinson's Disease in Africa." *Parkinsonism & Related Disorders* 14 (3): 177–82.

Okubadejo, N. U., J. H. Bower, W. A. Rocca., et al. 2006. "Parkinson's Disease in Africa: A Systematic Review of Epidemiologic and Genetic Studies." *Movement Disorders* 21 (12): 2150–56.

Olgiati, S., A. De Rosa, M. Quadri., et al. 2014. "PARK20 Caused by SYNJ1 Homozygous Arg258Gln Mutation in a New Italian Family." *Neurogenetics* 15 (3): 183–88.

Pabinger, S., A. Dander, M.Fischer., et al. 2013. "A Survey of Tools for Variant Analysis of next-Generation Genome Sequencing Data." *Briefings in Bioinformatics*, January, bbs086.

Pan, T., S. Kondo, W. Le., et al. 2008. "The Role of Autophagy-Lysosome Pathway in Neurodegeneration Associated with Parkinson's Disease." *Brain* 131 (8): 1969–78.

Parkinson, J. 1817. "An Essay on the Shaking Palsy." *Journal of Neuropsychiatry* 14 (2): 223–36.

Patel, Z. H., L. C. Kottyan, S. Lazaro., et al. 2014. "The Struggle to Find Reliable Results in Exome Sequencing Data: Filtering out Mendelian Errors." *Frontiers in Genetics* 5. Online publication.

Patterson, N., D. C. Petersen, V. Der Ross., et al. 2010. "Genetic Structure of a Unique Admixed Population: Implications for Medical Research." *Human Molecular Genetics* 19 (3): 411–19.

Pearce, J. M. 1989. "Aspects of the History of Parkinson's Disease." *Journal of Neurology, Neurosurgery & Psychiatry* 52 (Suppl): 6–10.

Pena, S., and R. Coimbra. 2015. "Ataxia and Myoclonic Epilepsy due to a Heterozygous New Mutation in KCNA2: Proposal for a New Channelopathy." *Clinical Genetics* 87 (2): e1–3.

Petit, A., T. Kawarai, E. Paitel., et al. 2005. "Wild-Type PINK1 Prevents Basal and Induced Neuronal Apoptosis, a Protective Effect Abrogated by Parkinson Disease-Related Mutations." *Journal of Biological Chemistry* 280 (40): 34025–32.

Polymeropoulos, M. H., J. J. Higgins, L. I. Golbe., et al. 1996. "Mapping of a Gene for Parkinson's Disease to Chromosome 4q21-q23." *Science* 274(5290): 1197-9.

Poulopoulos, M., O. A. Levy, and R. N. Alcalay. 2012. "The Neuropathology of Genetic Parkinson's Disease." *Movement Disorders* 27 (7): 831–42.

Pringsheim, T., N. Jette, A. Frolkis., et al. 2014. "The Prevalence of Parkinson's Disease: A Systematic Review and Meta-Analysis." *Movement Disorders* 29 (13): 1583–90.

Purcell, S., B. Neale, K. Todd-Brown., et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *The American Journal of Human Genetics* 81 (3): 559–75.

Quadri, M., M. Fang, M. Picillo., et al. 2013. "Mutation in the SYNJ1 Gene Associated with Autosomal Recessive, Early-Onset Parkinsonism." *Human Mutation* 34 (9): 1208–15.

Rackham, O. J. L., H. A. Shihab, M. R. Johnson., et al. 2014. "EvoTol: A Protein-Sequence Based Evolutionary Intolerance Framework for Disease-Gene Prioritization." *Nucleic Acids Research*. Online publication.

Ramirez, A., A. Heimbach, J. Gründemann., et al. 2006. "Hereditary Parkinsonism with Dementia Is Caused by Mutations in ATP13A2, Encoding a Lysosomal Type 5 P-Type ATPase." *Nature Genetics* 38 (10): 1184–91.

Ramjaun, A. R., and P. S. McPherson. 1996. "Tissue-Specific Alternative Splicing Generates Two Synaptojanin Isoforms with Differential Membrane Binding Properties." *Journal of Biological Chemistry* 271 (40): 24856–61.

Raudino, F. 2011. "The Parkinson Disease before James Parkinson." *Neurological Sciences* 33 (4): 945–48.

Reyes-Turcu, F. E., K. H. Ventii, and K. D. Wilkinson. 2009. "Regulation and Cellular Roles of Ubiquitin-Specific Deubiquitinating Enzymes." *Annual Review of Biochemistry* 78 (1): 363–97.

Robinson, P. N., P. Krawitz, and S. Mundlos. 2011. "Strategies for Exome and Genome Sequence Data Analysis in Disease-Gene Discovery Projects." *Clinical Genetics* 80 (2): 127–32.

Rohe, C. F., P. Montagna, G. Breedveld., et al. 2004. "Homozygous PINK1 C-Terminus Mutation Causing Early-Onset Parkinsonism." *Annals of Neurology* 56 (3): 427–31.

Roos, J. L., H. W. Pretorius, and M. Karayiorgou. 2009. "Clinical Characteristics of an Afrikaner Founder Population Recruited for a Schizophrenia Genetic Study." *Annals of the New York Academy of Sciences* 1151 (1): 85–101.

Rothfuss, O., H. Fischer, T. Hasegawa., et al. 2009. "Parkin Protects Mitochondrial Genome Integrity and Supports Mitochondrial DNA Repair." *Human Molecular Genetics* 18 (20): 3832–50.

Saux, O., K. Beck, C. Sachsinger., et al. 2002. "Evidence for a Founder Effect for Pseudoxanthoma Elasticum in the Afrikaner Population of South Africa." *Human Genetics* 111 (4-5): 331–38.

Savica R., W. A Rocca, and J. Ahlskog. 2010. "When Does Parkinson Disease Start?" *Archives of Neurology* 67 (7): 798–801.

Schapira, A. H. V., T. Warner, M. T. Gash., et al. 1997. "Complex I Function in Familial and Sporadic Dystonia." *Annals of Neurology* 41 (4): 556–59. doi:10.1002/ana.410410421.

Schapira, A.H.V. 2010. "Complex I: Inhibitors, Inhibition and Neurodegeneration." *Experimental Neurology* 224 (2): 331–35.

Schloss, J.A. 2008. "How to Get Genomes at One Ten-Thousandth the Cost." *Nature Biotechnology* 26 (10): 1113–15. doi:10.1038/nbt1008-1113.

Shimura, H., N. Hattori, S. Kubo., et al. 2000. "Familial Parkinson Disease Gene Product, Parkin, Is a Ubiquitin-Protein Ligase." *Nature Genetics* 25 (3): 302.

Sidransky, E., and G. Lopez. 2012. "The Link between the GBA Gene and Parkinsonism." *The Lancet Neurology* 11 (11): 986–98. doi:10.1016/S1474-4422(12)70190-4.

Siitonen, A., E. Majounie, M. Federoff., et al. 2013. "Mutations in EIF4G1 Are Not a Common Cause of Parkinson's Disease." *European Journal of Neurology* 20 (4): e59–e59.

Singleton, A. 2015. "A New Gene for Parkinson's Disease: Should We Care?" *The Lancet. Neurology* 14 (3): 238–39.

Song, D. D., C. W. Shults, A. Sisk., et al. 2004. "Enhanced Substantia Nigra Mitochondrial Pathology in Human A-Synuclein Transgenic Mice after Treatment with MPTP." *Experimental Neurology* 186 (2): 158–72.

Speed, D., and D. J. Balding. 2015. "Relatedness in the Post-Genomic Era: Is It Still Useful?" *Nature Reviews Genetics* 16 (1): 33–44.

Spencer, C. A., V. Plagnol, A. Strange., et al. 2011. "Dissection of the Genetics of Parkinson's Disease Identifies an Additional Association 5′ of SNCA and Multiple Associated Haplotypes at 17q21." *Human Molecular Genetics* 20 (2): 345–53.

Srour, M., F. F. Hamdan, Z. Gan-Or., et al. 2015. "A Homozygous Mutation in SLC1A4 in Siblings with Severe Intellectual Disability and Microcephaly." *Clinical Genetics*, 88 (1): e1-4.

Storz, G., and J. A. Imlayt. 1999. "Oxidative Stress." *Current Opinion in Microbiology* 2 (2): 188–94.

Surguchov, A., and K. W. Jeon. 2008. "Chapter 6 Molecular and Cellular Biology of Synucleins." In *International Review of Cell and Molecular Biology*, 270: 225–317.

Takahata, N. 1993. "Allelic Genealogy and Human Evolution." *Molecular Biology and Evolution* 10 (1): 2–22.

Tanner, C. M., R. Ottman, S. M. Goldman., et al. 1999. "Parkinson Disease in Twins: An Etiologic Study." *Journal of the American Medical Association* 281 (4): 341–46.

Tipping, A. J., T. Pearson, N. V. Morgan., et al. 2001. "Molecular and Genealogical Evidence for a Founder Effect in Fanconi Anemia Families of the Afrikaner Population of South Africa." *Proceedings of the National Academy of Sciences* 98 (10): 5734–39.

Trempe, J-F., and E. A. Fon. 2013. "Structure and Function of Parkin, PINK1, and DJ-1, the Three Musketeers of Neuroprotection." *Frontiers in Neurodegeneration* 4: 38..

Trinh, J., and M. Farrer. 2013. "Advances in the Genetics of Parkinson Disease." *Nature Reviews Neurology* 9 (8): 445.

Tucci, A., G. Charlesworth, U.-M. Sheerin., et al. 2012. "Study of the Genetic Variability in a Parkinson's Disease Gene: EIF4G1." *Neuroscience Letters* 518 (1): 19–22.

Valente, E. M., P. M. Abou-Sleiman, V. Caputo., et al. 2004. "Hereditary Early-Onset Parkinson's Disease Caused by Mutations in PINK1." *Science (New York, N.Y.)* 304 (5674): 1158–60.

Van de Warrenburg, B. P. C., M. Lammens, C.B. Lucking., et al. 2001. "Clinical and Pathologic Abnormalities in a Family with Parkinsonism and Parkin Gene Mutations." *Neurology February 27, 2001* 56 (4): 555–57.

Van Dijk, E. L., H. Auger, Y. Jaszczyszyn., et al. 2014. "Ten Years of next-Generation Sequencing Technology." *Trends in Genetics* 30 (9): 418–26.

Vandrovcova, J., F. Anaya, V. Kay., et al. 2010. "Disentangling the Role of the Tau Gene Locus in Sporadic Tauopathies." *Current Alzheimer Research* 7 (8): 726–34.

Van Duijn, C. M., M. C. Dekker, V. Bonifati., et al. 2001. "Park7, a Novel Locus for Autosomal Recessive Early-Onset Parkinsonism, on Chromosome 1p36." *American Journal of Human Genetics* 69 (3): 629–34.

Venter, C. J., M. D. Adams, E. W. Myers., et al. 2001. "The Sequence of the Human Genome." *Science* 291 (5507): 1304–51.

Verstraeten, A., J. Theuns, and C. Van Broeckhoven. 2015. "Progress in Unraveling the Genetic Etiology of Parkinson Disease in a Genomic Era." *Trends in Genetics* 31 (3): 140–49.

Vilariño- Güell, C., A. I. Soto, S. J. Lincoln., et al. 2008. "ATP13A2 Variability in Parkinson Disease." *Human Mutation* 30 (3): 406–10.

Vilariño-Güell, C., C. Wider, O. A. Ross., et al. 2011. "VPS35 Mutations in Parkinson Disease." *American Journal of Human Genetics* 89 (1): 162–67.

Vucic, D., V. M. Dixit, and I. E. Wertz. 2011. "Ubiquitylation in Apoptosis: A Post-Translational Modification at the Edge of Life and Death." *Nature Reviews Molecular Cell Biology* 12 (7): 439–52.

Wang, J. L., X. Yang, K.Xia., et al. 2010. "TGM6 Identified as a Novel Causative Gene of Spinocerebellar Ataxias Using Exome Sequencing." *Brain: A Journal of Neurology* 133 (Pt 12): 3510–18.

Wilson, M. A., C. V. St. Amour, J. L. Collins., et al. 2004. "The 1.8-Å Resolution Crystal Structure of YDR533Cp from Saccharomyces Cerevisiae: A Member of the DJ-1/ThiJ/PfpI Superfamily." *Proceedings of the National Academy of Sciences of the United States of America* 101 (6): 1531–36.

Ye, M. H., J. L. Chen, G. P. Zhao., et al. 2010. "Sensitivity and Specifcity of High-Resolution Melting Analysis in Screening Unknown Snps and Genotyping a Known Mutation." *Animal Science Papers and Reports* 28 (2): 161–70.

Zhang, Y., Y. Xu, Q. Zhao., et al. 2012. "The Structural Characteristics of Human Preprotein Translocase of the Inner Mitochondrial Membrane Tim23: Implications for Its Physiological Activities." *Protein Expression and Purification* 82 (2): 255–62.

Zhou, L., X. Ma, and F. Sun. 2008. "The Effects of Protein Interactions, Gene Essentiality and Regulatory Regions on Expression Variation." *BMC Systems Biology* 2 (1): 54.

Zimprich, A., S. Biskup, P. Leitner. 2004. "Mutations in LRRK2 Cause Autosomal-Dominant Parkinsonism with Pleomorphic Pathology." *Neuron* 44 (4): 601–7.

Zimprich, A., A. Benet-Pagès, W. Struhal., et al. 2011. "A Mutation in VPS35, Encoding a Subunit of the Retromer Complex, Causes Late-Onset Parkinson Disease." *The American Journal of Human Genetics* 89 (1): 168–75.

URL Reference List

- Virtual MedStudent.com: [http://www.virtualmedstudent.com/links/neurological/parkinsons html] Date accessed: 03/03/2013
- CCDS Report for Consensus CDS : [http://www.ncbi nlm nih.gov/CCDS/CcdsBrowse.cgi] Date accessed: 10/02/2013
- Wellcome Trust Sanger Institute: [http://www.sanger.ac.uk/gencode/] Date accessed: 10/02/2013
- Agilent Technologies: [http://www.genomics.agilent.com/CollectionSubpage.aspx?PageType=Product&SubPageType=ProductData&PageID=2318] Date accessed: 10/02/2013
- NCBI Reference Sequence: [http://www ncbi.nlm nih.gov/RefSeq/] Date accessed: 10/02/2013
- Illumina.com: [http://www.illumina.com/documents/products/datasheets/datasheet_truseq_exome_enrichment_kit.pdf] Date accessed: 10/02/2014
- Gencodegenes.org: [http://www.gencodegenes.org/] Date accessed: 10/02/2014
- Mirbase.org: [http://www.mirbase.org/] Date accessed: 10/02/2014
- dbSNP: [http://www.ncbi.nlm.nih.gov/projects/SNP/)] Date accessed: 10/02/2015
- Ensembl.org:[www.ensembl.org] Date accessed:10/07/2014
- Roche NimbleGen: [http://www nimblegen.com/products/seqcap/ez/v3/index html] Date accessed: 10/07/2014
- Variant Effect Predictor Tool: [http://www.ensembl.org/tools.htm] Date accessed: 25/09/2013
- SMART Protein Analyser: [http://smart.embl-heidelberg.de/] Date accessed: 25/03/2015
- Real Time PCR: [www.bio.davidson.edu/courses/molbio/molstudents/spring2003/pierce/realtimepcr htm] Date accessed: 06/04/2013
- QuantiTect® Multiplex PCR Handbook [www.qiagen.com/hb/quantitectmultiplexpcr] Date accessed: 18/09/2013
- DNA Software: [http://dnasoftware.com/FretAssays/tabid/139/Default.aspx] Date accessed: 18/09/2014
- GEOPD: [http://www.geopd.org] Date accessed: 20/10/2014
- World Federation of Neurology: [http://www.wfneurology.org/the-african-experience] Date accessed: 20/10/2014
- Kernan Health [http://health.kernan.org/imagepages/19515.htm] Date accessed 06/07/2014
- Cell Signal Technology [http://www.cellsignal.com/reference/pathway/Ubiquitin_Proteasome.html] Date accessed 06/07/2014
- 1000 Genome Project [http://www.1000genomes.org/] Date accessed 13/03/2013
- ESP6500 [http://evs.gs.washington.edu/EVS/] Date accessed 13/03/2013

Appendix I Diagnostic criteria for Parkinson's disease

Step 1: Diagnosis of Parkinsonian Syndrome

- Bradykinesia and at least two of the following:
  - Muscular rigidity
  - Resting tremor of 4-6 Hertz
  - Postural instability that is *not* by primary visual, vestibular, cerebellar or proprioceptive dysfunction

Step 2: Exclusion criteria for PD

- History of repeated strokes with stepwise progression of Parkinsonian features
- History of repeated head injury and definite encephalitis
- Oculogyric crises
- Neuroleptic treatment at onset of symptoms
- More than one affected relative
- Sustained remission
- Strictly unilateral features after 3 years
- Supranuclear gaze palsy
- Cerebellar signs
- Early severe autonomic involvement
- Early severe dementia with disturbances of memory, language, and praxis
- Babinski sign
- Presence of cerebral tumour or communication hydrocephalus on imaging study
- Negative response to large doses of levodopa in absence of malabsorption
- 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) exposure

Step 3: Supportive prospective positive criteria for PD

- Three or more are required for the diagnosis of definite PD in conjunction with Step 1:
  - Unilateral onset
  - Resting tremor must be present
  - Progressive disorder
  - Persistent asymmetry affecting the side of onset most
  - Excellent response to levodopa (70-100%) with response for 5 years
  - Severe levodopa induced chorea

Appendix II

Supplementary Table 1 List of the demographic and clinical information on the 40 South African Afrikaner probands and their family members which are available for genetic studies.

Family No	*Family ID*	*Gender*	*Relationship*	*Affected*	*AAO*	*Pedigree Pattern*
1	ZA 64	Female	Proband	Y	63	AD
		Male	Son	N		
2	ZA 95	Female	Proband	Y	55	AD
		Male	Son	N		
		Female	Sister	N		
		Female	Daughter	N		
3	ZA 233	Male	Proband	Y	36	AR
		Female	Mother	N		
		Female	Sister	N		
4	ZA 134	Male	Proband	Y	62	AD
		Female	Daughter	N		
5	ZA 272	Male	Proband	Y	62	AD
6	ZA 72	Male	Proband	Y	68	AD
7	ZA 263	Male	Proband	Y	58	AD
8	ZA 213	Male	Proband	Y	73	AR
		Female	Daughter	N		
		Female	Wife	N		
9	ZA 112	Female	Proband	Y	53	AD
		Male	Son	N		
		Male	Son	N		
10	ZA 89	Female	Proband	Y	47	AD
		Male	Husband	N		
		Male	Son	N		
		Female	Daughter	N		
		Female	Daughter	N		
		Female	Sister	N		
		Male	Son	N		
11	ZA 86	Male	Proband	Y	50	AD
		Male	Son	N		
		Female	Daughter	N		
		Female	Sister	N		

12	ZA 241	Male	Proband	Y	37	AD
		Female	Sister	N		
13	ZA 92	Female	Proband	Y	55	AR
		Female	Daughter	N		
		Male	Husband	N		
		Female	Daughter	N		
		Female	Daughter	N		
		Male	Nephew	N		
		Male	Brother	Y	57	
		Female	Sister in-law	N		
		Female	Sister	N		
14	ZA 34	Male	Proband	Y	72	AR
15	ZA 140	Male	Proband	Y	48	AD
		Female	Daughter	N		
		Male	Son	N		
		Female	Mother	N		
		Male	Brother	N		
		Female	Daughter	N		
		Male	Son	N		
16	ZA 252	Male	Proband	Y	39	AD
		Male	Father	N		
		Female	Sister	N		
		Female	Mother	N		
17	ZA 194	Male	Proband	Y	55	AR
		Male	Brother	N		
		Female	Daughter	N		
18	ZA 142	Male	Proband	Y	50	AD
		Male	Cousin	N		
		Male	Brother	N		
		Female	Sister	N		
		Female	Mother	N		
		Male	Son	N		
		Male	Son	N		
		Male	Son	N		
		Male	Son	N		

		Female	Daughter	N		
		Female	Sister	Y	53	
19	ZA 413	Female	Proband	Y	54	AR
		Male	Cousin	Y	67	
20	ZA 103	Male	Proband	Y	41	AD
		Female	Wife	N		
		Female	Sister	N		
		Female	Sister	N		
		Male	Cousin	N		
		Male	Cousin	N		
		Male	Cousin	N		
		Male	Cousin	N		
		Male	Uncle	Y	66	
21	ZA 116	Male	Proband	Y	40	AR
		Female	Sister	N		
		Female	Daughter	N		
		Female	Daughter	N		
		Female	Sister	N		
22	ZA 68	Male	Proband	Y	49	AD
		Female	Wife	N		
		Female	Daughter	N		
		Male	Son	N		
		Female	Sister	N		
		Female	Sister	N		
23	ZA 5	Male	Proband	Y	40	AR
24	ZA 190	Female	Proband	Y	47	AD
25	ZA 76	Male	Proband	Y	65	AD
		Male	Son	N		
		Male	Son	N		
26	ZA 253	Male	Proband	Y	40	AD
		Male	Nephew	Y	37	
		Female	Sister	N		
		Male	Brother	Y	48	
		Male	Nephew	N		
		Male	Cousin	N		

		Male	Brother-in-law	N		
27	ZA 411	Male	Proband	Y	65	AR
28	ZA 24	Female	Proband	Y	69	AR
		Female	Sister	Y	61	
29	ZA 106	Male	Proband	Y	49	AR
		Male	Son	N		
		Female	Daughter	N		
		Female	Sister	N		
		Female	Second cousin	N		
		Male	Second cousin	N		
		Female	Second cousin	N		
		Female	Second cousin	N		
30	ZA 78	Male	Proband	Y	54	AD
		Female	Daughter	N		
		Male	Son	N		
		Female	Wife	N		
31	ZA 4	Male	Proband	Y	44	AR
32	ZA 216	Male	Proband	Y	34	AD
		Female	Daughter	N		
		Female	Daughter	N		
33	ZA 340	Male	Proband	Y	68	AR
		Male	Brother	Y	48	
		Male	Brother	Y	58	
		Male	Brother	N		
		Female	Sister	N		
		Male	Brother	N		
		Female	Sister	N		
34	ZA 175	Male	Proband	Y	55	AD
		Female	Sister	N		
		Female	Daughter	N		
		Male	Brother	N		
35	ZA 316	Male	Proband	Y	58	AR
36	ZA 111	Male	Proband	Y	58	AD
		Female	Daughter	N		

		Male	Brother	N		
		Female	Sister	N		
		Male	Son	N		
		Female	Sister	N		
		Male	Son	N		
		Female	Daughter	N		
		Male	Twin	N		
		Female	Niece	N		
		Female	Wife	N		
37	ZA 16	Female	Proband	Y	27	AR
		Female	Sister	Y	27	
		Male	Father	N		
		Female	Mother	N		
		Female	Niece	N		
		Male	Nephew	N		
		Male	Son	N		
		Male	Brother	N		
		Female	Sister	N		
38	ZA 150	Female	Proband	Y	50	AR
39	ZA 256	Male	Proband	Y	45	AD
40	ZA 128	Female	Proband	Y	38	AD
		Female	Daughter	N		
		Female	Daughter	N		
		Female	Sister	N		
		Female	Mother	N		

PD = Parkinson's disease; ID = identification; AAO = age at onset of Parkinson's disease in years; F = female; M = male; Y = Yes; N = No; AR = autosomal recessive; AD = autosomal dominant

Appendix III – Supplementary information pertaining to IBD

Supplementary Table 2 Quality control performed on the six related probands traced back to a common founder couple.

Family ID	Individual ID	MISS_PHENO	N_MISS	N_GENO	F_MISS	HET_RATE
ZA106	78.67	N	279	306 670	0.0009098	0.404873188
ZA111	78.95	N	6770	306 670	0.02175	0.415456146
ZA134	81.65	N	632	306 670	0.002061	0.402841418
ZA78	67.82	N	330	306 670	0.001076	0.407993171
ZA89	68.16	N	948	306 670	0.003111	0.404157634
ZA92	68.27	N	826	306 670	0.00271	0.407397957

MISS_PHENO – missing phenotype (Yes/No); N_MISS – number of missing SNPs; N_GENO – number of non-obligatory genotypes; F_MISS – proportion of missing SNPs (genotype call rate); HET_RATE – heterozygosity rate

Supplementary Table 3 Shared segments between the affected sibling pair of family ZA92.

CHR	BP1	BP2	SNP1	SNP2	NSNP	KB
1	16773880	116717619	rs821000	rs6658845	1698	99943.7
1	160152057	236404391	rs1935306	rs2695057	1554	76252.3
1	244965588	249081330	rs2365687	rs4926506	101	4115.74
2	34739570	54708408	rs10176097	rs10171331	577	19968.8
2	57658970	77947988	rs2310825	rs11676317	431	20289
2	108033742	242955426	rs266243	rs13389823	2380	134922
3	70895	12286720	rs9682794	rs9850825	402	12215.8
3	16911330	54542933	rs9852648	rs7372541	719	37631.6
3	74345173	81997510	rs9867036	rs6548786	137	7652.34
3	88708334	99722396	rs7638288	rs6440967	122	11014.1
3	154494952	164537423	rs6440957	rs4560300	166	10042.5
3	170533587	175171066	rs6790486	rs1549108	113	4637.48
3	176087269	182547256	rs301230	rs2049283	116	6459.99
3	191020382	195573324	rs293862	rs903196	111	4552.94
4	85422	63509409	rs7667153	rs12500171	1264	63424
4	86652248	95155505	rs2062098	rs17309887	141	8503.26
4	110409978	142570472	rs6856291	rs1519551	606	32160.5
4	165129362	170237660	rs4502640	rs10017932	107	5108.3
5	48534	180690937	rs10039735	rs1279912	3462	180642
6	904145	4058782	rs2756313	rs12198921	107	3154.64
6	21347410	63457733	rs1555083	rs2474878	859	42110.3
6	88671941	108147644	rs632385	rs4946872	388	19475.7
6	139807600	170886531	rs6916887	rs8770	819	31078.9
7	44935	4708134	rs7456436	rs17828856	116	4663.2
7	37712109	55238268	rs4723693	rs10228436	420	17526.2

7	69935641	105456869	rs4717530	rs740309	631	35521.2
7	130154485	156711661	kgp9503409	rs2969124	525	26557.2
8	5857354	72116724	rs890027	rs10957540	1274	66259.4
8	77507613	146245372	rs7846606	rs35756786	1319	68737.8
9	185632	137530346	rs2992854	rs7021140	2242	137345
10	135656	135430043	rs10904561	rs4628635	2683	135294
11	17530484	30452181	rs7104083	rs514644	323	12921.7
11	75428958	109883416	rs554202	rs1648136	731	34454.5
11	116383064	134926754	rs11599994	rs6590788	509	18543.7
12	191619	21635232	rs11063263	rs3213212	498	21443.6
12	28231087	55287990	rs10843085	rs9943768	468	27056.9
12	102376027	118749798	rs4764862	rs461499	359	16373.8
12	124854903	133754245	rs3782257	rs12320759	304	8899.34
13	58000353	78044267	rs9569686	rs4885469	440	20043.9
13	106210736	115025398	rs7991826	rs11617448	272	8814.66
14	19465246	95108821	kgp9056199	rs11160197	1577	75643.6
15	20184600	29418573	rs12906138	rs2672680	135	9233.97
15	49832378	92684742	rs12232355	rs4777789	807	42852.4
16	6724106	52625613	rs9929593	rs12920540	631	45901.5
16	64439176	79906401	rs8054941	rs12102675	337	15467.2
16	82303566	85370416	rs2967321	rs2326526	153	3066.85
17	72487	9764531	rs12450662	rs2001486	274	9692.04
17	12574501	81051007	rs1519251	rs7502442	1094	68476.5
18	13034	3952770	rs12455984	rs6506142	143	3939.74
18	57181744	78014582	rs4558500	rs12456851	517	20832.8
19	288374	29219850	rs12981067	rs4474806	532	28931.5
20	127720	7671888	rs753217	rs6055258	239	7544.17
20	12693139	62892739	rs1333400	rs6062357	1089	50199.6
21	16229925	36219566	rs2822964	rs2154450	491	19989.6
22	16504399	49559242	kgp1568720	rs5769440	753	33054.8

CHR – chromosome; BP1 – start of the physical position of the segment (base pair); BP2 - end of the physical position of the segment (base pair 2); SNP1 – start of the SNP segment; SNP2 – end of the SNP segment; NSNP – number of SNPs in the segment; KB – physical length of the segment

Supplementary Table 4 Individual per sample comparison between the 40 probands.

FID1	IID1	FID2	IID2	Z0	Z1	Z2	PI(HAT)
ZA103	7268	ZA106	7867	0.9981	0	0.0019	0.0019
ZA103	7268	ZA112	7897	1	0	0	0
ZA103	7268	ZA116	7929	1	0	0	0
ZA103	7268	ZA128	8158	0.9945	0	0.0055	0.0055
ZA103	7268	ZA134	8165	0.9832	0.0168	0	0.0084
ZA103	7268	ZA140	8174	1	0	0	0
ZA103	7268	ZA142	8239	1	0	0	0

ZA103	7268	ZA150	5645	1	0	0	0
ZA103	7268	ZA16	10181	0.9986	0	0.0014	0.0014
ZA103	7268	ZA175	8316	1	0	0	0
ZA103	7268	ZA190	8338	1	0	0	0
ZA103	7268	ZA194	8415	1	0	0	0
ZA103	7268	ZA213	8430	0.989	0.0095	0.0016	0.0063
ZA103	7268	ZA216	8874	0.98	0.02	0	0.01
ZA103	7268	ZA233	5473	0.9714	0.0275	0.0011	0.0149
ZA103	7268	ZA24	9087	0.9639	0.0361	0	0.018
ZA103	7268	ZA241	9200	0.9893	0.0091	0.0016	0.0062
ZA103	7268	ZA252	9231	0.9818	0.0165	0.0017	0.0099
ZA103	7268	ZA253	9260	1	0	0	0
ZA103	7268	ZA256	9434	1	0	0	0
ZA103	7268	ZA263	9458	1	0	0	0
ZA103	7268	ZA272	9564	1	0	0	0
ZA103	7268	ZA316	5581	1	0	0	0
ZA103	7268	ZA34	7594	0.0134	0	0	0.0066
ZA103	7268	ZA4	3836	0.9767	0.0233	0	0.0116
ZA103	7268	ZA411	10110	1	0	0	0
ZA103	7268	ZA413	10141	0.9877	0.0123	0	0.0062
ZA103	7268	ZA5	3823	1	0	0	0
ZA103	7268	ZA64	6734	1	0	0	0
ZA103	7268	ZA68	6764	0.9754	0.0246	0	0.0123
ZA103	7268	ZA72	6768	0.9953	0.0021	0.0026	0.0036
ZA103	7268	ZA76	6782	0.9857	0.0143	0	0.0072
ZA103	7268	ZA78	6772	1	0	0	0
ZA103	7268	ZA86	6807	1	0	0	0
ZA103	7268	ZA89	6816	0.9883	0.0107	0.001	0.0063
ZA103	7268	ZA92	6827	0.9823	0.0177	0	0.0088
ZA103	7268	ZA95	6924	0.9837	0.0163	0	0.0082
ZA106	7867	ZA112	7897	1	0	0	0
ZA106	7867	ZA116	7929	0.9943	0.0057	0	0.0029
ZA106	7867	ZA128	8158	1	0	0	0
ZA106	7867	ZA134	8165	0.9807	0.0193	0	0.0096
ZA106	7867	ZA140	8174	1	0	0	0
ZA106	7867	ZA142	8239	1	0	0	0
ZA106	7867	ZA150	5645	1	0	0	0
ZA106	7867	ZA16	10181	0.9972	0	0.0028	0.0028
ZA106	7867	ZA175	8316	1	0	0	0
ZA106	7867	ZA190	8338	1	0	0	0
ZA106	7867	ZA194	8415	1	0	0	0
ZA106	7867	ZA213	8430	1	0	0	0
ZA106	7867	ZA216	8874	1	0	0	0
ZA106	7867	ZA233	5473	0.9816	0.0184	0	0.0092
ZA106	7867	ZA24	9087	1	0	0	0

ZA106	7867	ZA241	9200	1	0	0	0
ZA106	7867	ZA252	9231	1	0	0	0
ZA106	7867	ZA253	9260	1	0	0	0
ZA106	7867	ZA256	9434	1	0	0	0
ZA106	7867	ZA263	9458	1	0	0	0
ZA106	7867	ZA272	9564	1	0	0	0
ZA106	7867	ZA316	5581	1	0	0	0
ZA106	7867	ZA34	7594	1	0	0	0
ZA106	7867	ZA4	3836	0.9753	0.0247	0	0.0124
ZA106	7867	ZA411	10110	1	0	0	0
ZA106	7867	ZA413	10141	1	0	0	0
ZA106	7867	ZA5	3823	0.9573	0.0427	0	0.0213
ZA106	7867	ZA64	6734	1	0	0	0
ZA106	7867	ZA68	6764	1	0	0	0
ZA106	7867	ZA72	6768	1	0	0	0
ZA106	7867	ZA76	6782	0.9888	0.0112	0	0.0056
ZA106	7867	ZA78	6772	1	0	0	0
ZA106	7867	ZA86	6807	0.9838	0.0162	0	0.0081
ZA106	7867	ZA89	6816	1	0	0	0
ZA106	7867	ZA92	6827	1	0	0	0
ZA106	7867	ZA95	6924	0.9614	0.0386	0	0.0193
ZA112	7897	ZA116	7929	1	0	0	0
ZA112	7897	ZA128	8158	1	0	0	0
ZA112	7897	ZA134	8165	1	0	0	0
ZA112	7897	ZA140	8174	1	0	0	0
ZA112	7897	ZA142	8239	1	0	0	0
ZA112	7897	ZA150	5645	0.134	0	0	0.0066
ZA112	7897	ZA16	10181	1	0	0	0
ZA112	7897	ZA175	8316	0.9947	0.0022	0.0031	0.0042
ZA112	7897	ZA190	8338	1	0	0	0
ZA112	7897	ZA194	8415	0.9933	0	0.0067	0.0067
ZA112	7897	ZA213	8430	0.9856	0.0119	0.0025	0.0085
ZA112	7897	ZA216	8874	1	0	0	0
ZA112	7897	ZA233	5473	0.996	0.004	0	0.002
ZA112	7897	ZA24	9087	1	0	0	0
ZA112	7897	ZA241	9200	0.9613	0.0387	0	0.0194
ZA112	7897	ZA252	9231	1	0	0	0
ZA112	7897	ZA253	9260	1	0	0	0
ZA112	7897	ZA256	9434	0.9848	0.0152	0	0.0076
ZA112	7897	ZA263	9458	1	0	0	0
ZA112	7897	ZA272	9564	1	0	0	0
ZA112	7897	ZA316	5581	1	0	0	0
ZA112	7897	ZA34	7594	1	0	0	0
ZA112	7897	ZA4	3836	0.9931	0.001	0.0059	0.0064
ZA112	7897	ZA411	10110	1	0	0	0

ZA112	7897	ZA413	10141	1	0	0	0
ZA112	7897	ZA5	3823	0.9952	0	0.0048	0.0048
ZA112	7897	ZA64	6734	1	0	0	0
ZA112	7897	ZA68	6764	0.9957	0.0043	0	0.0021
ZA112	7897	ZA72	6768	0.9838	0.0162	0	0.0081
ZA112	7897	ZA76	6782	1	0	0	0
ZA112	7897	ZA78	6772	0.9953	0.0034	0.0013	0.003
ZA112	7897	ZA86	6807	0.9877	0.0123	0	0.0061
ZA112	7897	ZA89	6816	1	0	0	0
ZA112	7897	ZA92	6827	1	0	0	0
ZA112	7897	ZA95	6924	1	0	0	0
ZA116	7929	ZA128	8158	1	0	0	0
ZA116	7929	ZA134	8165	1	0	0	0
ZA116	7929	ZA140	8174	0.9892	0.0108	0	0.0054
ZA116	7929	ZA142	8239	1	0	0	0
ZA116	7929	ZA150	5645	1	0	0	0
ZA116	7929	ZA16	10181	0.9944	0.0056	0	0.0028
ZA116	7929	ZA175	8316	1	0	0	0
ZA116	7929	ZA190	8338	1	0	0	0
ZA116	7929	ZA194	8415	1	0	0	0
ZA116	7929	ZA213	8430	0.9746	0.0254	0	0.0127
ZA116	7929	ZA216	8874	1	0	0	0
ZA116	7929	ZA233	5473	1	0	0	0
ZA116	7929	ZA24	9087	1	0	0	0
ZA116	7929	ZA241	9200	0.9883	0.0117	0	0.0058
ZA116	7929	ZA252	9231	1	0	0	0
ZA116	7929	ZA253	9260	1	0	0	0
ZA116	7929	ZA256	9434	1	0	0	0
ZA116	7929	ZA263	9458	1	0	0	0
ZA116	7929	ZA272	9564	1	0	0	0
ZA116	7929	ZA316	5581	0.991	0.009	0	0.0045
ZA116	7929	ZA34	7594	1	0	0	0
ZA116	7929	ZA4	3836	1	0	0	0
ZA116	7929	ZA411	10110	1	0	0	0
ZA116	7929	ZA413	10141	1	0	0	0
ZA116	7929	ZA5	3823	0.9886	0.0114	0	0.0057
ZA116	7929	ZA64	6734	1	0	0	0
ZA116	7929	ZA68	6764	0.9881	0.0105	0.0014	0.0067
ZA116	7929	ZA72	6768	1	0	0	0
ZA116	7929	ZA76	6782	0.9948	0.0052	0	0.0026
ZA116	7929	ZA78	6772	1	0	0	0
ZA116	7929	ZA86	6807	0.9817	0.0183	0	0.0092
ZA116	7929	ZA89	6816	1	0	0	0
ZA116	7929	ZA92	6827	1	0	0	0
ZA116	7929	ZA95	6924	1	0	0	0

ZA128	8158	ZA134	8165	1	0	0	0
ZA128	8158	ZA140	8174	0.9927	0.0037	0.0036	0.0055
ZA128	8158	ZA142	8239	1	0	0	0
ZA128	8158	ZA150	5645	1	0	0	0
ZA128	8158	ZA16	10181	0.9891	0	0.0109	0.0109
ZA128	8158	ZA175	8316	1	0	0	0
ZA128	8158	ZA190	8338	0.9871	0.0094	0.0035	0.0082
ZA128	8158	ZA194	8415	0.9869	0	0.0131	0.0131
ZA128	8158	ZA213	8430	0.9677	0.0323	0	0.0162
ZA128	8158	ZA216	8874	0.9882	0.0118	0	0.0059
ZA128	8158	ZA233	5473	0.9922	0	0.0078	0.0078
ZA128	8158	ZA24	9087	0.9844	0.0145	0.0011	0.0084
ZA128	8158	ZA241	9200	0.9651	0.034	0.0008	0.0178
ZA128	8158	ZA252	9231	0.9884	0	0.0116	0.0116
ZA128	8158	ZA253	9260	1	0	0	0
ZA128	8158	ZA256	9434	1	0	0	0
ZA128	8158	ZA263	9458	1	0	0	0
ZA128	8158	ZA272	9564	1	0	0	0
ZA128	8158	ZA316	5581	0.9701	0.0255	0.0044	0.0172
ZA128	8158	ZA34	7594	1	0	0	0
ZA128	8158	ZA4	3836	0.9868	0	0.0132	0.0132
ZA128	8158	ZA411	10110	0.9893	0	0.0107	0.0107
ZA128	8158	ZA413	10141	0.9748	0.0252	0	0.0126
ZA128	8158	ZA5	3823	0.9987	0	0.0013	0.0013
ZA128	8158	ZA64	6734	1	0	0	0
ZA128	8158	ZA68	6764	0.9889	0.0111	0	0.0055
ZA128	8158	ZA72	6768	0.984	0.0137	0.0023	0.0092
ZA128	8158	ZA76	6782	0.9792	0.0193	0.0015	0.0111
ZA128	8158	ZA78	6772	1	0	0	0
ZA128	8158	ZA86	6807	0.976	0.024	0	0.012
ZA128	8158	ZA89	6816	0.9659	0.0341	0	0.017
ZA128	8158	ZA92	6827	1	0	0	0
ZA128	8158	ZA95	6924	0.982	0.018	0	0.009
ZA134	8165	ZA140	8174	0.9911	0.008	0.0009	0.0049
ZA134	8165	ZA142	8239	1	0	0	0
ZA134	8165	ZA150	5645	1	0	0	0
ZA134	8165	ZA16	10181	0.9907	0	0.0093	0.0093
ZA134	8165	ZA175	8316	1	0	0	0
ZA134	8165	ZA190	8338	1	0	0	0
ZA134	8165	ZA194	8415	1	0	0	0
ZA134	8165	ZA213	8430	1	0	0	0
ZA134	8165	ZA216	8874	1	0	0	0
ZA134	8165	ZA233	5473	0.9938	0.0062	0	0.0031
ZA134	8165	ZA24	9087	1	0	0	0
ZA134	8165	ZA241	9200	1	0	0	0

ZA134	8165	ZA252	9231	1	0	0	0
ZA134	8165	ZA253	9260	1	0	0	0
ZA134	8165	ZA256	9434	1	0	0	0
ZA134	8165	ZA263	9458	1	0	0	0
ZA134	8165	ZA272	9564	0.9869	0.009	0.0041	0.0086
ZA134	8165	ZA316	5581	0.9804	0.0088	0.0109	0.0153
ZA134	8165	ZA34	7594	1	0	0	0
ZA134	8165	ZA4	3836	1	0	0	0
ZA134	8165	ZA411	10110	1	0	0	0
ZA134	8165	ZA413	10141	0.9778	0.0222	0	0.0111
ZA134	8165	ZA5	3823	0.9897	0.0078	0.0026	0.0065
ZA134	8165	ZA64	6734	1	0	0	0
ZA134	8165	ZA68	6764	0.9791	0.0209	0	0.0104
ZA134	8165	ZA72	6768	0.9914	0.0086	0	0.0043
ZA134	8165	ZA76	6782	0.9944	0	0.0056	0.0056
ZA134	8165	ZA78	6772	1	0	0	0
ZA134	8165	ZA86	6807	0.9847	0.0153	0	0.0077
ZA134	8165	ZA89	6816	1	0	0	0
ZA134	8165	ZA92	6827	0.9905	0.0095	0	0.0047
ZA134	8165	ZA95	6924	1	0	0	0
ZA140	8174	ZA142	8239	1	0	0	0
ZA140	8174	ZA150	5645	1	0	0	0
ZA140	8174	ZA16	10181	0.9714	0.0286	0	0.0143
ZA140	8174	ZA175	8316	1	0	0	0
ZA140	8174	ZA190	8338	1	0	0	0
ZA140	8174	ZA194	8415	0.9809	0.0191	0	0.0096
ZA140	8174	ZA213	8430	0.9847	0.0148	0.0005	0.0079
ZA140	8174	ZA216	8874	0.9527	0.0473	0	0.0237
ZA140	8174	ZA233	5473	0.9972	0.0005	0.0023	0.0025
ZA140	8174	ZA24	9087	0.9815	0.0185	0	0.0093
ZA140	8174	ZA241	9200	0.9586	0.0414	0	0.0207
ZA140	8174	ZA252	9231	0.9906	0.0052	0.0042	0.0068
ZA140	8174	ZA253	9260	0.983	0.017	0	0.0085
ZA140	8174	ZA256	9434	1	0	0	0
ZA140	8174	ZA263	9458	0.9948	0.0013	0.0039	0.0045
ZA140	8174	ZA272	9564	0.9835	0.0165	0	0.0082
ZA140	8174	ZA316	5581	1	0	0	0
ZA140	8174	ZA34	7594	1	0	0	0
ZA140	8174	ZA4	3836	1	0	0	0
ZA140	8174	ZA411	10110	1	0	0	0
ZA140	8174	ZA413	10141	0.9705	0.0295	0	0.0147
ZA140	8174	ZA5	3823	0.9847	0.0153	0	0.0076
ZA140	8174	ZA64	6734	1	0	0	0
ZA140	8174	ZA68	6764	0.9685	0.0315	0	0.0158
ZA140	8174	ZA72	6768	0.9655	0.0322	0.0024	0.0184

ZA140	8174	ZA76	6782	0.9832	0.0168	0	0.0084
ZA140	8174	ZA78	6772	1	0	0	0
ZA140	8174	ZA86	6807	0.9901	0.0099	0	0.0049
ZA140	8174	ZA89	6816	1	0	0	0
ZA140	8174	ZA92	6827	0.9856	0.0144	0	0.0072
ZA140	8174	ZA95	6924	0.9917	0.0083	0	0.0041
ZA142	8239	ZA150	5645	1	0	0	0
ZA142	8239	ZA16	10181	0.9921	0.0079	0	0.004
ZA142	8239	ZA175	8316	1	0	0	0
ZA142	8239	ZA190	8338	1	0	0	0
ZA142	8239	ZA194	8415	1	0	0	0
ZA142	8239	ZA213	8430	1	0	0	0
ZA142	8239	ZA216	8874	0.9874	0.0126	0	0.0063
ZA142	8239	ZA233	5473	1	0	0	0
ZA142	8239	ZA24	9087	0.9903	0.0028	0.0069	0.0083
ZA142	8239	ZA241	9200	1	0	0	0
ZA142	8239	ZA252	9231	1	0	0	0
ZA142	8239	ZA253	9260	0.9876	0.0072	0.0053	0.0088
ZA142	8239	ZA256	9434	0.9834	0.0166	0	0.0083
ZA142	8239	ZA263	9458	1	0	0	0
ZA142	8239	ZA272	9564	0.993	0.007	0	0.0035
ZA142	8239	ZA316	5581	1	0	0	0
ZA142	8239	ZA34	7594	1	0	0	0
ZA142	8239	ZA4	3836	1	0	0	0
ZA142	8239	ZA411	10110	1	0	0	0
ZA142	8239	ZA413	10141	1	0	0	0
ZA142	8239	ZA5	3823	1	0	0	0
ZA142	8239	ZA64	6734	1	0	0	0
ZA142	8239	ZA68	6764	0.9933	0.0067	0	0.0033
ZA142	8239	ZA72	6768	1	0	0	0
ZA142	8239	ZA76	6782	0.9953	0	0.0047	0.0047
ZA142	8239	ZA78	6772	1	0	0	0
ZA142	8239	ZA86	6807	0.9885	0.0111	0.0004	0.0059
ZA142	8239	ZA89	6816	1	0	0	0
ZA142	8239	ZA92	6827	1	0	0	0
ZA142	8239	ZA92	6924	1	0	0	0
ZA150	5645	ZA16	10181	1	0	0	0
ZA150	5645	ZA175	8316	1	0	0	0
ZA150	5645	ZA190	8338	1	0	0	0
ZA150	5645	ZA194	8415	1	0	0	0
ZA150	5645	ZA213	8430	1	0	0	0
ZA150	5645	ZA216	8874	1	0	0	0
ZA150	5645	ZA233	5473	1	0	0	0
ZA150	5645	ZA24	9087	1	0	0	0
ZA150	5645	ZA241	9200	1	0	0	0

ZA150	5645	ZA252	9231	1	0	0	0
ZA150	5645	ZA253	9260	1	0	0	0
ZA150	5645	ZA256	9434	1	0	0	0
ZA150	5645	ZA263	9458	1	0	0	0
ZA150	5645	ZA272	9564	1	0	0	0
ZA150	5645	ZA316	5581	1	0	0	0
ZA150	5645	ZA34	7594	1	0	0	0
ZA150	5645	ZA4	3836	1	0	0	0
ZA150	5645	ZA411	10110	1	0	0	0
ZA150	5645	ZA413	10141	1	0	0	0
ZA150	5645	ZA5	3823	1	0	0	0
ZA150	5645	ZA64	6734	1	0	0	0
ZA150	5645	ZA68	6764	0.987	0.013	0	0.0065
ZA150	5645	ZA72	6768	1	0	0	0
ZA150	5645	ZA76	6782	1	0	0	0
ZA150	5645	ZA78	6772	1	0	0	0
ZA150	5645	ZA86	6807	1	0	0	0
ZA150	5645	ZA89	6816	1	0	0	0
ZA150	5645	ZA92	6827	1	0	0	0
ZA150	5645	ZA95	6924	1	0	0	0
ZA16	10181	ZA175	8316	1	0	0	0
ZA16	10181	ZA190	8338	1	0	0	0
ZA16	10181	ZA194	8415	0.9932	0	0.0068	0.0068
ZA16	10181	ZA213	8430	0.9927	0.0073	0	0.0036
ZA16	10181	ZA216	8874	0.957	0.043	0	0.0215
ZA16	10181	ZA233	5473	0.9992	0	0.0008	0.0008
ZA16	10181	ZA24	9087	1	0	0	0
ZA16	10181	ZA241	9200	0.9918	0.0082	0	0.0041
ZA16	10181	ZA252	9231	0.9971	0	0.0029	0.0029
ZA16	10181	ZA253	9260	1	0	0	0
ZA16	10181	ZA256	9434	1	0	0	0
ZA16	10181	ZA263	9458	0.9968	0.0001	0.003	0.0031
ZA16	10181	ZA272	9564	1	0	0	0
ZA16	10181	ZA316	5581	1	0	0	0
ZA16	10181	ZA34	7594	1	0	0	0
ZA16	10181	ZA4	3836	0.9957	0	0.0043	0.0043
ZA16	10181	ZA411	10110	1	0	0	0
ZA16	10181	ZA413	10141	1	0	0	0
ZA16	10181	ZA5	3823	1	0	0	0
ZA16	10181	ZA64	6734	1	0	0	0
ZA16	10181	ZA68	6764	0.9869	0.0131	0	0.0066
ZA16	10181	ZA72	6768	0.9685	0.0315	0	0.0157
ZA16	10181	ZA76	6782	0.9863	0.0114	0.0023	0.008
ZA16	10181	ZA78	6772	1	0	0	0
ZA16	10181	ZA86	6807	0.9677	0.0323	0	0.0162

ZA16	10181	ZA89	6816	0.9826	0.0174	0	0.0087
ZA16	10181	ZA92	6827	1	0	0	0
ZA16	10181	ZA95	6924	0.9893	0.0107	0	0.0053
ZA175	8316	ZA190	8338	1	0	0	0
ZA175	8316	ZA194	8415	1	0	0	0
ZA175	8316	ZA213	8430	1	0	0	0
ZA175	8316	ZA216	8874	1	0	0	0
ZA175	8316	ZA233	5473	0.9948	0	0.0052	0.0052
ZA175	8316	ZA24	9087	0.9889	0.0111	0	0.0056
ZA175	8316	ZA241	9200	0.9905	0.0054	0.0041	0.0068
ZA175	8316	ZA252	9231	1	0	0	0
ZA175	8316	ZA253	9260	0.9901	0.0099	0	0.0049
ZA175	8316	ZA256	9434	1	0	0	0
ZA175	8316	ZA263	9458	1	0	0	0
ZA175	8316	ZA272	9564	1	0	0	0
ZA175	8316	ZA316	5581	1	0	0	0
ZA175	8316	ZA34	7594	1	0	0	0
ZA175	8316	ZA4	3836	1	0	0	0
ZA175	8316	ZA411	10110	1	0	0	0
ZA175	8316	ZA413	10141	1	0	0	0
ZA175	8316	ZA5	3823	1	0	0	0
ZA175	8316	ZA64	6734	1	0	0	0
ZA175	8316	ZA68	6764	0.981	0.019	0	0.0095
ZA175	8316	ZA72	6768	1	0	0	0
ZA175	8316	ZA76	6782	0.9971	0	0.0029	0.0029
ZA175	8316	ZA78	6772	1	0	0	0
ZA175	8316	ZA86	6807	1	0	0	0
ZA175	8316	ZA89	6816	0.9921	0	0.0079	0.0079
ZA175	8316	ZA92	6827	1	0	0	0
ZA175	8316	ZA95	6924	1	0	0	0
ZA190	8338	ZA194	8415	1	0	0	0
ZA190	8338	ZA213	8430	1	0	0	0
ZA190	8338	ZA216	8874	1	0	0	0
ZA190	8338	ZA233	5473	0.9952	0	0.0048	0.0048
ZA190	8338	ZA24	9087	1	0	0	0
ZA190	8338	ZA241	9200	0.9951	0.0021	0.0028	0.0039
ZA190	8338	ZA252	9231	1	0	0	0
ZA190	8338	ZA253	9260	1	0	0	0
ZA190	8338	ZA256	9434	1	0	0	0
ZA190	8338	ZA263	9458	1	0	0	0
ZA190	8338	ZA272	9564	1	0	0	0
ZA190	8338	ZA316	5581	1	0	0	0
ZA190	8338	ZA34	7594	1	0	0	0
ZA190	8338	ZA4	3836	1	0	0	0
ZA190	8338	ZA411	10110	1	0	0	0

ZA190	8338	ZA413	10141	1	0	0	0
ZA190	8338	ZA5	3823	1	0	0	0
ZA190	8338	ZA64	6734	1	0	0	0
ZA190	8338	ZA68	6764	0.9944	0.0056	0	0.0028
ZA190	8338	ZA72	6768	0.9929	0.0056	0.0015	0.0043
ZA190	8338	ZA76	6782	1	0	0	0
ZA190	8338	ZA78	6772	1	0	0	0
ZA190	8338	ZA86	6807	1	0	0	0
ZA190	8338	ZA89	6816	1	0	0	0
ZA190	8338	ZA92	6827	1	0	0	0
ZA190	8338	ZA95	6924	1	0	0	0
ZA194	8415	ZA213	8430	0.9783	0.0217	0	0.0109
ZA194	8415	ZA216	8874	1	0	0	0
ZA194	8415	ZA233	5473	1	0	0	0
ZA194	8415	ZA24	9087	0.9994	0.0005	0.0001	0.0004
ZA194	8415	ZA241	9200	0.9876	0.0124	0	0.0062
ZA194	8415	ZA252	9231	1	0	0	0
ZA194	8415	ZA253	9260	0.9894	0.0093	0.0014	0.006
ZA194	8415	ZA256	9434	0.9942	0.0056	0.0002	0.003
ZA194	8415	ZA263	9458	1	0	0	0
ZA194	8415	ZA272	9564	1	0	0	0
ZA194	8415	ZA316	5581	1	0	0	0
ZA194	8415	ZA34	7594	1	0	0	0
ZA194	8415	ZA4	3836	0.9944	0.0056	0	0.0028
ZA194	8415	ZA411	10110	1	0	0	0
ZA194	8415	ZA413	10141	1	0	0	0
ZA194	8415	ZA5	3823	1	0	0	0
ZA194	8415	ZA64	6734	1	0	0	0
ZA194	8415	ZA68	6764	0.9783	0.0217	0	0.0108
ZA194	8415	ZA72	6768	1	0	0	0
ZA194	8415	ZA76	6782	0.9847	0.0153	0	0.0076
ZA194	8415	ZA78	6772	1	0	0	0
ZA194	8415	ZA86	6807	0.9811	0.0189	0	0.0094
ZA194	8415	ZA89	6816	0.9619	0.0381	0	0.019
ZA194	8415	ZA92	6827	0.9971	0.0029	0	0.0014
ZA194	8415	ZA95	6924	0.984	0.016	0	0.008
ZA213	8430	ZA216	8874	0.981	0.019	0	0.0095
ZA213	8430	ZA233	5473	0.9868	0.0132	0	0.0066
ZA213	8430	ZA24	9087	0.9959	0.0024	0.0017	0.0029
ZA213	8430	ZA241	9200	0.9896	0.0081	0.0024	0.0064
ZA213	8430	ZA252	9231	1	0	0	0
ZA213	8430	ZA253	9260	1	0	0	0
ZA213	8430	ZA256	9434	1	0	0	0
ZA213	8430	ZA263	9458	0.9825	0.0152	0.0023	0.0099
ZA213	8430	ZA272	9564	0.9862	0.0138	0	0.0069

ZA213	8430	ZA316	5581	1	0	0	0
ZA213	8430	ZA34	7594	1	0	0	0
ZA213	8430	ZA4	3836	0.9928	0	0.0072	0.0072
ZA213	8430	ZA411	10110	1	0	0	0
ZA213	8430	ZA413	10141	0.9921	0	0.0079	0.0079
ZA213	8430	ZA5	3823	0.9768	0.0232	0	0.0116
ZA213	8430	ZA64	6734	1	0	0	0
ZA213	8430	ZA68	6764	0.9524	0.0476	0	0.0238
ZA213	8430	ZA72	6768	0.98	0.02	0	0.01
ZA213	8430	ZA76	6782	0.9884	0.0108	0.0008	0.0062
ZA213	8430	ZA78	6772	1	0	0	0
ZA213	8430	ZA86	6807	0.9668	0.0332	0	0.0166
ZA213	8430	ZA89	6816	0.9866	0.0134	0	0.0067
ZA213	8430	ZA92	6827	0.9629	0.0371	0	0.0186
ZA213	8430	ZA95	6924	1	0	0	0
ZA216	8874	ZA233	5473	0.9602	0.0398	0	0.0199
ZA216	8874	ZA24	9087	0.9741	0.0259	0	0.013
ZA216	8874	ZA241	9200	0.9948	0.0052	0	0.0026
ZA216	8874	ZA252	9231	0.9878	0.0122	0	0.0061
ZA216	8874	ZA253	9260	1	0	0	0
ZA216	8874	ZA256	9434	0.9864	0.0136	0	0.0068
ZA216	8874	ZA263	9458	1	0	0	0
ZA216	8874	ZA272	9564	1	0	0	0
ZA216	8874	ZA316	5581	1	0	0	0
ZA216	8874	ZA34	7594	1	0	0	0
ZA216	8874	ZA4	3836	0.9935	0.0065	0	0.0033
ZA216	8874	ZA411	10110	1	0	0	0
ZA216	8874	ZA413	10141	0.9861	0.0139	0	0.007
ZA216	8874	ZA5	3823	0.9712	0.0288	0	0.0144
ZA216	8874	ZA64	6734	1	0	0	0
ZA216	8874	ZA68	6764	0.9585	0.0415	0	0.0207
ZA216	8874	ZA72	6768	0.956	0.044	0	0.022
ZA216	8874	ZA76	6782	0.983	0.017	0	0.0085
ZA216	8874	ZA78	6772	0.9871	0.0129	0	0.0065
ZA216	8874	ZA86	6807	1	0	0	0
ZA216	8874	ZA89	6816	0.966	0.034	0	0.017
ZA216	8874	ZA92	6827	0.9781	0.0219	0	0.0109
ZA216	8874	ZA95	6924	0.9587	0.0413	0	0.0207
ZA233	5473	ZA24	9087	0.9409	0.0591	0	0.0296
ZA233	5473	ZA241	9200	0.9722	0.0278	0	0.0139
ZA233	5473	ZA252	9231	1	0	0	0
ZA233	5473	ZA253	9260	1	0	0	0
ZA233	5473	ZA256	9434	1	0	0	0
ZA233	5473	ZA263	9458	1	0	0	0
ZA233	5473	ZA272	9564	1	0	0	0

ZA233	5473	ZA316	5581	1	0	0	0
ZA233	5473	ZA34	7594	1	0	0	0
ZA233	5473	ZA4	3836	0.9975	0.0021	0.0005	0.0015
ZA233	5473	ZA411	10110	1	0	0	0
ZA233	5473	ZA413	10141	0.97	0.03	0	0.015
ZA233	5473	ZA5	3823	1	0	0	0
ZA233	5473	ZA64	6734	1	0	0	0
ZA233	5473	ZA68	6764	0.9729	0.0271	0	0.0136
ZA233	5473	ZA72	6768	0.9789	0.0182	0.0029	0.012
ZA233	5473	ZA76	6782	0.9733	0.0267	0	0.0134
ZA233	5473	ZA78	6772	1	0	0	0
ZA233	5473	ZA86	6807	1	0	0	0
ZA233	5473	ZA89	6816	1	0	0	0
ZA233	5473	ZA92	6827	0.9628	0.0372	0	0.0186
ZA233	5473	ZA95	6924	0.9705	0.0295	0	0.0147
ZA24	9087	ZA241	9200	0.9634	0.0366	0	0.0183
ZA24	9087	ZA252	9231	0.994	0	0.006	0.006
ZA24	9087	ZA253	9260	0.9987	0.0013	0	0.0006
ZA24	9087	ZA256	9434	0.997	0	0.003	0.003
ZA24	9087	ZA263	9458	0.9962	0.0038	0	0.0019
ZA24	9087	ZA272	9564	0.9824	0.0176	0	0.0088
ZA24	9087	ZA316	5581	1	0	0	0
ZA24	9087	ZA34	7594	1	0	0	0
ZA24	9087	ZA4	3836	0.9883	0.0099	0.0018	0.0068
ZA24	9087	ZA411	10110	1	0	0	0
ZA24	9087	ZA413	10141	0.9908	0.0064	0.0028	0.006
ZA24	9087	ZA5	3823	0.9722	0.0278	0	0.0139
ZA24	9087	ZA64	6734	1	0	0	0
ZA24	9087	ZA68	6764	1	0	0	0
ZA24	9087	ZA72	6768	1	0	0	0
ZA24	9087	ZA76	6782	0.9643	0.0357	0	0.0178
ZA24	9087	ZA78	6772	0.9872	0.0128	0	0.0064
ZA24	9087	ZA86	6807	0.9772	0.0228	0	0.0114
ZA24	9087	ZA89	6816	0.985	0.015	0	0.0075
ZA24	9087	ZA92	6827	0.9969	0.0019	0.0012	0.0021
ZA24	9087	ZA95	6924	0.9712	0.0288	0	0.0144
ZA241	9200	ZA252	9231	0.9937	0.0063	0	0.0031
ZA241	9200	ZA253	9260	0.9838	0.0162	0	0.0081
ZA241	9200	ZA256	9434	0.9785	0.0215	0	0.0107
ZA241	9200	ZA263	9458	1	0	0	0
ZA241	9200	ZA272	9564	1	0	0	0
ZA241	9200	ZA316	5581	1	0	0	0
ZA241	9200	ZA34	7594	1	0	0	0
ZA241	9200	ZA4	3836	0.9842	0.0158	0	0.0079
ZA241	9200	ZA411	10110	0.9938	0.0062	0	0.0031

ZA241	9200	ZA413	10141	0.9797	0.0198	0.0004	0.0104
ZA241	9200	ZA5	3823	0.958	0.042	0	0.021
ZA241	9200	ZA64	6734	1	0	0	0
ZA241	9200	ZA68	6764	1	0	0	0
ZA241	9200	ZA72	6768	0.9573	0.0427	0	0.0214
ZA241	9200	ZA76	6782	0.9907	0.0077	0.0016	0.0055
ZA241	9200	ZA78	6772	1	0	0	0
ZA241	9200	ZA86	6807	0.9908	0.0092	0	0.0046
ZA241	9200	ZA89	6816	0.9782	0.0218	0	0.0109
ZA241	9200	ZA92	6827	0.9756	0.0244	0	0.0122
ZA241	9200	ZA95	6924	1	0	0	0
ZA252	9231	ZA253	9260	1	0	0	0
ZA252	9231	ZA256	9434	1	0	0	0
ZA252	9231	ZA263	9458	1	0	0	0
ZA252	9231	ZA272	9564	1	0	0	0
ZA252	9231	ZA316	5581	0.9958	0.001	0.0031	0.0037
ZA252	9231	ZA34	7594	1	0	0	0
ZA252	9231	ZA4	3836	0.9839	0.0076	0.0085	0.0123
ZA252	9231	ZA411	10110	1	0	0	0
ZA252	9231	ZA413	10141	1	0	0	0
ZA252	9231	ZA5	3823	0.9965	0.0035	0	0.0017
ZA252	9231	ZA64	6734	1	0	0	0
ZA252	9231	ZA68	6764	0.9767	0.0233	0	0.0117
ZA252	9231	ZA72	6768	0.9931	0.0015	0.0054	0.0061
ZA252	9231	ZA76	6782	0.991	0	0.009	0.009
ZA252	9231	ZA78	6772	1	0	0	0
ZA252	9231	ZA86	6807	0.9759	0.0241	0	0.012
ZA252	9231	ZA89	6816	1	0	0	0
ZA252	9231	ZA92	6827	1	0	0	0
ZA252	9231	ZA95	6924	1	0	0	0
ZA253	9260	ZA256	9434	0.9893	0.0107	0	0.0053
ZA253	9260	ZA263	9458	1	0	0	0
ZA253	9260	ZA272	9564	0.991	0.009	0	0.0045
ZA253	9260	ZA316	5581	0.9936	0.0053	0.0011	0.0037
ZA253	9260	ZA34	7594	1	0	0	0
ZA253	9260	ZA4	3836	1	0	0	0
ZA253	9260	ZA411	10110	1	0	0	0
ZA253	9260	ZA413	10141	1	0	0	0
ZA253	9260	ZA5	3823	0.9561	0.0439	0	0.022
ZA253	9260	ZA64	6734	1	0	0	0
ZA253	9260	ZA68	6764	0.9703	0.0297	0	0.0149
ZA253	9260	ZA72	6768	1	0	0	0
ZA253	9260	ZA76	6782	0.9964	0.0023	0.0013	0.0024
ZA253	9260	ZA78	6772	1	0	0	0
ZA253	9260	ZA86	6807	1	0	0	0

ZA253	9260	ZA89	6816	1	0	0	0
ZA253	9260	ZA92	6827	1	0	0	0
ZA253	9260	ZA95	6924	0.9681	0.0319	0	0.0159
ZA256	9434	ZA263	9458	1	0	0	0
ZA256	9434	ZA272	9564	1	0	0	0
ZA256	9434	ZA316	5581	1	0	0	0
ZA256	9434	ZA34	7594	1	0	0	0
ZA256	9434	ZA4	3836	1	0	0	0
ZA256	9434	ZA411	10110	1	0	0	0
ZA256	9434	ZA413	10141	0.9984	0.0016	0	0.0008
ZA256	9434	ZA5	3823	1	0	0	0
ZA256	9434	ZA64	6734	1	0	0	0
ZA256	9434	ZA68	6764	1	0	0	0
ZA256	9434	ZA72	6768	1	0	0	0
ZA256	9434	ZA76	6782	0.9901	0.0099	0	0.005
ZA256	9434	ZA78	6772	1	0	0	0
ZA256	9434	ZA86	6807	0.9498	0.0502	0	0.0251
ZA256	9434	ZA89	6816	0.978	0.022	0	0.011
ZA256	9434	ZA92	6827	0.9963	0.0037	0	0.0019
ZA256	9434	ZA95	6924	1	0	0	0
ZA263	9458	ZA272	9564	1	0	0	0
ZA263	9458	ZA316	5581	1	0	0	0
ZA263	9458	ZA34	7594	1	0	0	0
ZA263	9458	ZA4	3836	1	0	0	0
ZA263	9458	ZA411	10110	1	0	0	0
ZA263	9458	ZA413	10141	1	0	0	0
ZA263	9458	ZA5	3823	0.9814	0.0186	0	0.0093
ZA263	9458	ZA64	6734	1	0	0	0
ZA263	9458	ZA68	6764	1	0	0	0
ZA263	9458	ZA72	6768	0.9611	0.0389	0	0.0195
ZA263	9458	ZA76	6782	1	0	0	0
ZA263	9458	ZA78	6772	1	0	0	0
ZA263	9458	ZA86	6807	1	0	0	0
ZA263	9458	ZA89	6816	1	0	0	0
ZA263	9458	ZA92	6827	1	0	0	0
ZA263	9458	ZA95	6924	1	0	0	0
ZA272	9564	ZA316	5581	1	0	0	0
ZA272	9564	ZA34	7594	1	0	0	0
ZA272	9564	ZA4	3836	1	0	0	0
ZA272	9564	ZA411	10110	1	0	0	0
ZA272	9564	ZA413	10141	1	0	0	0
ZA272	9564	ZA5	3823	0.9933	0.0067	0	0.0033
ZA272	9564	ZA64	6734	1	0	0	0
ZA272	9564	ZA68	6764	0.9756	0.0244	0	0.0122
ZA272	9564	ZA72	6768	1	0	0	0

ZA272	9564	ZA76	6782	1	0	0	0
ZA272	9564	ZA78	6772	1	0	0	0
ZA272	9564	ZA86	6807	1	0	0	0
ZA272	9564	ZA89	6816	1	0	0	0
ZA272	9564	ZA92	6827	0.9994	0	0.0006	0.0006
ZA272	9564	ZA95	6924	1	0	0	0
ZA316	5581	ZA34	7594	1	0	0	0
ZA316	5581	ZA4	3836	0.9884	0.0116	0	0.0058
ZA316	5581	ZA411	10110	1	0	0	0
ZA316	5581	ZA413	10141	1	0	0	0
ZA316	5581	ZA5	3823	1	0	0	0
ZA316	5581	ZA64	6734	1	0	0	0
ZA316	5581	ZA68	6764	0.9975	0.0025	0	0.0013
ZA316	5581	ZA72	6768	0.986	0.0117	0.0022	0.0081
ZA316	5581	ZA76	6782	0.9742	0.0258	0	0.0129
ZA316	5581	ZA78	6772	1	0	0	0
ZA316	5581	ZA86	6807	1	0	0	0
ZA316	5581	ZA89	6816	0.9902	0.004	0.0058	0.0078
ZA316	5581	ZA92	6827	0.9967	0.0033	0	0.0016
ZA316	5581	ZA95	6924	0.9978	0.0015	0.0007	0.0014
ZA34	7594	ZA4	3836	1	0	0	0
ZA34	7594	ZA411	10110	1	0	0	0
ZA34	7594	ZA413	10141	1	0	0	0
ZA34	7594	ZA5	3823	1	0	0	0
ZA34	7594	ZA64	6734	1	0	0	0
ZA34	7594	ZA68	6764	1	0	0	0
ZA34	7594	ZA72	6768	1	0	0	0
ZA34	7594	ZA76	6782	1	0	0	0
ZA34	7594	ZA78	6772	1	0	0	0
ZA34	7594	ZA86	6807	1	0	0	0
ZA34	7594	ZA89	6816	1	0	0	0
ZA34	7594	ZA92	6827	1	0	0	0
ZA34	7594	ZA95	6924	1	0	0	0
ZA4	3836	ZA411	10110	1	0	0	0
ZA4	3836	ZA413	10141	0.9797	0.0203	0	0.0102
ZA4	3836	ZA5	3823	1	0	0	0
ZA4	3836	ZA64	6734	1	0	0	0
ZA4	3836	ZA68	6764	0.9662	0.0338	0	0.0169
ZA4	3836	ZA72	6768	0.9712	0.0206	0.0082	0.0185
ZA4	3836	ZA76	6782	0.9854	0.0146	0	0.0073
ZA4	3836	ZA78	6772	1	0	0	0
ZA4	3836	ZA86	6807	0.9907	0.0085	0.0009	0.0051
ZA4	3836	ZA89	6816	0.9888	0.0034	0.0078	0.0095
ZA4	3836	ZA92	6827	0.9981	0.0016	0.0003	0.0011
ZA4	3836	ZA95	6924	1	0	0	0

ZA411	10110	ZA413	10141	1	0	0	0
ZA411	10110	ZA5	3823	1	0	0	0
ZA411	10110	ZA64	6734	1	0	0	0
ZA411	10110	ZA68	6764	1	0	0	0
ZA411	10110	ZA72	6768	1	0	0	0
ZA411	10110	ZA76	6782	1	0	0	0
ZA411	10110	ZA78	6772	1	0	0	0
ZA411	10110	ZA86	6807	1	0	0	0
ZA411	10110	ZA89	6816	1	0	0	0
ZA411	10110	ZA92	6827	1	0	0	0
ZA411	10110	ZA95	6924	0.982	0.018	0	0.009
ZA413	10141	ZA5	3823	0.9887	0.0113	0	0.0057
ZA413	10141	ZA64	6734	1	0	0	0
ZA413	10141	ZA68	6764	0.9869	0.0131	0	0.0066
ZA413	10141	ZA72	6768	0.9844	0.0049	0.0107	0.0132
ZA413	10141	ZA76	6782	0.9599	0.0371	0.0031	0.0216
ZA413	10141	ZA78	6772	0.9772	0.0211	0.0017	0.0122
ZA413	10141	ZA86	6807	0.9823	0.0156	0.0021	0.0099
ZA413	10141	ZA89	6816	0.9689	0.0311	0	0.0155
ZA413	10141	ZA92	6827	1	0	0	0
ZA413	10141	ZA95	6924	0.9855	0.0145	0	0.0072
ZA5	3823	ZA64	6734	1	0	0	0
ZA5	3823	ZA68	6764	0.9639	0.0361	0	0.0181
ZA5	3823	ZA72	6768	0.9886	0.0114	0	0.0057
ZA5	3823	ZA76	6782	0.9771	0.0229	0	0.0115
ZA5	3823	ZA78	6772	0.9917	0.0083	0	0.0041
ZA5	3823	ZA86	6807	1	0	0	0
ZA5	3823	ZA89	6816	0.9818	0.0182	0	0.0091
ZA5	3823	ZA92	6827	0.9933	0.0067	0	0.0034
ZA5	3823	ZA95	6924	0.9881	0.0119	0	0.0059
ZA64	6734	ZA68	6764	1	0	0	0
ZA64	6734	ZA72	6768	1	0	0	0
ZA64	6734	ZA76	6782	0.0175	0	0	0.0351
ZA64	6734	ZA78	6772	1	0	0	0
ZA64	6734	ZA86	6807	1	0	0	0
ZA64	6734	ZA89	6816	1	0	0	0
ZA64	6734	ZA92	6827	1	0	0	0
ZA64	6734	ZA95	6924	0.0175	0	0	0.0351
ZA68	6764	ZA72	6768	1	0	0	0
ZA68	6764	ZA76	6782	0.9837	0.0163	0	0.0081
ZA68	6764	ZA78	6772	1	0	0	0
ZA68	6764	ZA86	6807	0.9809	0.0191	0	0.0095
ZA68	6764	ZA89	6816	1	0	0	0
ZA68	6764	ZA92	6827	0.9662	0.0338	0	0.0169
ZA68	6764	ZA95	6924	0.9688	0.0312	0	0.0156

ZA72	6768	ZA76	6782	0.971	0.0253	0.0037	0.0163
ZA72	6768	ZA78	6772	0.9738	0.0262	0	0.0131
ZA72	6768	ZA86	6807	0.9792	0.0208	0	0.0104
ZA72	6768	ZA89	6816	0.9939	0.0031	0.003	0.0046
ZA72	6768	ZA92	6827	1	0	0	0
ZA72	6768	ZA95	6924	0.9634	0.0366	0	0.0183
ZA76	6782	ZA78	6772	1	0	0	0
ZA76	6782	ZA86	6807	0.9681	0.0319	0	0.016
ZA76	6782	ZA89	6816	0.9972	0	0.0028	0.0028
ZA76	6782	ZA92	6827	1	0	0	0
ZA76	6782	ZA95	6924	0.9801	0.0199	0	0.01
ZA78	6772	ZA86	6807	0.9932	0.0068	0	0.0034
ZA78	6772	ZA89	6816	1	0	0	0
ZA78	6772	ZA92	6827	1	0	0	0
ZA78	6772	ZA95	6924	0.9806	0.0194	0	0.0097
ZA86	6807	ZA89	6816	0.9868	0.0132	0	0.0066
ZA86	6807	ZA92	6827	1	0	0	0
ZA86	6807	ZA95	6924	0.9742	0.0258	0	0.0129
ZA89	6816	ZA92	6827	1	0	0	0
ZA89	6816	ZA95	6924	1	0	0	0
ZA92	6827	ZA950	6924	0.9915	0.0085	0	0.0043

FID – Family ID; IID – Individual ID; Z0 – probability that individuals at a specific marker will share no alleles; Z1 – probability that individuals at a specific marker will share 1 allele; Z2 – probability that individuals at a specific marker will share 2 alleles.

Appendix IV – Novel and rare non-synonymous variants identified in four WES PD probands using the hypothesis based approach

Gene	SNP	Frequency	rs number	SIFT	PolyPhen2	MutationTaster
ANAPC1	Q451H	0	rs79100806	T	B	0.360565
AQP7	Y115H	0	rs74668961	D	B	0.999905
AQP7	E202D	0	rs114937176	T	B	0.064028
CASP1	G85E	0	rs2509649	T	B	0.097257
CCDC144NL	X222Q	0	rs4605228	NA	NA	0
CLIP1	L271F	0	rs79909185	D	P	0.974668
GXYLT1	Y233X	0	rs77044712	NA	NA	1
GXYLT1	Y234C	0	rs79044728	D	D	0.998509
HBD	S87A	0	-	T	NA	0.162278
HNRNPCL1,LOC649330	D255Y	0	rs141207681	D	B	0.727625
HNRNPCL1,LOC649330	S286R	0	rs148930640	D	B	0.410105
HNRNPCL1,LOC649330	A248V	0	rs149302457	D	B	0.094496
HNRNPCL1,LOC649330	T287A	0	rs149796618	T	B	0.090357
HNRNPCL1,LOC649330	D265N	0	rs2359486	T	B	0.001468
HNRNPCL1,LOC649330	P253L	0	rs150590256	T	B	3.44E-4
HNRNPCL1,LOC649330	E245D	0	rs146075045	T	B	1.59E-4
HNRNPCL1,LOC649330	V258D	0	rs2076063	T	B	5.1E-5
IL28A	F137L	0	-	NA	B	1.9E-5
IL32	D172G	0	rs2981599	D	P	0.981577
KATNAL2	S301C	0	rs76539063	D	P	0.9725871
KIR3DL1	S239N	0	-	T	NA	3.9E-5
MLL3	S772L	0	rs4024453	D	P	0.453495
MLL3	T316S	0	rs10454320	T	P	0.436638
NOTCH2NL	T158I	0	rs75586173	T	D	0.997294
PABPC1	R374C	0	-	D	D	0.996115
PABPC1	E372G	0	-	D	D	0.988773
PABPC3	A181T	0	rs112107735	T	B	0.984775
PABPC3	I195V	0	rs76861216	T	B	0.972619
PABPC3	P191T	0	rs76264750	T	B	0.004658
PABPC3	Q172R	0	rs75475407	T	B	6.0E-5
PLIN4	A883T	0	rs80238130	-	-	-
PPIAL4G	F112L	0	rs6604511	D	P	0.702685
PRAMEF4	G94E	0	-	T	NA	0.004287
PRAMEF4	L19V	0	-	D	NA	9.8E-4

PRB2	R357Q	0	-	NA	NA	0.001769
PRSS1	D218Y	0	-	T	B	0.356489
PRSS1	V213I	0	-	T	B	0.284652
PRSS3	T60S	0	-	T	B	0.00535
PRSS3	K152Q	0	-	T	B	0.003122
PRSS3	S175N	0	-	T	B	1.34E-4
PRSS3	K72E	0	rs151192741	T	B	2.3E-5
PRSS3	A145T	0	rs855581	D	P	0.999736
TIMM23	N9D	0	rs4935252	D	P	0.999742
TPSD1	H143R	0	rs72775466	T	B	0.102904
AQP7	A100T	0.50/0.50 (n = 2)	rs77962308	T	B	0.008256
BCLAF1	T837N	0.50/0.50 (n = 2)	rs62431284	D	P	5.0E-6
C22orf42	M120I	0.958/0.042 (n = 3420)	rs144597334	D	B	2.91E-4
CCDC144NL	S217Y	1.00/0.00 (n = 120)	rs2318592	D	NA	0.00422
CDC27	W644R	0.50/0.50 (n = 2)	rs74348171	D	NA	0.999988
CDC27	Y641C	0.50/0.50 (n = 2)	rs62075618	D	NA	0.999985
CDC27	H615Q	0.50/0.50 (n = 2)	rs75661039	D	NA	0.999889
CDC27	H615R	0.50/0.50 (n = 2)	rs76926116	D	NA	0.999889
CDC27	R631X	0.993/0.007 (n = 129)	rs77685276	NA	NA	1
CNTN5	S23A	0.50/0.50 (n = 2)	rs10790978	-	-	-
CNTN5	L70R	0.50/0.50 (n = 2)	rs7125822	-	-	-
GPRIN2	W91R	0.50/0.50 (n = 2)	rs3127820	D	B	4.0E-6
GXYLT1	R227L	0.50/0.50 (n = 2)	rs76555438	D	D	0.998482
GXYLT1	R230S	0.50/0.50 (n = 2)	rs74583427	D	D	0.995639
GXYLT1	E218G	0.50/0.50 (n = 2)	-	D	P	0.994118
GXYLT1	Y233N	0.50/0.50 (n = 2)	-	D	D	0.993666
GXYLT1	E218K	0.50/0.50 (n = 2)	rs77582546	D	D	0.992851
GXYLT1	N226S	0.50/0.50 (n = 2)	rs78536827	T	P	0.971514
KCNJ12,KCNJ18	R118Q	0.50/0.50 (n = 2)	rs1657740	T	B	0.021368
MAP2K3	Q73X	0.50/0.50 (n = 2)	rs55796947	NA	NA	1
MLL3	L291F	1.00/0.00 (n=120)	rs56850341	T	D	0.885477
MYO5B	V1703A	0.981/0.019 (n = 259)	rs138128932	NA	B	0.145661
PABPC3	R469Q	0.50/0.50 (n = 2)	rs140135080	T	B	0.996872
PABPC3	K444M	0.50/0.50 (n = 2)	rs75484271	T	B	0.002961
PABPC3	G451A	0.919/0.081 (n = 333)	rs113617207	T	B	7.6E-5

PABPC3	I448T	0.989/0.011 (n = 375)	rs112901832	T	B	0.92257
PABPC3	S446G	0.998/0.002 (n = 325)	rs78778235	T	B	7.6E-4
PPIAL4G	A101V	0.50/0.50 (n = 2)	rs2490183	D	B	0.99409
PRAMEF1	R213H	0.50/0.50 (n = 2)	rs1063769	T	B	1.6E-4
PRAMEF1	R265G	0.50/0.50 (n = 4)	rs74937070	T	P	0.003856
PRAMEF1	A257T	0.996/0.004 (n = 939)	rs1063779	T	B	0.001722
PSG9	T410I	1.00/0.00 (n = 2)	rs1063001	D	B	0.004537
PSG9	H397R	1.00/0.00 (n = 2)	rs2072285	T	B	1.25E-4
TIMM23	G23E	0.50/0.50 (n = 2)	rs373071373	D	P	0.85697
YY1AP1	Q424R	0.998/0.002 (n = 1315)	rs113197997	T	B	0.294756

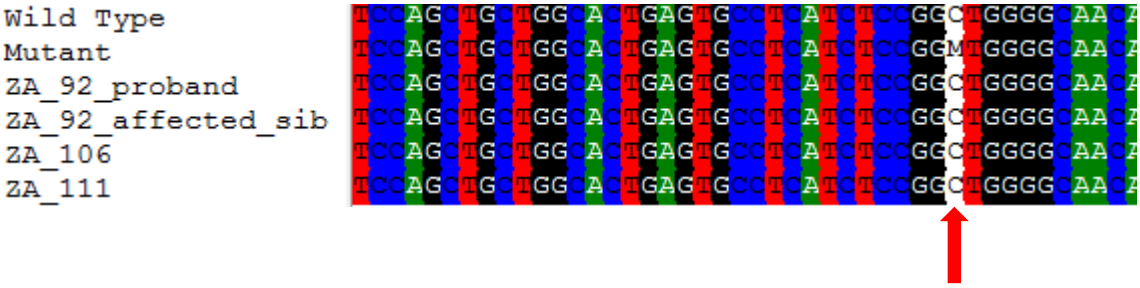
P - pathogenic; D - damaging; T - tolerated; B - benign; NA - stop/gain mutation; n - number of chromosomes

Appendix V- Primer sequences designed for Sanger sequencing validation

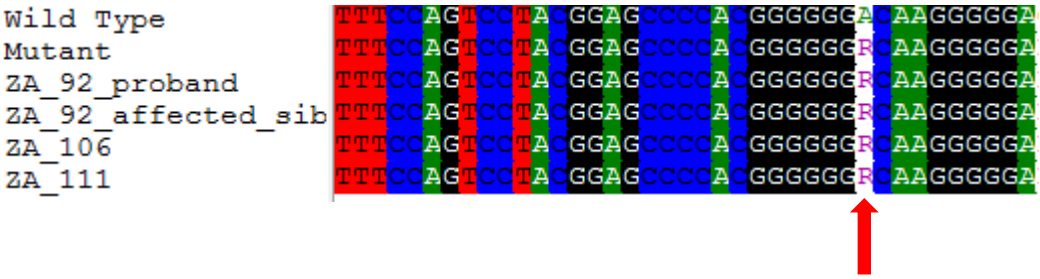
Gene	SNP	Primer sequence (5' – 3')	% GC	Tm (°C)	PCR conditions (Ta in °C)	Size of fragment (bp)
ANAPC1	Q451H	For: TGG TGA TGG TTG TAT AAC ACT GTA Rev: GAC CAG TGC TGA CAG TTA CTT	37.50 47.62	57.64 57.89	52	327
BCLAF1	T837N	For: TTC ACC CAC CAA ACC TAT ATC AA Rev: AGG AAG AGG TCG TGG TAC TT	39.15 50.00	57.48 57.69	52	232
C22orf42	M120I	For: CTG AAT CAT GAT GGC AGT CAA A Rev: GCC CAC CTC CAC ATA CTC	40.91 61.11	57.02 56.72	55	268
GPRIN2	W91R	For: CCG CCT GAG AGC ATG AA Rev: CTT CCG AGC ACC ACT GT	58.82 58.82	56.52 56.36	55	213
NOTCH2NL	T158I	For: ATG TGC CTC AGG GTT TAC AG Rev: GCT AAA TAA AGG TAT CTG CTG AAG G	50.00 40.00	57.50 57.80	55	201
MAP2K3	Q73X	For: AGT GAT CAG TAC CGA GGC TA Rev: CGT AGA AGG TGA CAG TGT AGA AA	50.00 43.48	57.27 57.88	55	235
PRSS3	A145T	For: CTG GGA GAG CAC AAC ATC AA Rev: GGG AGG CAA GAG GAT TCA AAT A	50.00 45.45	57.81 57.83	55	328
IL32	D172G	For: GTT CTG GCC TGG GTG AAG Rev: AGG TGG TGT CAG TAT CTT CAT TT	61.11 39.13	57.60 57.49	55	209
TIMM23	G23E	For: CCC GCT GTT ATT GAG GAG TAA Rev: AGC CAT GCA GAT ACA CTA ACC	47.62 47.62	57.46 57.80	55	341
KATNAL2	S301C	For: GCG TCC GGA TTG TTC CT Rev: CCA ACT ACG ATC TGC TGT CC	58.82 55.00	56.87 55.00	55	214
CNTN5	L70R	For: CCA CTT CAT ATG CTG CTT TGT T Rev: TCT CTC CTA CTG GAA TAG TCT TGA	40.90 41.70	53.80 54.10	55	296
TIMM23	D9N	For: GAC GCG CAA CTT AGT GTA GA Rev: GGG TTA CCC GCT GTT ATT GA	50.00 50.00	58.03 57.59	55	261

Appendix VI – Validation of candidate variants through Sanger sequencing and High Resolution Melt analysis

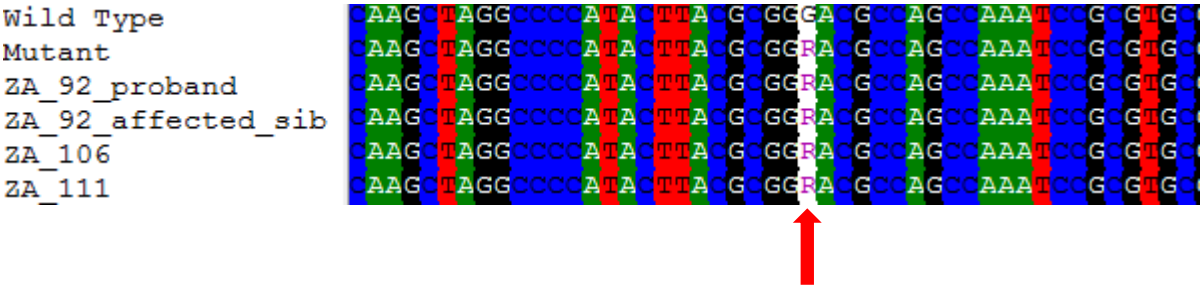
(A) ***PRSS3*** **(A145T)**



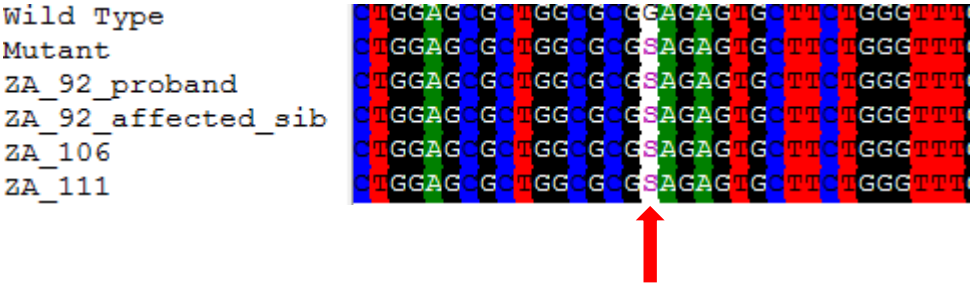
(B) ***IL32*** **(D172G)**



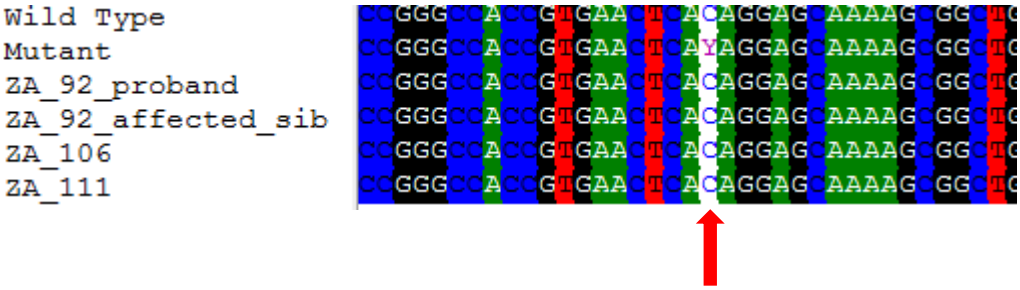
(C) ***TIMM23*** **(G23E)**



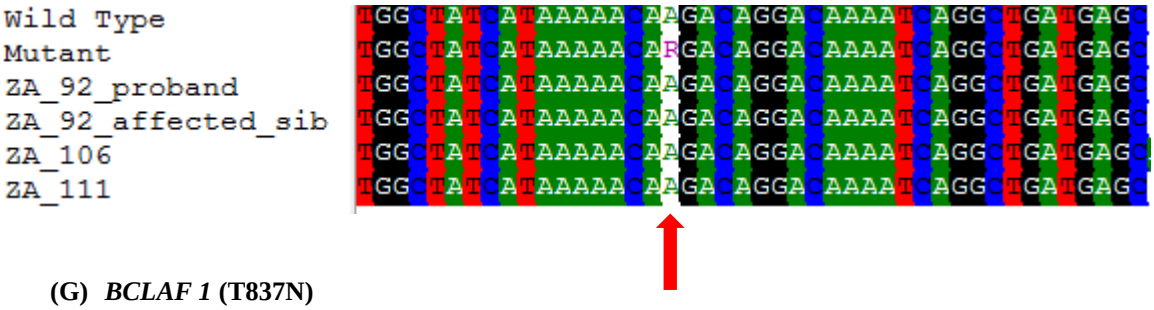
(D) ***KATNAL2*** **(S301C)**



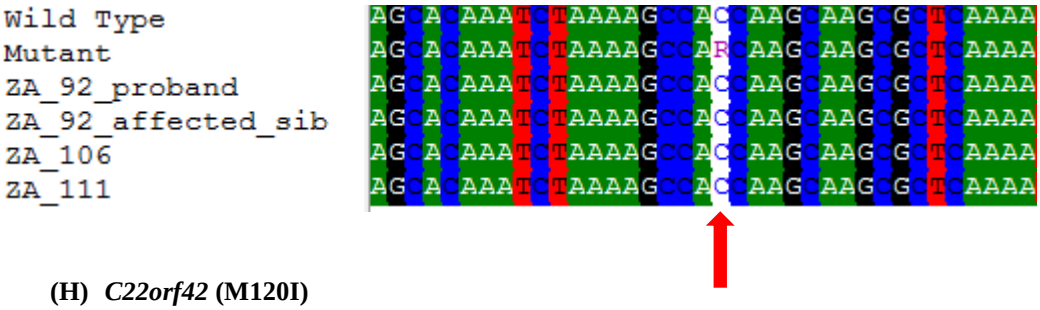
(E) ***MAP2K3*** **(Q73X)**



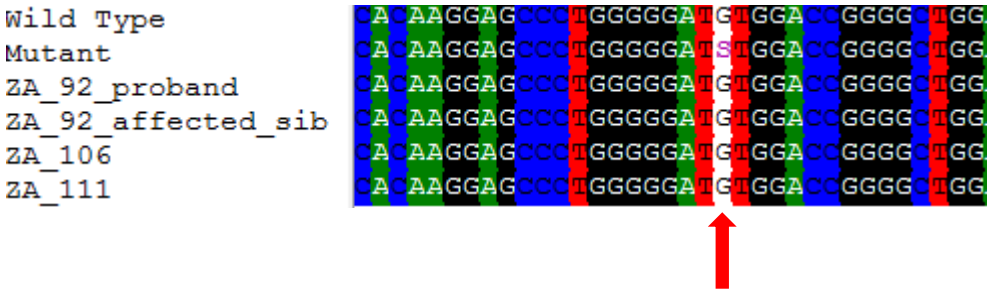
(F) ***ANAPC1*** **(Q451H)**



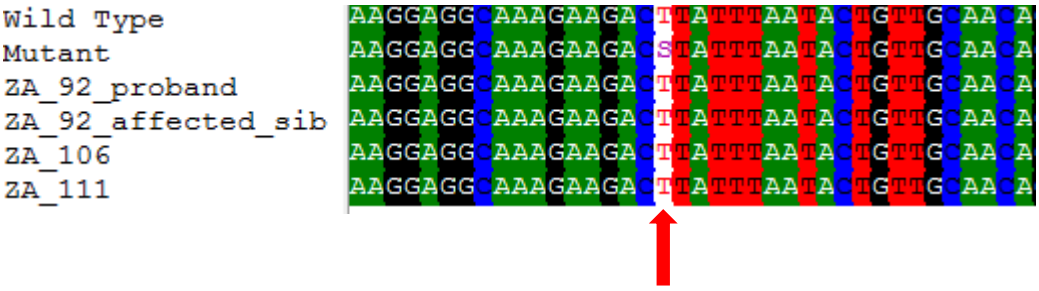
(G) ***BCLAF 1*** **(T837N)**



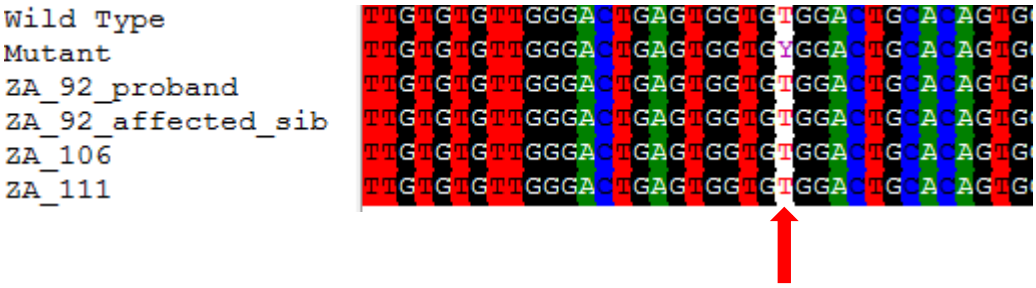
(H) ***C22orf42*** **(M120I)**



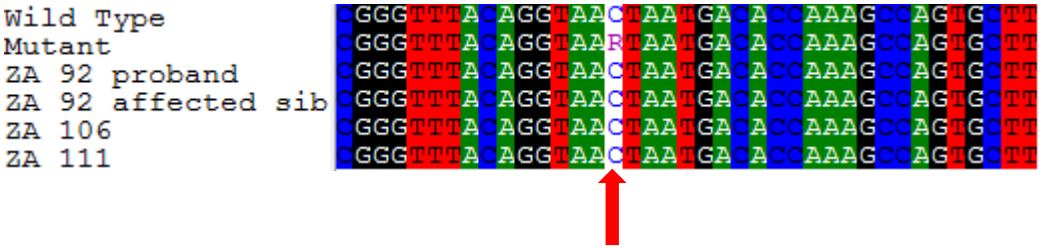
(I) ***CNTN5*** **(L70R)**



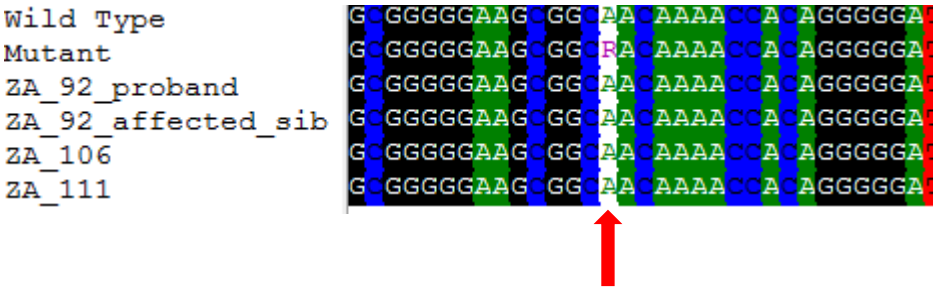
(J) ***GPRIN2*** **(W92R)**



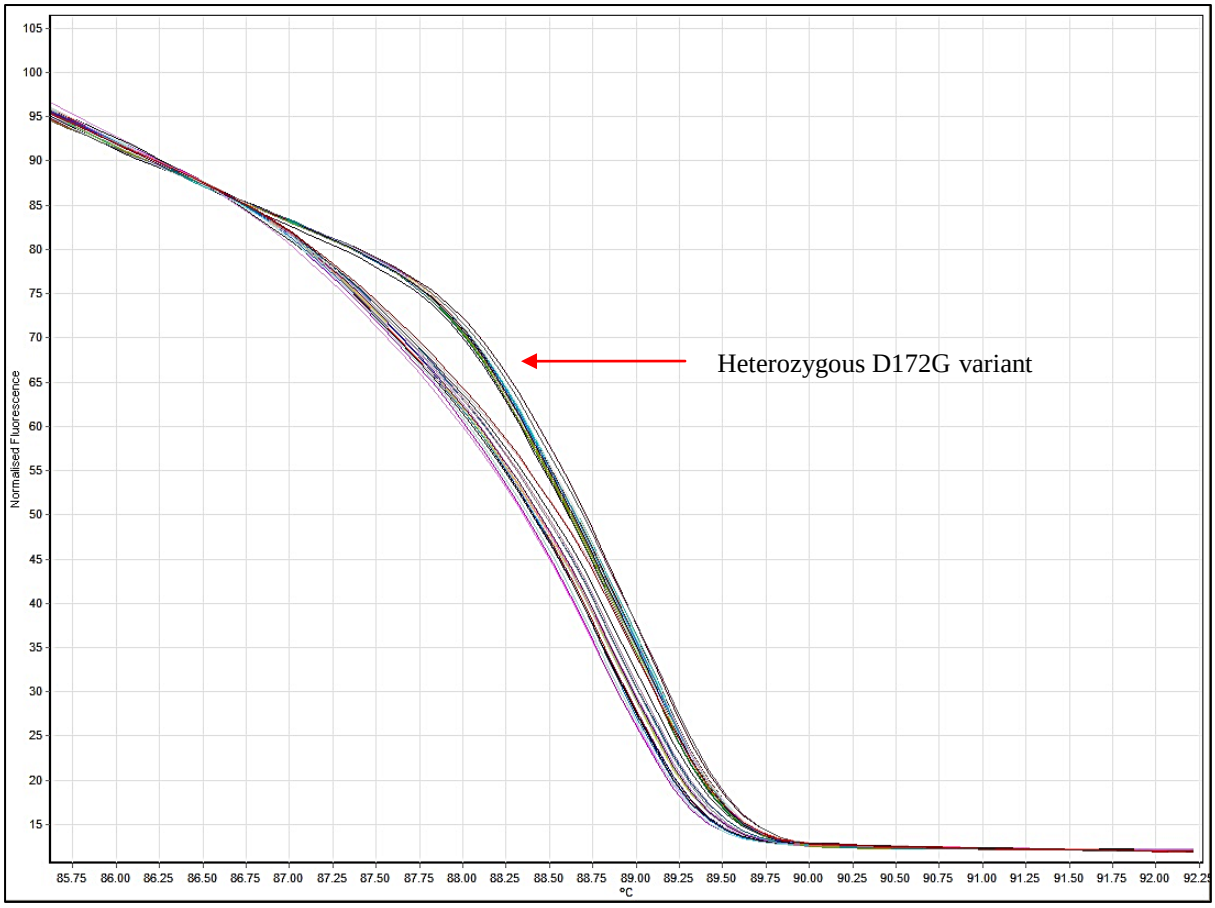
(K) ***NOTCH2NL*** **(T158I)**



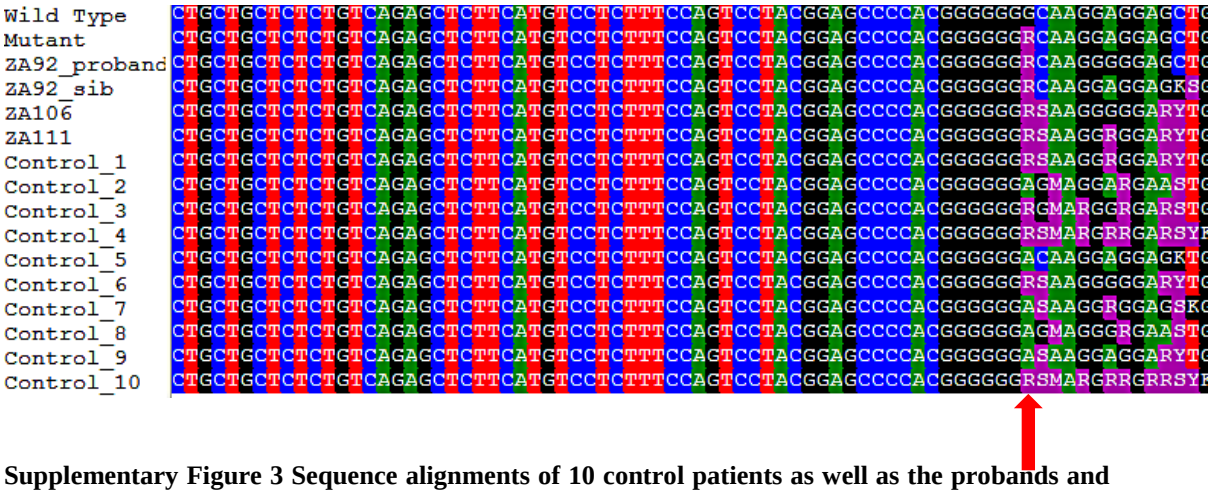
(L) ***TIMM23*****(N9D)**



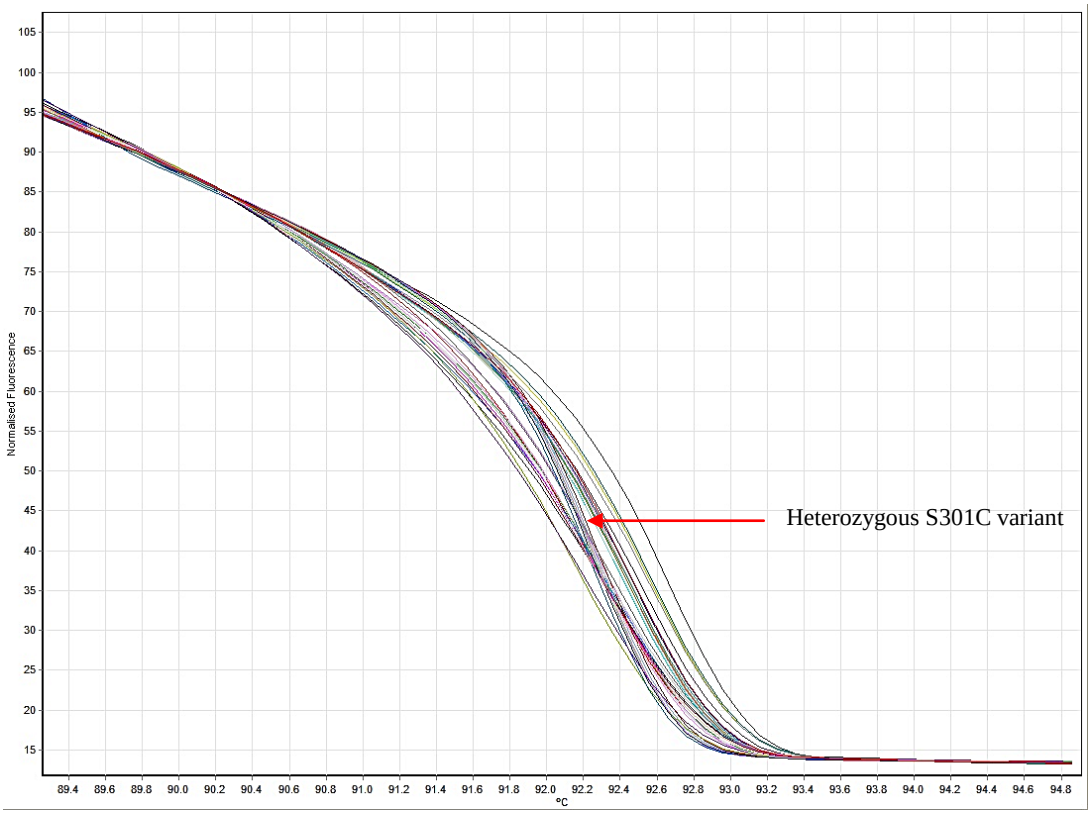
Supplementary Figure 1 Sequence alignments of the prioritized variants common across three PD probands and one affected sibling. The location of the specific SNP is indicated by the red arrow. The wild type is the reference sample, the mutant sample is a construct indicating the position of the variant in a heterozygous state.



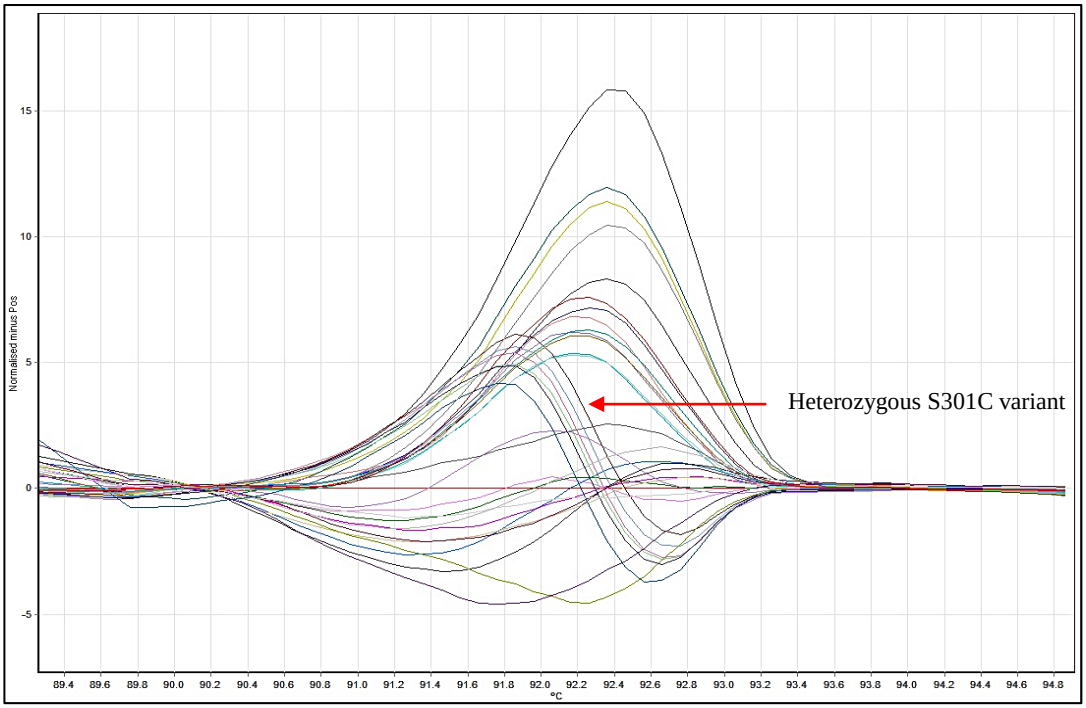
Supplementary figure 2 HRM normalized graph indicating the heterozygous D172G variant in *IL32* the positive controls as well as Afrikaner controls.



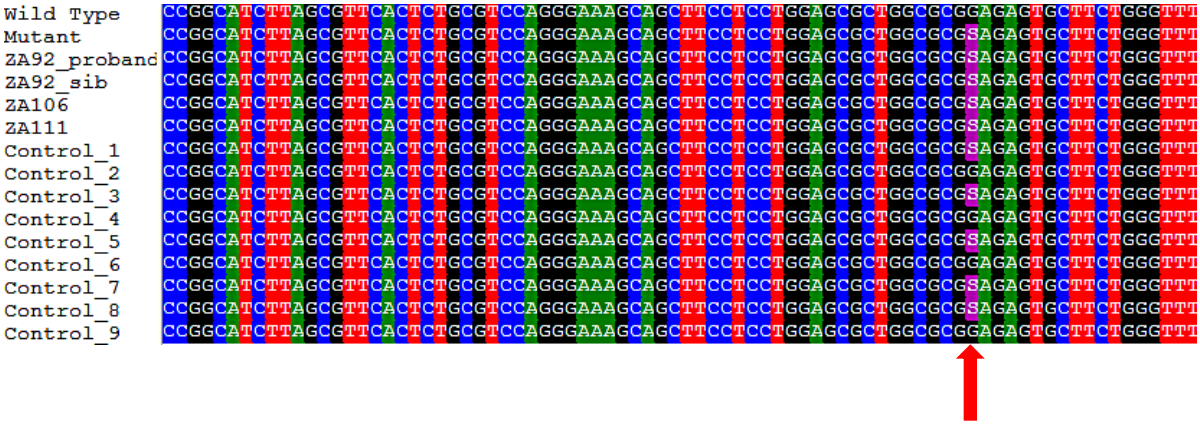
Supplementary Figure 3 Sequence alignments of 10 control patients as well as the probands and affected sibling for *IL32* (D172G). The location of the SNP is indicated by the red arrow. The wild type is the reference sample, the mutant is the sample in which a heterozygous change would be present.



Supplementary Figure 4 HRM normalized graph indicating the heterozygous S301C variant in *KATNAL2* in the positive controls as well as Afrikaner controls.



Supplementary Figure 5 HRM difference graph indicating the heterozygous S301C variant in *KATNAL2* in the positive controls as well as Afrikaner controls.



Supplementary Figure 6 Sequence alignments of nine control patients as well as the probands and affected sibling in *KATNAL2* (S301C). The location of the SNP is indicated by the red arrow. The wild type is the reference sample, the mutant is the sample in which a heterozygous change would be present.

Appendix VII: WES results obtained using the hypothesis-free approach

Supplementary Table 1 Stepwise breakdown of results obtained by TAPER™ for already existing WES results.

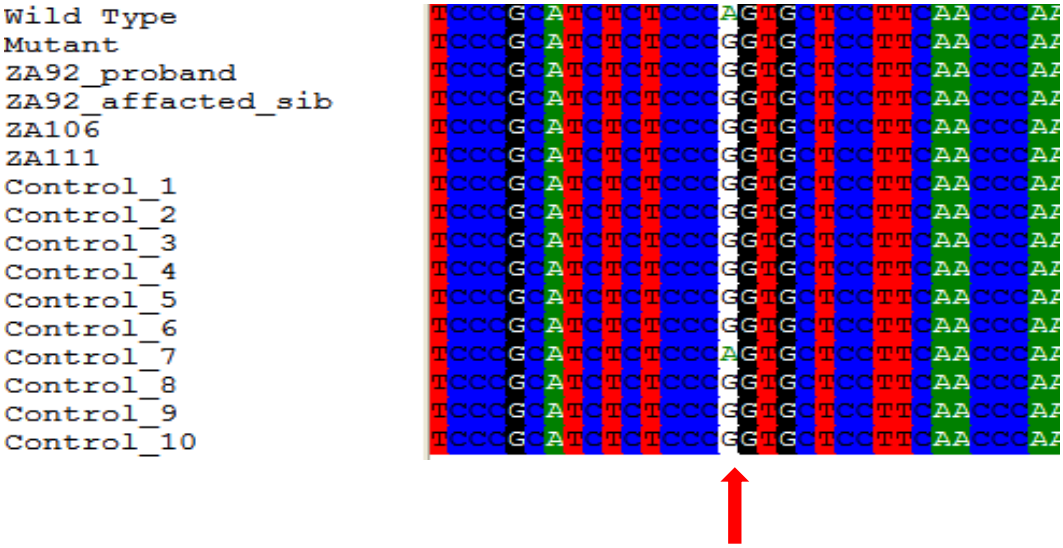
	Parkinson's disease dataset 1 – L34R in *FBOX7*			**Intellectual disability and microcephaly dataset 1 – E256K in *SLC1A4***		**Ataxia and myoclonic epilepsy dataset 1 – R297Q in *KCNA2***	**Parkinson's disease dataset 2 – R275W and M432V in *PARK2***		
	Individual_1	**Individual_2**	**Individual_3**	**Individual_1**	**Individual_2**	**Individual_1**	**Individual_1**	**Individual_2**	**Individual_3**
Total number of variants in VCF file	55 726	55 336	55 289	54 426	54 574	60 128	104 307	108 243	97 833
STEP 1: Total number of variants assigned to exonic regions by wANNOVAR	19 727	19 969	20 353	24 573	24 425	23 747	19 850	19 972	19 863
STEP 2: All synonymous and non-frameshifts removed	9 465	9 544	9 766	12 227	12 248	11 693	9 752	9 777	9 838
STEP 3: Remove all variants with a frequency >1% in 1KGP	1 281	934	966	2 177	2 153	1 377	1 771	1 681	1 932
STEP 4: Remove all variants with a frequency >1% in ESP6500	917	797	819	1 928	1 906	941	1 335	1 445	1 575
STEP 5: Remove all variants with negative GERP+++ scores	718	615	651	1 243	1 261	688	1 014	1 126	1 232
STEP 6: Remove all variants with FATHMM scores greater than 1.0	262	224	240	252	231	257	413	301	328
STEP 7: Variants removed from X and Y chromosome	252	221	236	241	221	202	398	262	292
STEP 8: Variants linked to relevant diseases	34	37	31	56	57	23	2	2	2
Variant of interest in shortlist?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Supplementary Table 2 Primer sequences for the six candidate variants identified through the use of TAPER™

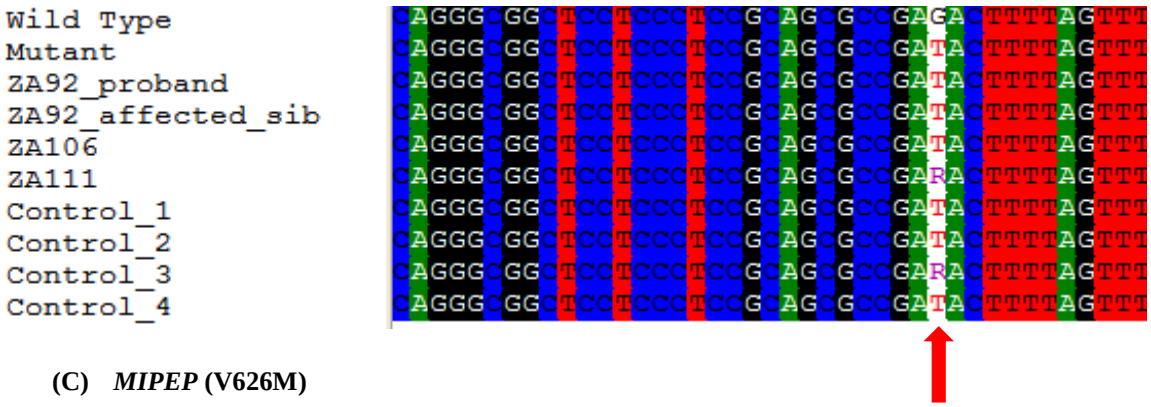
Gene	SNP	Primer sequence (5' – 3')	% GC	Tm (°C)	PCR conditions (Ta in °C)	Size of fragment (bp)
WNK1	S719G	For: GGGAGAGGATGGAACATTTCTT Rev: CAGGACACTCTCCAATGCTTTA	45.5 45.5	62.0 62.0	65.0	301
CASP7	E43D	For: GGACACGGGTCGCTTTG Rev: TAGGAAGCCGGCAGGAA	64.7 56.8	63.0 63.0	60.0	299
SYNJ1	V1405I	For: AACCCATTTAGAGCCAAGTCTG Rev: CCATCCTTTCGGGTTGCTAAT	45.53 47.62	58.31 58.35	55.0	252
USP17	C357S	For: CTTGTCTATGTCCTCTATGCTGTG Rev: GTGTCTTTCCCATTCACTCTTCT	45.8 43.6	62.0 62.0	Touchdown PCR (Temperature range 62 – 50)	201
MIPEP	V626M	For: CTGTGTCTCCACTGTGTTCT Rev: CCAGGTTTGGCTGTGTTATG	50.0 50.0	61.0 61.0	62.0	321

Appendix VIII - Validation of candidate variants through Sanger sequencing and High Resolution Melt analysis

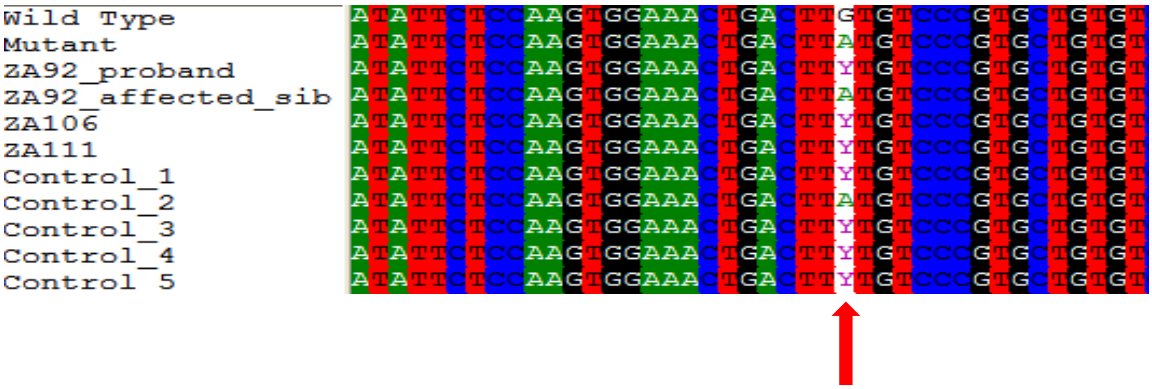
(A) *WNK1* (S719G)



(B) *CASP7* (E43D)



(C) *MIPEP* (V626M)



Supplementary Figure 11 Sequence alignments of the prioritized variants common across three PD probands and one affected sibling. The location of the SNP is indicated by the red arrow. The wild type is the reference sample, the mutant sample is a construct indicating the position of the variant in a homozygous state.

Appendix IX – DNA isolation from blood using the phenol/chloroform method

Extraction of nuclei from whole blood

Blood from two 5ml EDTA tubes per patient was transferred into a 50ml Falcon tube. The tube was then filled to 20 ml with ice-cold lysis buffer and inverted gently a few times. Subsequently, the sample was incubated on ice for 5-10 min. The sample was then centrifuged at 2500-3000 rpm at room temperature in a Beckman model TJ-6 centrifuge (Scotland, UK). The supernatant was discarded and the pellet was resuspended in 20ml, ice-cold lysis buffer that was then followed by another round of incubation and centrifugation. The supernatant was discarded and the pellet resuspended in DNA extraction buffer, after which the nuclei were either immediately used for DNA extraction, or stored at -70°C until DNA was required for genetic testing.

Extraction of DNA from nuclei

A total volume of 100µl of proteinase K (10µg/ml) was added to newly prepared or defrosted nuclei and the mixture was incubated overnight at 37°C. After this step, 2ml distilled water, 500µl 3M sodium-acetate and 25µl phenol/chloroform were added to the sample. The tubes were subsequently inverted and mixed gently for 10 min on a Voss rotator (Voss of Maldon, England) at 4°C. The mixture was then transferred to a glass Corex tube so that the aqueous phase could be clearly distinguished from the organic phase, followed by centrifugation in a Sorvall RC-5B refrigerated super-speed centrifuge (rotor SS 34, Dupont Instruments) at 8000rpm at 4°C for 10 min. The upper aqueous phase contained the DNA and was transferred to a clean Corex tube using a sterile plastic Pasteur pipette, while taking care not to disturb the interface or the organic phase. Approximately 25ml chloroform/octanol was added to the aqueous phase after which the tube was closed with a polypropylene stopper and gently inverted for 10 min. This mixture was centrifuged at 4°C, followed by the removal of the upper aqueous phase as described earlier. The DNA was then ethanol precipitated by adding two volumes of ice-cold 96% ethanol and inverting gently until DNA strands appeared as a white precipitate. The DNA strands were removed using a yellow-tipped Gilson pipette and placed in a clean,1.5ml Eppendorf microfuge tube. One millilitre 70% ethanol was then added to the DNA and the mixture centrifuged in a Beckman microfuge for 3 min at 13000 rpm. The ethanol was carefully decanted and the 70% ethanol wash step was repeated one more time in order to remove any excess salts. After careful removal of most of the ethanol, the DNA pellet was air-dried for 30-60 min at room temperature by inverting the Eppendorf microfuge tube on Carlton paper. Two hundred microlitres Tris-EDTA buffer was added and the DNA was resuspended, initially by stationary incubation at 37°C overnight and subsequently by gentle mixing in a Voss rotator at 4°C for a further 3 days. This was followed by stationary incubation at 4°C until the DNA had been fully resuspended.

After 1-2 weeks, when the DNA had completely resuspended in the buffer, the optical density(OD) of the DNA was determined in a Milton Roy series 120i spectrophotometer (USA) at 260nm (OD260). The DNA concentration, in □g/□l, was determined by diluting 10µl of DNA in 500µl of TE and multiplying the measured OD260 by a factor of 2.5, while the purity of the DNA was monitored by the OD260//OD280 ratio, which should be approximately 1.8 for pure DNA.

Appendix X – Reagents and Solutions

Cresol Loading Dye

2% (v/v) 10mg/ml cresol stock solution

0.9933M sucrose

10x TBE Electrophoresis Buffer (pH 8.3)

0.0890M Trizma Base

0.0890M Boric Acid

0.0020M EDTA