Identification of genetic modifiers of Parkinson's disease age-at-onset for *LRRK2* G2019S carriers

Submitted to University College London in partial fulfilment of the requirements for the Master's of Research in Translational Neurology

Supervisors: Dr Alan Pittman and Dr Mie Rizig
Word count: 16468

Module: Research project

Candidate Number: QPGK2

Statement of originality

I confirm that the work in this thesis is my own. Where data or information was obtained from other sources it has been indicated in the text.

Acknowledgements

I am very grateful to Dr Alan Pittman, who has been extremely generous in providing day-to-day guidance on both the project and the wider field of neurogenetics and bioinformatics, as well as in allowing me to beta-test the Illumina Truseq Neurodegeneration Panel which he was involved in designing. Equally I am greatly appreciative to Dr Mie Rizig for her advice on the project, and for inviting me to attend clinic, and witness first-hand the variable penetrance and expressivity of the G2019S mutation. Thanks go to Debbie Hughes for providing instruction on the technical aspects of Next-generation sequencing. Professor Huw Morris and Professor Nicholas Wood were generous in contributing advice and guidance. Thanks also go to Hallgeir Jonvik, Manuela Tan, Demis Kia, David Zhang, Lea R Bibo, Michael Thor, Heather Ging and anyone else who introduced me to new techniques or advised me in carrying out practical aspects, enabling the construction of this thesis.

Table of Contents

List of Figures

List of tables

1 Gene abbreviations

LRRK2	Leucine-rich repeat kinase 2
DNM3	Dynamin 3
GBA	Glucocerebrosidase
ATP13A2	ATPase 13A2
SNCA	Synuclein alpha
DJ-1	Deglycase DJ-1
PINK1	PTEN-induced putative kinase 1
UCLH1	Ubiquitin carboxyl-terminal esterase L1
HTRA2	HtrA serine peptidase 2
PLA2G6	Phospholipase A2 Group VI
FBXO7	F-box only protein 7
VPS35	Vacuolar protein sorting-associated protein 35
EIF4G1	Eukaryotic translation initiation factor 4 gamma 1

Identification of genetic modifiers of Parkinson's disease age-of-onset in *LRRK2* G2019S carriers

2 Abstract

Background

The Gly2019Ser (c.6055G->A) missense mutation in the *leucine-rich repeat kinase (LRRK2)* gene is the most common single mutation linked to Parkinson's disease (PD), with prevalence worldwide estimated at 1-4% of PD cases. Prevalence varies greatly between different ethnic populations: North Africans (34-41%) Ashkenazi Jewish (10-25%), and European (1-2%). G2019S exhibits age-dependent incomplete penetrance and variable phenotype severity. Recent studies have estimated penetrance at 25-42.5% at age 80 (Lee et al., 2017), suggesting a significant contribution from other genetic and environmental modifiers.

The heritability in AAO for G2019S PD has not been established, although a number of regions have been implicated but not independently replicated, including *Dynamin 3 (DNM3)* and the LRRK2 trans allele (Trinh et al., 2016).

The work presented here uses both targeted and whole-exome next-generation sequencing (NGS) techniques and supplementary Sanger sequencing in AAO genetic modifier validation and discovery.

Hypotheses

i) An undiscovered functional variant in *DNM3* is associated with AAO.

ii) The trans *LRRK2* haplotype influences AAO.

iii) Other genetic modifiers of AAO penetrance and phenotype severity exist, both rare and common.

Methods

Discovery cohort

123 Parkinson's disease patients and first degree relatives (predominantly Ashkenazi Jewish and North African) were Sanger sequenced for the G2019S mutation.

41 G2019S mutation carriers underwent Illumina Truseq Targeted Neurodegeneration Panel next-generation sequencing (NGS), and 38 were Sanger sequenced for a tag SNP (rs2206543) in *DNM3*.

Archive cohorts

35 European Caucasian G2019S carriers were extracted from archive NGS data:

1. Targeted PD gene (re-sequencing) NGS data (n = 22)
2. PD exomes (n = 13)

Analyses

Logistic regression, Kaplan Meier survival curve analysis and Cox proportional hazards modelling was used in assessing AAO modification of *LRRK2* and *DNM3*, and for discovery of exome-wide modifiers. *LRRK2* haplotypes were characterised.

Results and conclusions

SNPs in *LRRK2* alternative haplotype, *DNM3* rs2206543, and genome-wide data were not significantly associated with altered AAO. However, AAO bimodal distribution supports the existence of genetic modifiers. Pathogenic *LRRK2* haplotypes were all characterised as European-MENA (haplotype 1).

3 Introduction

3.1 Parkinson's disease

Parkinson's disease (PD) is a complex multifactorial neurodegenerative disease, affecting an estimated 7-10 million people worldwide. Clinically, PD is defined by the combination of bradykinesia (slowness of movement) and at least one other of the following principal motor signs: tremor, muscular rigidity of the limbs and trunk, and postural instability, with a range of other symptoms clarifying (i.e. unilateral onset, >5 year positive response to levodopa treatment), or precluding diagnosis (i.e. history of repeated head injury, Babinski sign (Ali and Morris, 2015)). Patients gradually develop motor impairments, caused by the loss of dopaminergic neurons of the substantia nigra pars compacta (SNpc) (Fearnley and Lees, 1991b). This loss is generally associated with accumulation of fibrillary aggregates composed of alpha-synuclein (Spillantini et al., 1998) and other proteins (Wakabayashi et al., 2007), called Lewy bodies. Consequently, loss of the striatal projections of the dopaminergic neurons occurs (Fearnley and Lees, 1991a).

Understanding of PD has oscillated between genetic and environmental conceptions, with a historical emphasis on environmental influence. Observations of postencephalitic parkinsonism after the early 1900s influenza pandemic strengthened the environmental hypothesis (Bonifati, 2012), as did the discovery of a chemical compound, resemblant of a widely used pesticide, that induced permanent parkinsonism in humans, by inhibition of complex I of the electron transport chain in mitochondria: 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) (Hala et al., 1983). A landmark 1997 study saw the discovery of the first PD gene, alpha-synuclein (Polymeropoulos et al., 1997). However, it was the identification of *LRRK2* in 2004 that was instrumental in supplanting the long-held view of PD as non-genetic (Paisan-Ruiz et al., 2004, Zimprich et al., 2004b). Commonly occurring

mutations in *LRRK2* caused late-onset, typical disease with reduced penetrance, whereas previously discovered genes had caused rare, highly penetrant, early-onset and atypical disease. However additional environmental risks (pesticides, well-water consumption and head trauma), and protections (coffee intake, cigarette smoking (Hernan et al., 2002), and potentially, higher BMI (Noyce et al., 2017), have also been parsed. As indicated in Figure 1, PD is now considered to be caused by a complex interplay of genetic, environmental and epigenetic factors.

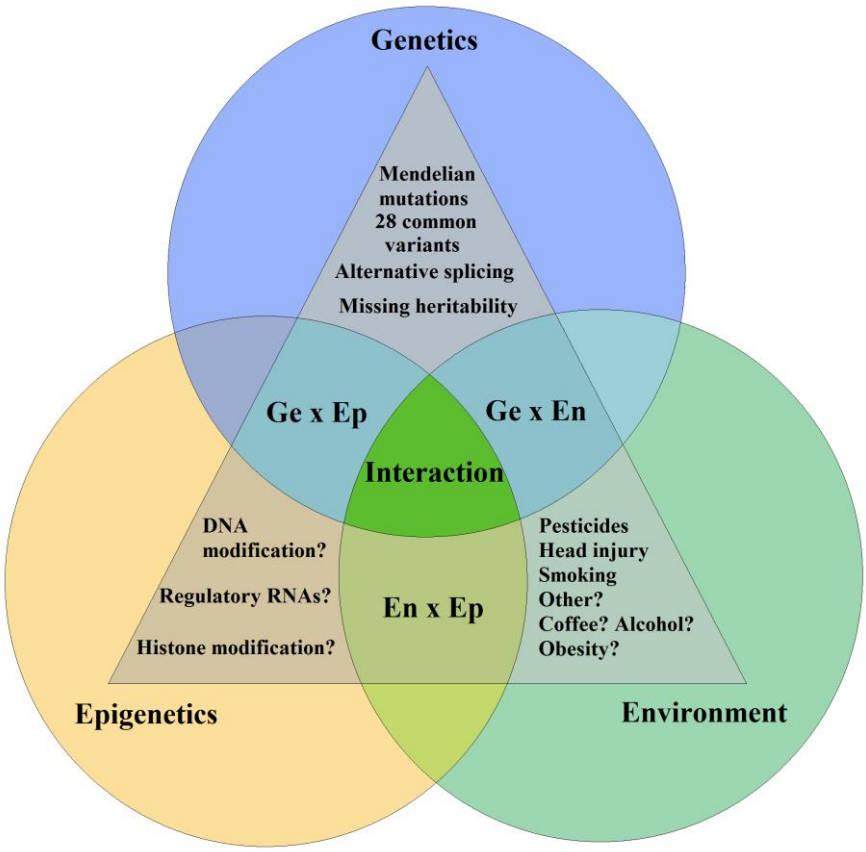


Figure 1. Depiction of the interaction between factors impacting PD. Adapted from Lill et al (2016)

In terms of genetic causes, 17 chromosomal locations (PARK1-18) have been identified in association with inherited PD, as shown in Table 1. Currently, autosomal dominant mutations in alpha-synuclein (*SNCA*) and *LRRK2*, and autosomal recessive mutations in *Parkin* (Kitada et al., 1998), *PINK1* (Valente et al., 2004), *DJ-1* (Bonifati et al., 2003) and *ATP13A2* (Ramirez et al., 2006) have been confirmed as conclusively Parkinson's disease causing. *Figure 2* shows the use of genetic screening in clinical practice for Mendelian mutations. There are also 28 common variants across 24 loci which act as risk factors for sporadic disease (Nalls et al., 2014). It therefore seems likely that understanding the roles of PD associated genes will bring insight into shared pathological mechanisms between familial and idiopathic disease.

Table 1.Chromosomal locations identified in association with inherited Parkinsonism

Locus	Chromosome	Gene	Inheritance	Phenotype	Additional Information
PARK1	4q21	*SNCA*	Autosomal dominant	Early-onset PD	Missense mutations, genomic multiplications
PARK2	6q25.2-q27	*Parkin*	Autosomal recessive	Early-onset PD	
PARK3	2p13	Unknown	Autosomal dominant	Classical PD	Unconfirmed
PARK4	was erroneously called; the family were later shown to have a SNCA triplication.				
PARK5	4p13	*UCLH1*	?	Classical PD	Unreplicated mutations in a single sibling pair
PARK6	1p36	*PINK1*	Autosomal recessive	Early-onset PD	
PARK7	1p36	*DJ-1*	Autosomal recessive	Early-onset PD	
PARK8	12q12	*LRRK2*	Autosomal dominant	Classical PD	
PARK9	1p36	*ATP13A2*	Autosomal recessive	Juvenile parkinsonism, pyramidal signs, dementia	
PARK10	1p32	Unknown	Risk factor	Late-onset parkinsonism	Unconfirmed
PARK11	2q37	Unknown	Risk factor	Late-onset parkinsonism	
PARK12	Xq21-25	Unknown	X-linked	Late-onset parkinsonism	Unconfirmed
PARK13	2p12	*HTRA2*	Autosomal dominant or risk factor	Late-onset parkinsonism	Unconfirmed
PARK14	22q12-q13	*PLA2G6*	Autosomal recessive	Early-onset parkinsonism-dystonia	
PARK15	22q12-q13	*FBXO7*	Autosomal recessive	Juvenile parkinsonism and pyramidal signs	
PARK16	1q32	Unknown	Risk factor	Late-onset parkinsonism	Unconfirmed
PARK17	16q11.2	*VPS35*	Autosomal dominant	Classical PD	
PARK18	3q27.1	*EIF4G1*	Autosomal dominant	Classical PD	Unconfirmed

Adapted from Sheerin et al (2014)

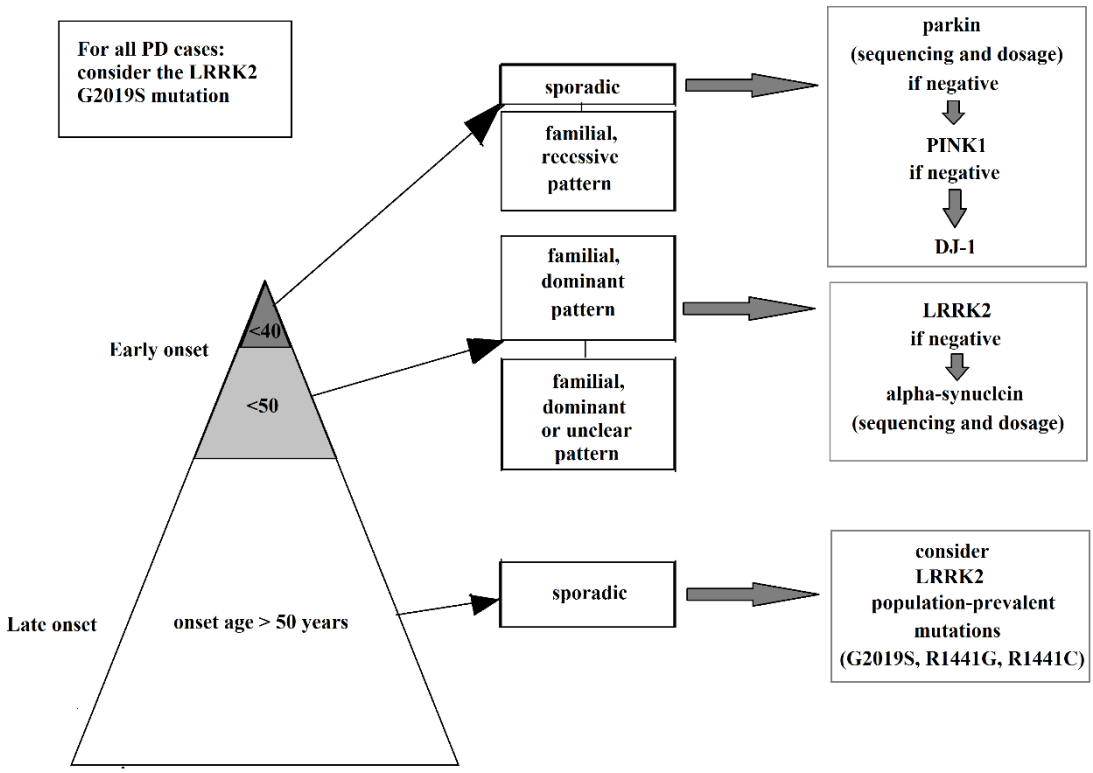


Figure 2. A potential algorithm for determination of appropriate genetic screening in PD. G2019S has entered clinical practice. Some clinicians recommend ubiquitously screening for G2019S mutation. An alternative economical approach is to screen for this mutation when disease is familial and late-onset, or in all patients of particular high prevalence ethnic populations. Image adapted from Bonifati and Wood (2012)

It is worth emphasizing the heterogeneity of PD at this stage, in terms of clinical presentation, underlying pathologies, and genetics. Geneticists have taken to quoting Tolstoy regarding the contribution of genetics to complex disease: "Every unhappy family is unhappy in its own way." A large proportion of the PD burden still has unexplained genetic inheritance: only 5-10% of patients (depending on the population) have a known PD mutation, and discovered common risk variants contribute less than 5% of genetic heritability (Keller et al., 2012).

Commentators note that in general, genome-wide association studies have not delivered a significant portion of the genetic basis of common, complex traits. It has been suggested that the genetic architecture of complex disease is influenced by many genetic variants each with individually small effects. Lubbe et al (2016) present evidence that multiple rare variants of currently unknown significance, inherited with known PD Mendelian causes, may modulate risk and AAO. Indeed, it has been shown that polygenic inheritance of common variants is associated with early-onset PD (Escott-Price et al., 2015).

All 3.08 billion base pairs (bp) in the genome of *Homo sapiens'* have now been sequenced (Gandhi and Wood, 2012). Due to population expansion, it is estimated that every single nucleotide polymorphism that is compatible with life exists in a human somewhere. Advances in the cost and feasibility of whole-genome, whole-exome and targeted sequencing have meant that genetic variability that impacts phenotype heterogeneity in Mendelian disease can be explored. Particularly, NGS technologies allow unprecedented unbiased access to both rare and common variants. It is the interpretation of these variants, particularly those which are rare, that poses a tremendous challenge. For instance, the genome-wide association analysis approach (where hundreds of thousands of genetic variations are tested for association with a phenotypic outcome), can be problematic because of the massive number of statistical tests performed concurrently, which causes unprecedented potential for false positive results. Large sample size and independent replication are important for drawing conclusions.

3.2 Current understanding of *LRRK2* function and dysfunction

Since its discovery, leucine-rich repeat kinase 2 (*LRRK2*) has been the focus of intense research. The cytosolic multi-domain enzyme belongs to the ROCO superfamily of proteins

which is characterized by the presence of Ras in complex G domain (Roc) (which exhibits intrinsic GTPase activity and has a unique dimeric structure (Deng et al., 2008)) and the spacer C-terminal of Roc (COR) domain. Immediately next to the Roc-COR tandem is a kinase domain, belonging to the serine/threonine kinases. Together, these three domains appear to dominate the proteins function and dysfunction, with all disease segregating single nucleotide variants (SNV) inhabiting this region. *LRRK2* is one of only three discovered proteins in the human proteome exhibiting dual kinase and guanosine triphosphatase (GTPase) activities (Cogo et al., 2017). The central domains are flanked by multiple protein-protein interaction domains and repetitive regions. Ankyrin-like repeat (ANK) and leucine-rich repeat (LRR) are located at the *LRRK2* N terminus. At the C-terminus is WD40, which contains seven WD40 repeats and is essential for protein folding. This WD40 domain therefore impacts *LRRK2* function and kinase activity. The multiple protein-protein interaction domains indicate the potential of *LRRK2* to form a multifarious signalling node with many binding partners (Lewis and Manzoni, 2012).

Leucine-rich repeat kinase is a conspicuously large gene spanning 1.4 Mb, consisting of 51 exons and encoding 2527 amino acids. It is highly conserved across species (Bardien et al., 2011a), which may enhance its investigation in model organisms. However, as a complex multidomain protein with multiple protein-protein interactions *LRRK2* may overreach its targets and interacting partners when investigated through overexpression in non-physiologically relevant cell lines or animal models (Lee et al., 2017). Earlier research had focused on the role of *LRRK2* in neurons, where endogenous expression is low (Schapansky et al., 2014). *LRRK2* has been reported as interacting with copious molecules in diverse settings: endosome vesicle trafficking (Shin et al., 2008), cytoskeleton reorganization (Meixner et al., 2011), mitochondrial function (Wang et al., 2012), regulation of ER/Golgi retromer complex (MacLeod et al., 2013), and lysosomal autophagy (Manzoni et al., 2013);

and in signaling pathways such as wingless/int (Berwick and Harvey, 2012), TNF-alpha/Fas ligand (FasL)/Fas-associated protein with death domain (Ho et al., 2009), mitogen activated protein kinase and nuclear factor k-light-chain-enhancer of activated B cells pathways (Lee et al., 2017).

A consistent role for *LRRK2* in innate immunity is emerging. Cook et al (2017) have indicated that *LRRK2* is overexpressed in PD and *LRRK2*-mutant immune cells during PD. *LRRK2* may thus be recruited when a cell undergoes inflammation and immune reaction. Recruitment appears insufficient to stem neurotoxic pathways (Cook et al., 2017). Notably other key PD-associated genes, SNCA, DJ-1 and GBA, are expressed in immune cells (Gardai et al., 2013). Dopamine neurons in the substantia nigra pars compacta (SNpc) are also intrinsically vulnerable to metabolic stress, particularly when it is caused by dopamine oxidation or mitochondrial dysfunction (Lee et al., 2017). SNpc is a brain region with the highest density of microglia and a relatively low density of astrocytes (Kim et al., 2000, Savchenko et al., 2000), suggesting chronic activation of microglia and damage of dopaminergic neurones would be particularly harmful in the SNpc (Lee et al., 2017b).

Six mutations in *LRRK2*, which all occur at evolutionarily conserved amino acid positions (as shown in Figure 3, have been conclusively identified as causing Mendelian autosomal dominant disease: p.R1441G/C/H (Paisán-Ruíz et al., 2004, Zabetian et al., 2005, Zimprich et al., 2004a), p.G2019S (Di Fonzo et al., 2005, Gilks et al., 2005, Nichols et al., 2005), p.Y1699C (Paisan-Ruiz et al., 2004, Zimprich et al., 2004a), and p.I2020T (Funayama et al., 2005, Zimprich et al., 2004b). A number of additional *LRRK2* single nucleotide variants (SNVs) have emerged as PD risk factors for sporadic disease, and both coding and non-coding variants implicated in PD were discovered in genome-wide association (GWA) (Trabzuni et al., 2013), putatively linking Mendelian disease with iPD. Multiple investigators

have shown kinase activity of *LRRK2* is required for the pathological changes leading to PD (Lee et al., 2012).

Only the G2019S mutation, which is thought to occur in the DYG motif of the kinase activation loop (Liu et al., 2013), has been demonstrated to cause significant alteration in kinase activity, with a three to fivefold increase in autophosphorylation or phosphorylation of peptide substrates (Covy and Giasson, 2009). Whereas the other *LRRK2* pathogenic mutations were not demonstrated to have consistent effects on kinase function, despite I2020T for instance, being situated immediately adjacent to G2019S. It has been hypothesized that the glycine G2019S-WT in the DYG segment of the activation loop provides conformational flexibility (Shan et al., 2009), with the replacement by serine in G2019S mutants altering *LRRK2* dynamics. Many isoforms found in the substantia nigra appear to have had exons 32-33 (Trabzuni et al., 2013), and 42 spliced out (Giesert et al., 2013); these exons contain a portion of the COR and kinase domain, perhaps indicating functional import to this splicing event. Though Giesert et al (2013) used murine embryos and *Mus musculus* rather than PD tissues, potentially limiting disease relevance. *LRRK2* is expressed most robustly in the occipital cortex in adult human brains, compared to the cerebellum and white matter. Whereas expression in the substantia nigra and putamen (areas most implicated in PD) is middling (Trabzuni et al., 2013). It has been hypothesised that alternative splicing (perhaps splicing out) of LRRK2 may modify AAO PD penetrance.

LRRK2 mediates the phosphorylation of itself and other kinases, at multiple domains throughout the protein (Rideout, 2017). This activity is kept in check by the action of specific cellular phosphatases (Taymans, 2017). 14-3-3 binds to a cluster of phosphorylated residues located in the N-terminal of *LRRK2*. Phosphorylation of these residues depends on *LRRK2* kinase activity (Rideout, 2017). Recent phosphoproteomics work has identified Rab GTPases

as key *LRRK2* substrates (Steger et al., 2016). All pathogenic *LRRK2* mutants, as well as a number of risk factors show elevated phosphorylation of certain Rab GTPases, perhaps indicating that this is a ubiquitous toxic pathway (Rideout, 2017). Although different *LRRK2* mutants have been noted to interact in dissimilar ways in model systems; it was discovered that R1441C and Y1699C but not G2019S or wild-type *LRRK2* preferentially associate with deacetylated microtubules in cell lines, seemingly altering axonal transport and inhibiting axonal transport in Drosophila in vivo (Godena et al., 2014). Evidence is accumulating that the Roc/GTPase domain, likely in a guanine-nucleotide bound state is important not only as an intramolecular regulator of *LRRK2* kinase activity but also as a platform for interaction with heterologous partners, including PKA, Sec16A, PAK6, Rab proteins and tubulins (Manzoni et al., 2015). A PD protective variant, R1398H has also been shown to enhance GTPase and Wnt signaling activity (Nixon-Abell et al., 2016).

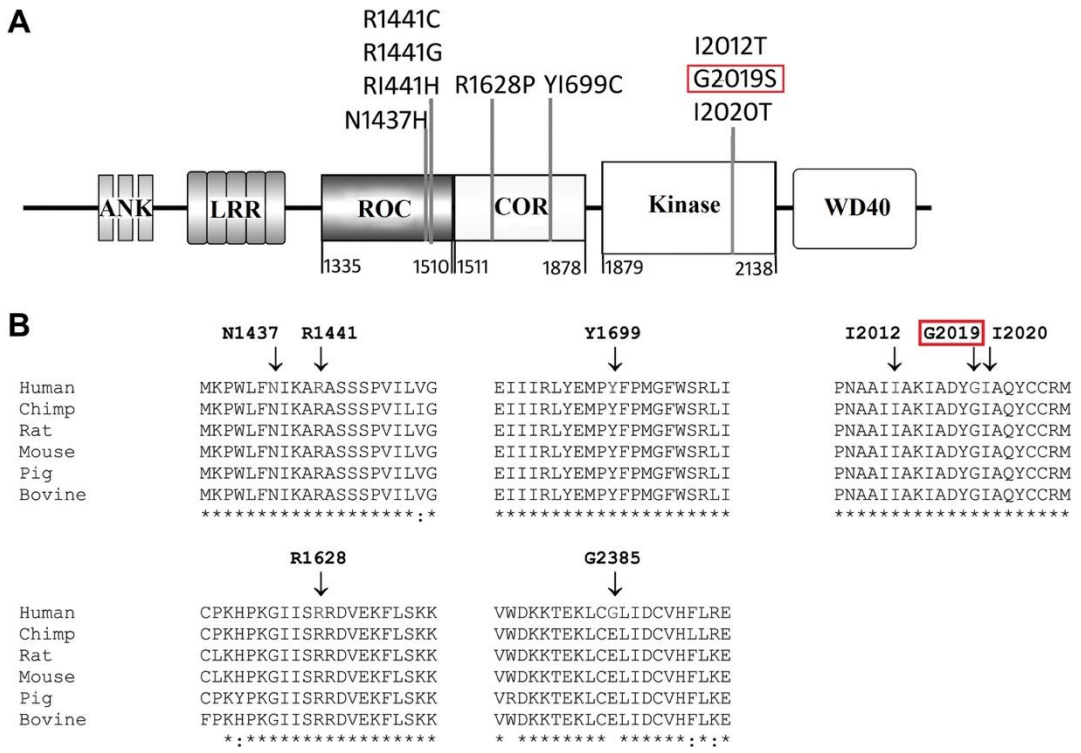


Figure 3. Pathogenic mutations are located at evolutionarily conserved positions. Adapted from Bardien et al (2011).

LRRK2 is a pleiomorphic locus with GWAS studies identifying its involvement in multibacillary leprosy (Zhang et al., 2009), Crohn's disease (Franke et al., 2010), inflammatory bowel disease (IBD) (Liu et al., 2011), and certain cancers (Hassin-Baer et al., 2009) (see Figure 4). Amongst the reported signalling pathways there is incomplete agreement regarding physiological relevance in the neuronal environment. The diversity of reported signalling pathways and roles in various diseases, may indicate different roles in different tissues. With reported roles in diverse processes (autophagy, the immune system, vesicle dynamics, retromer function, and mitochondrial dynamics), the appropriate and specific targeting of *LRRK2* will be necessary for therapeutic intervention. In addition to the uncertainty regarding which interactions occur in vivo, it is not elucidated whether (or which) specific altered interactions initiate or contribute to a pathogenic mechanism. Finding *LRRK2*

interacting partners in genetic studies may be important in understanding its physiological function and dysfunction in tissues appropriate to PD.

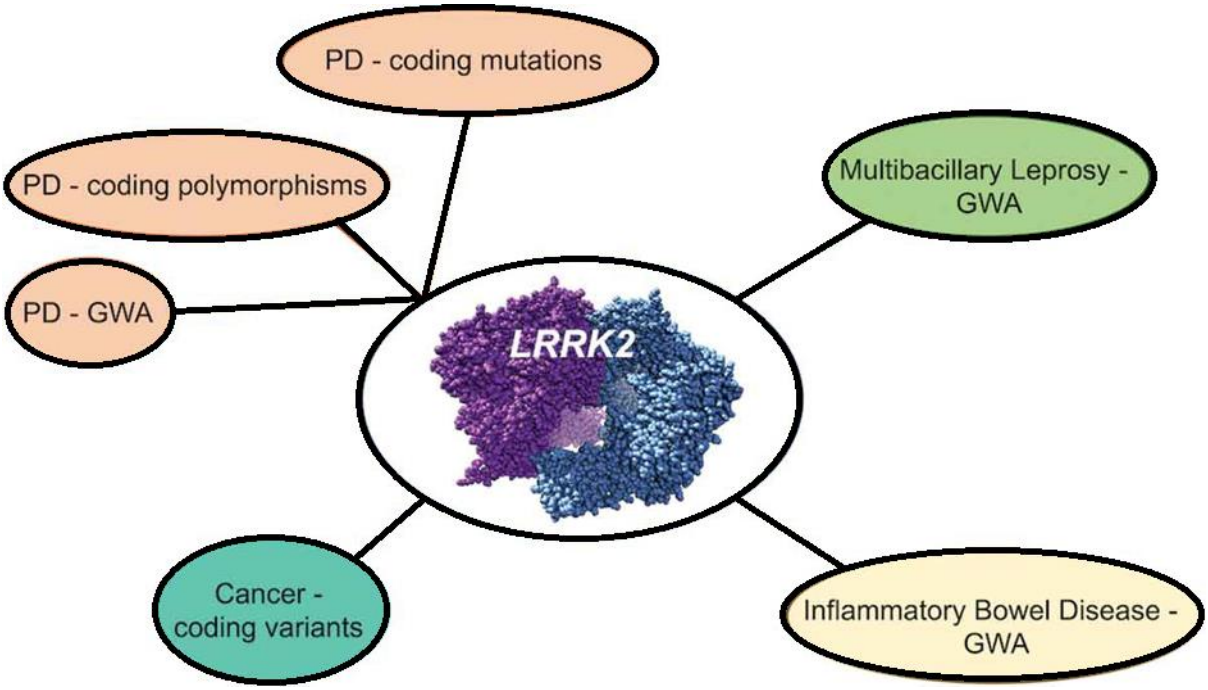


Figure 4. A 3D structural image of the LRRK2 dimer with various disease roles indicated. Image is adapted from Cogo et al (2017).

3.3 Worldwide prevalence, haplotypes, and natural history

3.3.1 G2019S prevalence across populations

G2019S, which is the most common worldwide PD mutation, has been reported as occurring at highest frequencies amongst familial PD cases of North African (37%), followed by Ashkenazi Jewish descent (23%) (Lesage et al., 2006, Ozelius et al., 2006). Figure 5 shows estimates of worldwide frequency for G2019S. Frequency has been high among sporadic cases (41% North Africans and 13% Ashkenazi Jews (Lesage et al., 2006, Ozelius et al., 2006)). The mutation is also regularly identified in healthy controls (3% North African and

1.3% Ashkenazi Jews) (Rideout, 2017). G2019S appears to demonstrate a subtle South-North gradient of frequency in Western Europe, with higher frequencies in Portuguese patients (9-16% of familial; 3-4% of sporadic) (Bras et al., 2005, Ferreira et al., 2007), followed by Catalonian (Gaig et al., 2006), Cantabrian (González-Fernández et al., 2007, Infante et al., 2006), Asturian (Mata et al., 2006), Galician (6-16% familial; 2-6% of sporadic) (González-Fernández et al., 2007), and Basque patients (1-2%) (Gorostidi et al., 2009). G2019S is notably very rare in Greece and Crete (Kalinderi et al., 2007, Papapetropoulos et al., 2008, Papapetropoulos et al., 2007, Spanaki et al., 2006). In the UK G2019S was identified in 2.5% familial and 0.3- 1.6% of sporadic disease (Gilks et al., 2005, Khan et al., 2005, Williams-Gray et al., 2006). More research is required into Middle Eastern and sub-Saharan African prevalence, for which data is lacking.

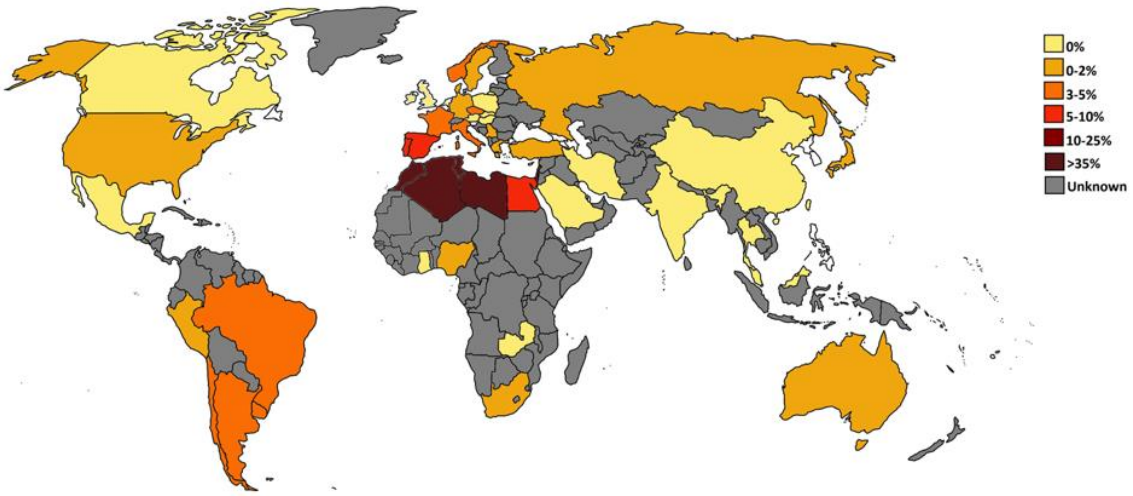


Figure 5. Estimates of worldwide prevalence of G2019S mutations. Adapted from Monfrini and Di Fonzo (2017)

3.3.2 G2019S haplotypes

A haplotype refers to a collection of specific alleles that are inherited in a tightly linked manner. Similarly, linkage disequilibrium (LD) describes two alleles at nearby sites co-

occurring on the same haplotype at greater frequency than predicted by chance (Devlin and Risch, 1995). LD patterns can be unpredictable and beset by noise; sometimes pairs of sites tens of kilobases apart are in complete LD, whereas nearby pairs of sites in the same region may have weak LD values. Two measures exist to evaluate the marker pairwise LD value: D' and r^2. Haplotypes are informative for uncovering the evolutionary history of a genetic mutation, as well as demographic effects: population growth, bottlenecks, admixture, and natural selection. The geographic centre of origin of a mutation typically corresponds to the area in which the mutation is most frequent (Ardlie et al., 2002).

The G2019S mutation has been found on three different haplotypes. These three haplotypes can be distinguished by markers only ~5 kb upstream and downstream of G2019S (Lesage et al., 2010), indicating that the mutation may have arisen independently in each haplotype. Although (as shown in Table 2 and Figure 6), haplotypes 2 and 3 exhibit far greater homology compared to haplotype 1, and are considered evolutionarily related. Network analysis has identified a non-carrier form of each of the three G2019S-carrier haplotypes at high frequency in the general population (Bardien et al., 2011b). Haplotype 2 is very rare and is shared by three European-American (Zabetian et al., 2006, Lesage et al., 2010) and two French families (Lesage et al., 2010). It is hypothesized to have occurred much more recently than haplotype 1. Haplotype 3 was found in Japanese individuals (Tomiyama et al., 2006, Zabetian et al., 2006) and one Turkish family (Pirkevi et al., 2009), and is similarly rare.

Table 2. Three haplotypes have been identified for G2019S, with SNPs numbered for convenience

SNP Number	Marker	Physical map	Haplotype 1	Haplotype 2	Haplotype 3
1	rs28903073	38939777	A	G	G
2	rs7966550	38974962	T	T	T
3	rs1427263	39000101	A	A	C
4	rs11176013	39000140	G	G	A
5	rs11564148	39000168	A	A	T
6	rs2404834	39015274	C	T	C
7	rs7302841	39015923	A	G	G
8	rs715402	39017481	A	A	A
9	rs6581667	39017773	C	G	G
	G2019S	39020469	A	A	A
10	rs10506155	39022206	G	A	A
11	rs919714	39022516	C	C	C
12	rs10784522	39026632	T	G	G
13	rs10878405	39028521	A	G	G
14	ss52051244	39043597	A	G	A

It is the common haplotype, haplotype 1 (also referred to as the European-MENA haplotype), which informs clinical genetic testing of PD patients. Haplotype 1 is most prevalent in Berbers (Change et al., 2008), followed by North African Arabs, Ashkenazi and Sephardic Jews. Haplotype 1, which spans 6.28-2.43 kb, depending on the markers genotyped, was identified in G2019S carriers from diverse populations including Italy (Marongiu et al., 2006), France (Lesage et al., 2005), Germany (Hedrich et al., 2006), Russia (Illarioshkin et al., 2007), Sardinia (Cossu et al., 2007, Floris et al., 2009), Spain (Mata et al., 2006), Portugal (Ferreira et al., 2007, Goldwurm et al., 2005), Brazil (Goldwurm et al., 2005), Chile (Perez-Pastene et al., 2007), Uruguay and Peru (Mata et al., 2009), and Australia (Huang et al., 2007).

It is hypothesized that this haplotype originated in North Africa or the Middle East and subsequently spread to other countries through migration patterns. Bar-Shira et al (2009) used a maximum-likelihood method on 77 G2019S predominantly Ashkenazi Jewish carriers and 50 AJ non-carriers to estimate that AJ with G2019S share a common ancestor who lived ~1830 years ago. Bar-Shira et al (2009) hypothesise that the common founder of G2019S lived before the most recent founder in the Ashkenazim, with a specific ancestral pattern of inheritance occurring in Ashkenazi Jews. The European-MENA haplotype has not been found in Iraqi Jews (Orr-Urtreger et al., 2007), who are descendants of the original Babylonian Jews (586 BCE) (Motulsky, 1995). Nor has the G2019S mutation been found in Yemenite Jews, who constitute a distinct ethnic Jewish group. These observations are considered consistent with the mutation occurring after or during the second Jewish diaspora (586 BC to 70 AD) (Bar-Shira et al., 2009).

Whereas Farrer et al (2008) propose that the mutation is Phoenician rather than Arabic in origin, and that it dates to the founding of ancient Carthage in 814BC. Supporting this theory, Al-Mubarak (2015) found no G2019S in 98 Saudi PD patients, 30 of whom had LOPD. Saudi Arabians have Phoenician heritage, and Berbers were not historically associated with the region. Hashad et al (2011) found that G2019S was common in 11/113 or 9.7% of sporadic PD patients in Egypt. Change et al (2008) found a higher frequency of *LRRK2* G2019S in Moroccan Berbers, compared to other North African populations, although it was not delineated to what extent the other North African populations were Berber or Arab. Farrer et al (2008) consider the admixture of these populations, which began in the 7th century AD with the arrival of Arabs in North Africa, so extensive as to render it unnecessary to distinguish. Whereas Bouhouche et al (2017) identified the shortest haplotype observed so far in a Berber-speaking PD patient, indicating that the mutation could have arisen in Berbers. It

seems most likely that the mutation originated in North Africa rather than the Middle East, although it would be pertinent to establish Lebanese, Syrian and Yemenite haplotype 1 prevalence, as well as to conduct a second genetic epidemiological study of Saudi Arabia PD.

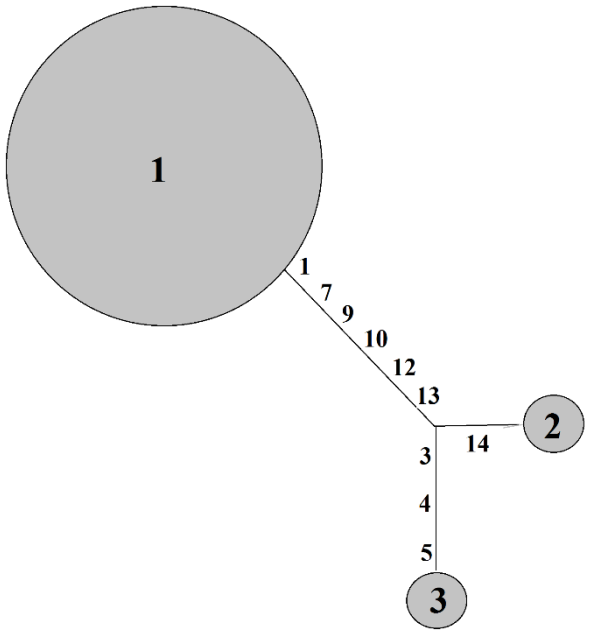


Figure 6. A haplotype network for haplotypes 1, 2 and 3. The SNPs that differ between each haplotype (as labelled in Table 2) are shown on the connecting lines.

3.4 Penetrance

Penetrance refers to the portion of cases in which a pathogenic mutation causes a clinical disease phenotype, with 100% indicating certainty of expression. From the time of initial genome-wide linkage analysis, incomplete penetrance at the PARK8 locus was suspected due to instances of the disease- associated haplotype in unaffected individuals. Penetrance varies among different mutations and variants in *LRRK2*. Penetrance of Gly2385Arg is low (Farrer et al., 2007), whereas that of Ile2020Thr is very high (Ho et al., 2016). Established

differences in penetrance are informative as they may provide insight into divergent mechanisms of mutant dysfunction.

Most penetrance estimates for PD *LRRK2* variants were performed on G2019S and have been highly variable, ranging between 24-100%. Earlier studies tended to have higher estimations (70-100%) (Funayama et al., 2002, Paisan-Ruiz et al., 2004), and were performed in large families with multiple affected members. Later studies utilised cohorts of PD patients not ascertained for familial history. These studies reported lower lifetime penetrance estimates (22-32%) and lower age-dependent estimates (2% at age 50 to 33% at age 80) (Goldwurm et al., 2005). Latourelle et al (2011) demonstrated that penetrance of *LRRK2* mutations in families with multiple affected members is substantially higher than in randomly ascertained idiopathic PD cases (67% versus 33% at age 85).

Sporadic cases of G2019S PD are noteworthy as they imply that some relatives were unaffected carriers, perhaps into advanced age. These carrier relatives could have genetic variation protective for disease, lacked additional risk variation, or avoided an environmental trigger. Neurologists have noted starkly variable age-of-onset, sometimes among family members. A portion of this variation arises from the fact that reported age-of-onset is a subjective measure. Trinh et al (2016) state the variance within families (median 56 years, IQR 47-65 years) is less than between unrelated carriers (median 57 years, IQR 40-74), which they state as evidence for the existence of genetic modifiers. Sample size in determining this variance was likely limited. Certain populations appear to display higher proportions of sporadic cases with no apparent family history of PD, as compared to those with family background of the disease (Lesage et al., 2006). Aside from ascertainment differences impacting estimates, other factors reputedly caused inconsistency in the penetrance data: small sample sizes, selection of statistical analysis, ethnicity and

environmental factors. It is suspected that longitudinal studies and further meta-analyses would be useful in terms of ascertaining more reproducible estimates of disease penetrance. Marder et al (2015) used the kin-cohort method to estimate penetrance of G2019S carriers in 2270 relatives of PD patients from New York and Israel, at 26% by age 80. It was noted that risk was almost three times as high for relatives predicted to be carriers compared to predicted non-carriers. Healy et al (2008) reported 28% risk at age 59 and 74% at age 79.

One relevant question regarding G2019S penetrance is whether it varies in different ethnicities. The current literature is suggestive of a similar penetrance between Ashkenazi Jewish PD patients and North Africans, the two populations with highest prevalence. Whereas research has highlighted lower penetrance for European Norwegian carriers (Hentati et al., 2014). Understanding of disease penetrance has value in the search for genetic modifiers, the design of clinical trials and genetic counselling. It is not currently possible to predict whether an unaffected G2019S carrier will develop PD in the future. Although *LRRK2* PD also exhibits an onset distribution very similar to that seen in idiopathic PD, it is notable that *LRRK2* PD appears to occur in females and males similarly, with some studies indicating a slightly increased penetrance in females (Cilia et al., 2014). This contrasts idiopathic PD where males are at greater risk (Wooten et al., 2004).

3.5 Modifiers of age-of-onset

Modifiers may act to inhibit the *LRRK2* toxic gain-of-function and unbiased genome-wide approaches, both linkage and association, are likely most relevant in identifying them. A number of studies have been conducted using similar techniques with various permutations, although none have been replicated independently. One overarching issue is likely one common to GWAS: the massive number of statistical tests performed causes great potential

for false positive results. Aspects of study design must therefore be closely attended to when reviewing studies.

Latourelle et al (2011) undertook a genomewide linkage and association study in white non-hispanic *LRRK2* mutation carriers, with the aim to identify modifiers of *LRRK2* PD penetrance. Two regions (1q32.1 and 16q12.1) were flagged for high LOD score in the linkage analysis, while no statistically significant SNPs were identified in subsequent association analysis (which were performed in areas flagged by linkage). Utilising multiple *LRRK2* mutation carriers can strategically increase sample size; Latourelle et al (2011) achieved a sample of 99 cases from 59 families for the association analysis. Latourelle et al (2011) perform adjustments for relatedness. Care would likely also need to be taken in ensuring LRRK2 mutations had similar age-at-onset penetrance distributions or were case/control matched for this variable, which Latourelle et al (2011) do not refer to. The most significant region identified through LOD score (1q32) was implicated in two GWAS for iPD, which the authors propose supports the hypothesis that PD risk factors are implicated in AAO heritability, and/or the reverse notion that genetic modifiers of *LRRK2* PD play some role in idiopathic disease.

Whereas Trinh et al (2016) focus efforts initially on a more homogeneous patient subset of North African G2019S carriers. Discovery based linkage analysis and subsequent association analysis were also used. A locus at *DNM3* that has association with AAO was identified in linkage analysis by a high LOD score. Subsequently three significant SNPs in *DNM3* (rs742510, rs2421947 and rs2206543) emerged in association analysis. Only rs2206543 was significant when using AAO as a quantitative trait. One caveat is that using AAO as a quantitative variable imposes a particular ordering on the categories, and specifies the relative

size of effects, which may not be present in the data. As described previously there may be uncertainty/subjectivity in estimates of AAO or in the distribution of the outcome data. In the qualitative association analysis a lower threshold than the generally accepted $5.0x10^{-8}$ was used ($p_{nominal}$ = 2.6 x 10-5), although this threshold is not an absolute (Fadista et al., 2016). The association analysis was, however, comparatively well-powered with 232 unrelated individuals.

Median AAO for *DNM3* rs2421947 GG homozygotes with positive carrier status for G2019S was 51.5 years, whereas for CC homozygotes it was 64 years. CG heterozygotes had AAO of 57 years. Confidence intervals around these estimates are not shown. These denominations are also quite small compared to the overall distribution of AAO in G2019S PD; smaller effect sizes require larger sample sizes and are more challenging to establish. Additionally, first motor symptom onset as mentioned previously is to an extent subjective and could be impacted by recall bias. PD is a disease with a lengthy prodromal phase (Pellicano et al., 2007) including symptoms such as sleep disorder: this also calls into question the validity of using age of motor symptom onset as a definitive quantitative trait. Trinh et al (2016) subsequently present evidence that in striatal brain tissue the *DNM3* expression varied as a function of the rs2421947 genotype. Also evidence is provided that *DNM3* is perturbed in *LRRK2* G2019S neurons. Although this adds another layer of evidence, it is of primary importance to replicate the *DNM3* genetic association independently. Replication was performed by Trinh et al (2016) on G2019S carriers from Algeria, France, Norway and North America. The sample size for Norwegian samples was smaller, with 64 G2019S carriers. Trinh et al (2016) note that these were convenience based samples containing only PD patients, and may suffer from ascertainment bias.

In addition to the potential *DNM3* modifier, Trinh et al (2016) also note a higher LOD score at the *LRRK2* locus, perhaps indicating the trans haplotype impacts AAO. Significant scores were obtained for chromosomes 17q25.3 and 21q21.2. However, evidence for these three linkage results was not found in subsequent association analyses. It is notable that neither Latourelle et al (2011) nor Trinh et al (2016) flag the same regions in their analyses.

Botta-Orfila et al (2012) provide evidence that *LRRK2*-Associated PD is modified by *SNCA* variants, which Trinh et al (2016) have stated that they have been unable to independently replicate. Some studies (Golub et al., 2009, Ziv Gan-Or et al, 2011) also presented evidence of modification by common variants in MAPT. Golub et al (2011) found AAO modification by a variant (rs2435207) in the MAPT gene region, although sample size was small at 44 PD patients from 19 families. In the study by Ziv Gan-Or et al (2011) the confidence intervals around AAO for carriers of different MAPT genotypes overlap significantly, highlighting the same issue described previously. Variants were selected in this study in advance (on the basis of published and unpublished GWAS, and using the SNP found associated by Golub et al (2011)). The SNP is not highly significant and not precisely reported ($p < 0.03$). Trinh et al (2016) also state they were unable to replicate the MAPT association with AAO.

Lubbe et al (2016) presented a paper assessing the level of additional rare variants of unknown significance in PD genes among cases and controls; the percentage of cases with additional variants were significantly higher (although not withstanding correction for multiple testing) at 33.33% in Known Mutation-PD compared to 15.4% of controls with additional variants. Similar results were replicated in NeuroX data from the same group. Lubbe and colleagues (2016) found that *ATP13A2* variants seemed enriched in NeuroX *LRRK2*-positive cases. The enrichment also occurred in the exome cohort. The effects of

ATP13A2 additional variants on AAO was only assessed using NeuroX data and no statistical difference was observed, indicating more data is required.

4 Added value of this study

This study aims to validate the modification factors evidenced elsewhere (*LRRK2* trans haplotype and *DNM3*), in a separate multi-ethnic cohort. Independent validation in the same and different populations will be essential in both confirming genetic modifiers and establishing how these vary across populations. This study will utilise a cohort of European Caucasian, Ashkenazi Jewish, and North African G2019S carriers in targeting genes already flagged for association with AAO. Additionally, discovery-based association analysis for common variants in genes implicated in neurodegeneration more broadly will be used.

Trinh et al (2016) state that it would be prudent to consider introducing *DNM3* in genetic testing for North African G2019S patients. However, caution should be taken in this regard, considering the overlap of onset age in carriers of different rs2421947 alleles. Additionally GWAS have been fraught with false associations, highlighting the need for independent replication.

If it could be predicted how a patient will progress, then the impact of therapeutic intervention could be more effectively parsed in clinical trials. Genetic modifiers could also, at some stage, enter clinical practice through screening and genetic counselling. Further exploration of G2019S prevalence and G2019S haplotypes could improve understanding of the evolutionary origins of the disease.

5 Power calculation

Power needed to find an effect was calculated, as described in the methods section below. The calculation indicated that for a modifier with very strong effect size then 100 EOPD cases and 100 LOPD controls would be fairly well powered for a GWAS logistic regression analysis to reach a $5x10^{-8}$ significance level (shown in Table 3). However, this may be over-optimistic, especially considering the lack of concordance in results for previously published genome-wide studies; the discordancy may indicate that such a large effect modifier variant does not exist.

Table 3. Power calculation results

	Common variant/ strong effect	Common variant/ weaker effect	Relative rare variant/ moderate strong effect
Disease allele frequency	0.3	0.3	0.15
Genotype relative risk	1.75	1.2	1.5
Expected power for a one-stage study	0.77	0.00	0.034

During the course of this project PD patients and UR were actively recruited and characterised, meaning final sample size was unknown.

6 Methods

Power calculation

Power calculations were performed in GAS Power calculator (http://csg.sph.umich.edu/abecasis/cats/gas_power_calculator/index.html). An additive model was assumed. Power calculations were constructed to find an effect with significance level of $p = 5x10^{-8}$ in 100 EOPD G2019S and 100 LOPD (and unaffected proband G2019S carriers) using logistic regression. Prevalence of EOPD in G2019S PD carriers was estimated at 50% in calculations, for the purposes of the model.

6.1 Cohorts

Data from three cohorts were included in this study, with details of the cohorts described in Table 4. All participants gave their informed consent to have their biological samples used in clinical research.

Table 4. Cohort characteristics

	Truseq cohort	**Reseq cohort**	**Exome cohort**
Recruitment & data type	Data collected during the course of this project, and continuing afterwards; NGS sequenced on Illumina targeted neurodegeneration panel during this project	NGS data already available; Illumina custom-made panel targeting PD genes	NGS data already available; exome data
Sample size	123 samples	3337 individuals NGS data (n = 2137 PD cases; n = 1161 controls)	~1000 NGS exomes
Ethnicity	North African, Ashkenazi Jewish, European Caucasian	European	Not available
Country of data collection	Not available	United States (n = 1777), United Kingdom (n = 1157), Germany (n = 253), the Netherlands (n = 111)	Predominantly UK and n = 13 exomes from Greece
Studies	TRCN	Not available	PROBAND Tracking Parkinson's study, Parkinson's Families Project, Molecular Studies of Neurodegenerative Disorders study, PDDNAP, Queen's Square Brain Bank, and others (unknown)
Inclusion criteria	PD patient or unaffected relative	PD patient with no known mutation in PD gene	Various depending on the study samples were obtained from. PROBAND, PFP, and PDDNAP specified early onset (≤45) or familial PD.
Previous known publications	None/unknown (some samples may have been included previously)	NeuroX array (Nalls et al., 2015)	Unknown

Pseudonymised clinical data on the cohorts was provided by Dr Mie Rizig, Hallgeir Jonvik (Institute of Neurology data manager), and Manuela Tan, from records collected at the UCL Institute of Neurology (IoN), or databases that had been shared with IoN researchers. Informed consent was given by participants for access to records for the purposes of clinical research.

7 DNA preparation and Sanger sequencing

Samples and reagents were handled with sterile technique throughout to prevent contamination.

Genomic DNA was extracted from blood or saliva from participating individuals:

1. **DNA extraction from blood**

   Blood was collected in EDTA (ethylenediaminetetraacetic acid) bottles. Genomic DNA was extracted from whole blood in the diagnostic genetics laboratory at the UCL Institute of Neurology using a FlexiGene© kit (Quiagen) according to manufacturer instructions.

2. **DNA extraction from saliva**

   Genomic DNA was extracted from saliva at a laboratory in Germany using standard methods.

7.1.1 DNA quantification and qualification

Two quantification methods were used in order to ensure that the samples were of an adequate concentration for sequencing and in order to perform dilutions for normalisation: Qubit and Nanodrop. All samples underwent Qubit quantitation as this method gives a more accurate reading of quantity than the Nanodrop technique. Nanodrop was used on a supplementary basis for assessment of contaminants and DNA quality.

7.1.2 Qubit quantification

Genomic DNA samples were run on the Qubit 2.0 fluorometer (Life Technologies, UK). The Qubit fluorometer detects the amount of fluorescent dye bound to DNA (or RNA) and in this way provides a direct measurement of DNA quantity. Both broad range and high sensitivity specifications were used. Broad range quantification was used when expected concentration was larger, whereas high sensitivity was used for very low concentration DNA samples. Samples were initially standardized to 50ng/µl and diluted to 10ng/µl DNA.

7.1.3 Nanodrop quantification and quality control

Genomic DNA was measured using the NanoDrop Spectrophotometer equipped with ND-1000 software. The NanoDrop measures the absorbance of the sample at 260nm wavelength (the wavelength DNA absorbs at); greater absorbance occurs at greater quantities of nucleic acids. The ratios of 260/280 nm and 260/230nm were used to assess purity of DNA. Ratios of 260/280 = ~1.8 and 260/230 = 2.0-2.2 indicate pure DNA. Significantly lower ratios indicate the presence of contaminants. The NanoDrop spectrophotometer was loaded with 1ul of

sample and the DNA concentration was expressed in ng/µl. The Nanodrop was used during optimization and troubleshooting.

7.1.4 Polymerase chain reaction (PCR)

This technique was used to specifically amplify, by many orders of magnitude, two regions of the genome, in order for these regions to be Sanger sequenced. The first region contained exon 41 and the G2019S mutation. The second region contained a haplotype tag rs2206543 in *DNM3*. This required the design of two pairs of primer sequences that encapsulate the region of interest and fulfil a number of functional prerequisites. The process of PCR is summarised in Figure 7.

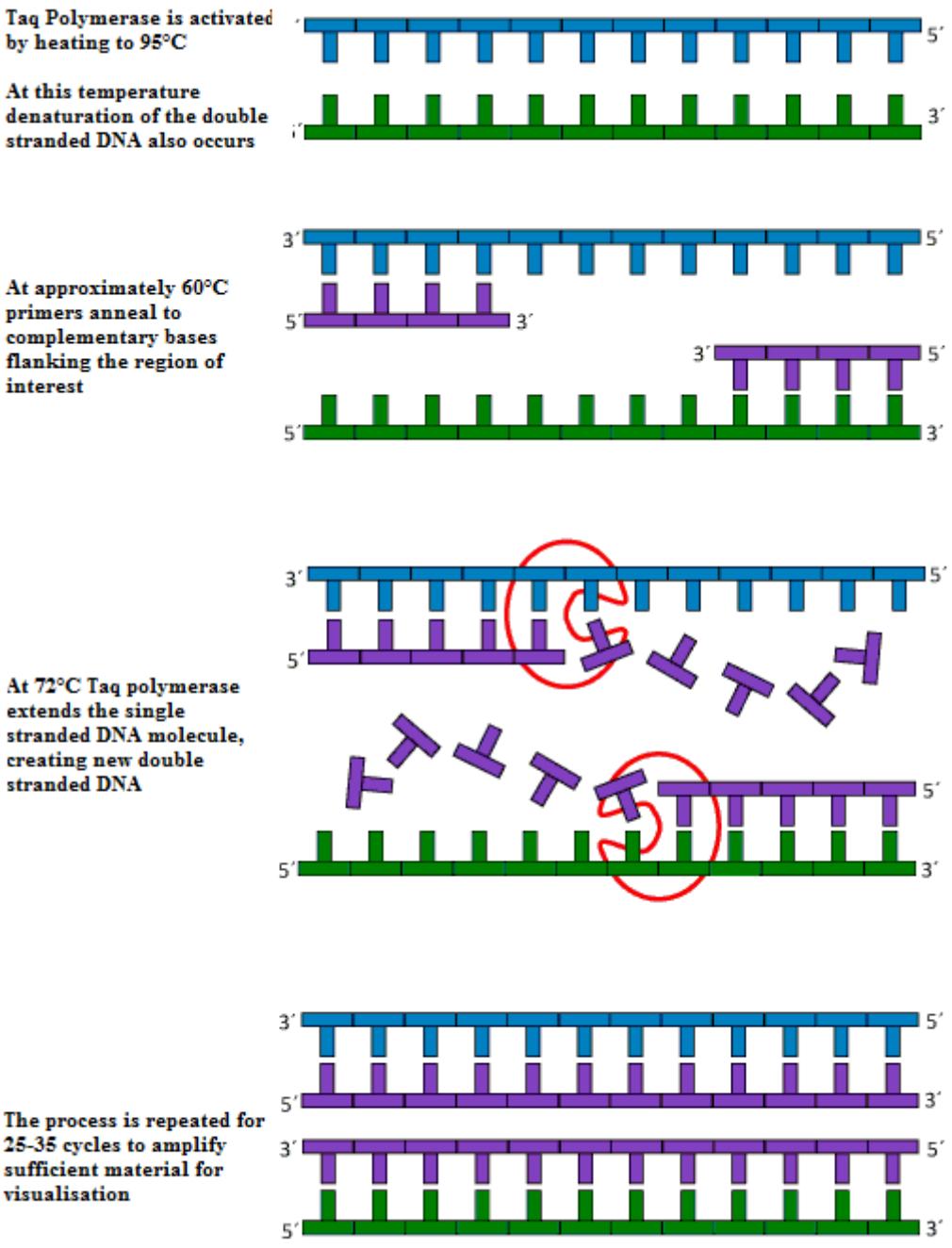


Figure 7. Diagram showing the process of PCR amplification. Adapted from (http://www2.le.ac.uk/departments/emfpu/to-be-deleted/explained/images/PCR-process.gif/view)

7.1.5 Primer design

The Primer3 online software was used to design oligonucleotide primers for PCR, sequences of which are shown in the Appendix. Primer3 software uses a variety of parameters and thermodynamic modelling to predict useful primer sequences. Input regions (areas intended for amplification) were obtained from the reference sequence on the UCSC Genome browser website (https://genome.ucsc.edu/index.html). The resulting primer sequences outputted by Primer3 were checked using the BLAT Search Genome tool on the UCSC website to check specificity for the desired location. Primers were selected based on a number of criteria known to affect functioning: sequence specificity, primer melting temperature of ~60-64°C, a product size of ~500, 40-60% GC content, and the absence of common SNPs and repeating elements within either primer sequence (as shown in Figure 8). Rs2421947 *DNM3* was refractory to amplification by this method, due to the large number of repeating elements, which likely precluded binding; instead primers for a SNP (rs2206543) in high linkage disequilibrium with rs2421947 were selected.

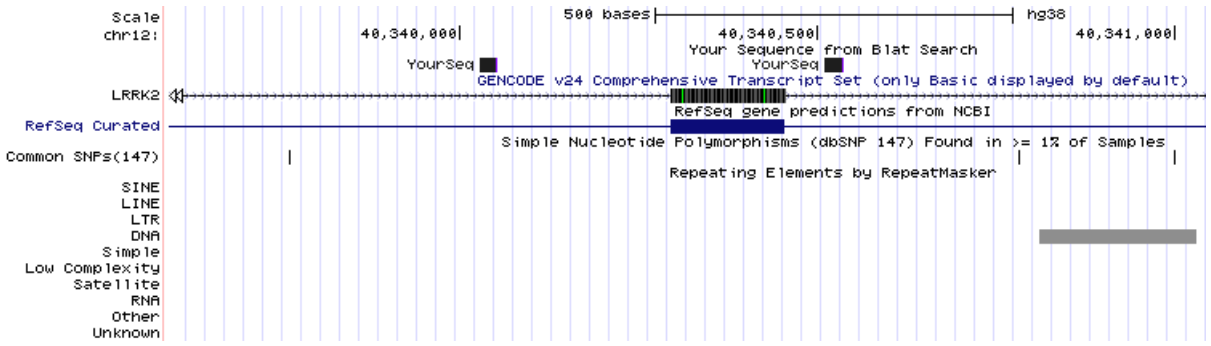


Figure 8. UCSC genome browser BLAT search output. Primer sequences ("YourSeq") are shown to encapsulate exon 41, which contains the G2019S mutation. There are no common SNPs or repeating elements (SINE/LINE) within the primer sequences.

7.1.6 SNP Annotation and Proxy Search (SNAP)

SNAP hosted by Broad Institute Version 2.2 was used to find a SNP in LD with rs2421947. The R^2 values of high LD SNPs were evaluated for different populations (MKK, CEU, JPT + CHB) using the SNP dataset HapMap3 (release 2). SNPs were required to have R^2 greater or equal to 0.8 with rs2421947 in CEU, MKK and JPT+CHB populations in order to be used as a surrogate. The steps outlined for primer design were then followed. Rs2206543 was the only SNP found to meet both R^2 value and primer design requirements. Figure 5 shows the R^2 values outputted by SNAP for rs2206543.

Table 5. R2 values with rs2421947 from HapMap3 (release 2) for different populations

	JPT+CHB	MKK	CEU	YRI	ASW	CHD	GIH	LWK	MEX	TSI	CEU+TSI	JPT+CHB
rs2206543	0.982	0.806	1.00	1.00	1.00	1.00	1.00	0.843	1.00	1.00	1.00	

7.1.7 Polymerase chain reaction (PCR)

PCR was performed using Roche Fastart PCR Master (400RXN/10ml) (Roche Applied Sciences). Figure 5 shows the reagent and DNA quantities added to each well of a 96 well plate.

Before use, aliquots of 3ng/µl primer concentration were made from the stock solution, which was standardised to 100ng/µl, in order to prevent repeated degradation by thawing.

Table 6. Reagents per well

	volume (µl)	concentration (ng/ µl)
Faststart Master mix	12.5	N/A
Forward primer	2.5	3.0
Reverse primer	2.5	3.0
Wash buffer (H_2O)	5.0	N/A
Sample DNA	2.5	10.0

After centrifugation of the above solution, heating on a thermal cycler was used to denature the double-stranded DNA into single strands. Subsequent cooling allowed annealing of the primers to the DNA template. DNA polymerase begins synthesizing new strands of DNA starting from the primers. Stages of denaturation, annealing and synthesis repeat many times to exponentially amplify the concentration of the DNA target. The thermal cycler was set with 65 to 55 touchdown programme with hot start. A hot start thermocycler option is used because the Taq polymerase in the FastStart PCR Master mix has been chemically modified so that it has no activity up to 75°C. This is to ensure that non-specifically annealed primers are not partially extended by the DNA polymerase during the ambient temperatures of PCR preparation. It requires a pre-incubation step of 95°C, 2-4 minutes (hot-start) in order to initiate its activity.

The touchdown programme starts at a temperature higher (65°C) than the optimal melting temperature (Tm) of the individual primers (which varies depending on numerous factors; calculations are only an approximation). It is then reduced over subsequent cycles (0.5°C reductions were selected for this experiment; smaller reductions allow the temperature closest to the unknown primer Tm to be reached and binding to occur).

7.1.8 Agarose gel electrophoresis

Agarose gel electrophoresis was used to quantitate the PCR product and verify it had the correct molecular weight. 5x TBE (tris-borate-EDTA) solution was prepared using 121.1g of Trizma base (Sigma), 61.8g Boric Acid, 7.4g of Ethylenediaminetetraacetic acid (EDTA) (Sigma), and dissolved in 1 litre of distilled water. Subsequently a one in five dilution was performed to bring the concentration to 1x TBE. A 2% gel was prepared using Ultrapure Agarose (Invitrogen) and TBE 1x buffer stained with 10µl of gel red (Cambridge bioscience). 3µl of PCR product and 3µl of x6 Orange DNA loading dye (Thermo scientific) were loaded into each gel well. A DNA ladder (Midrange 100-2000 bp) (Qiagen) was inserted into the two furthest wells of each row. The gel was run at 120mV for 30 minutes. DNA fragments are visualized using a UV transilluminator. Where bands appeared at appropriate positions relative to the ladder, the PCR product was progressed to Sanger sequencing.

7.1.9 Exosap clean-up

Exosap solution was made with Fast-Alkaline phosphatase (Thermo scientific), which removes unused dNTPs, and Exonuclease I (Thermo scientific), which removes ssDNA from PCR products. The enzyme mix is prepared 1ml at a time, in order to minimize freeze thaw degradation. It is then stored at -20°C. The procedure is shown in Figure 9 overleaf.

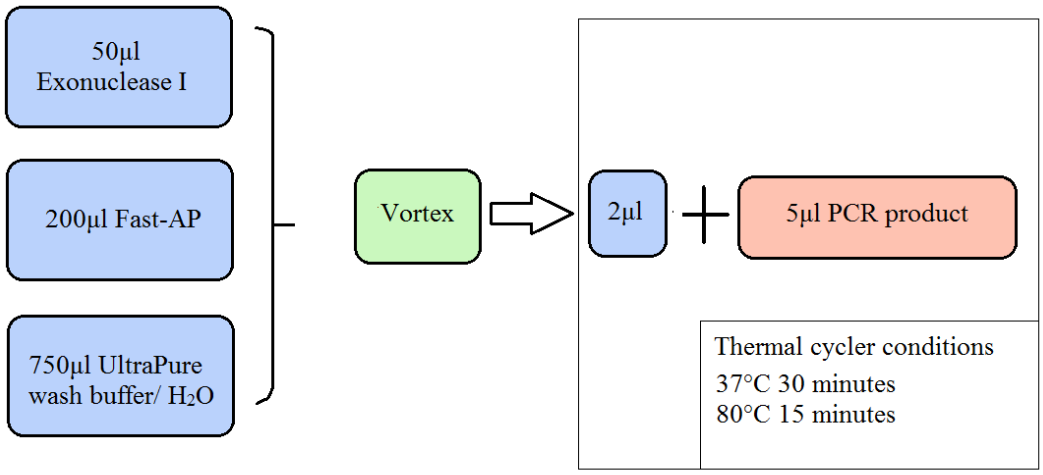


Figure 9. PCR purification protocoll. The ratio of 2µl of Exosap to 5µl PCR product was scaled up as required.

7.1.10 BigDye sequencing

BigDye® Terminator v3.1 Cycle Sequencing Kit was used in the sequencing reaction as shown in Figure 10. For G2019S, the reverse primer was used, as this was found to produce the clearest sequencing for the G2019S mutation (due to its position within the sequence). For rs2206543 the forward primer was used for clearest results. The standard program recommended by Applied Biosystems (as shown in Figure 10) was used to run the Sequencing reaction in the thermal cycler. Where results were ambiguous, the alternative primer was used in a separate reaction or the sequencing was repeated.

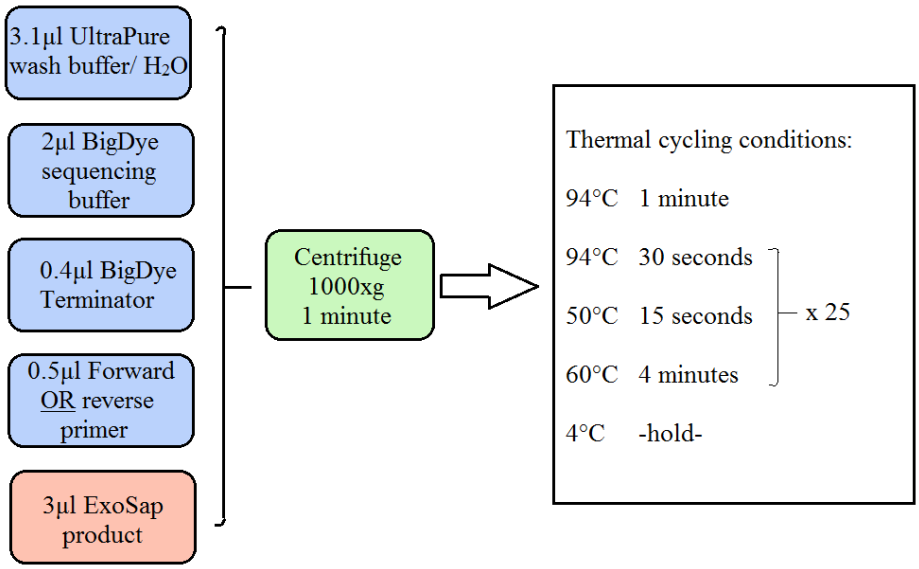


Figure 10. BigDye sequencing protocol

7.1.11 Sephadex purification

2.9g of Sephadex G-50 powder (Sigma-Aldrich) was added to 40ml of distilled water. It was mixed well and allowed to hydrate for at least 30 minutes at <4°C, before being mixed again immediately before use. The solution was reused for a maximum of one week when stored in the refrigerator. 350µl of Sephadex solution was added to each well of a Corning FiltrEXTM 96 well filter plates (0.66 mm glass fibre filter). The Corning FiltrEXTM filter plate was placed on top of an empty collection plate and centrifuged for 3 minutes at 750xg. The Corning FiltrEXTM was then placed on top of an unused 96 well non-skirted plate. The entire contents of the sequencing reaction were pipetted onto the Sephadex columns, taking care not to touch the Sephadex with the pipette tip. The plates were then centrifuged for 5 minutes at 910xg.

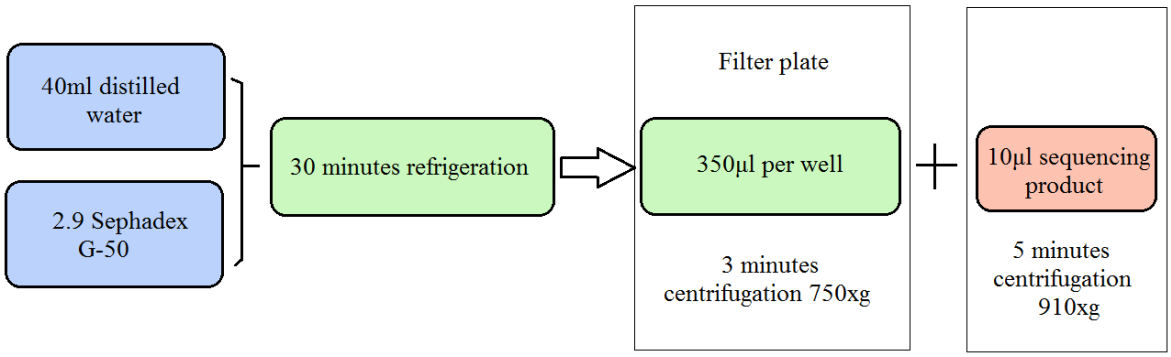


Figure 11. Sephadex filtration protocol

7.1.12 Sequencing 3730 DNA analyser; electropherogram visualisation

Sanger sequencing was performed on a 3730 DNA analyser (Applied Biosystems, Foster City, CA, USA) and electropherograms were visualised using CodonCode Aligner (CodonCode Corporation, MA, USA, version 7.0).

Samples identified as G2019S through the described method were subsequently sequenced for rs2206543.

8 High-throughput techniques

G2019S carriers identified through Sanger sequencing were advanced to next-generation sequencing using the following techniques.

8.1.1 Illumina Truseq Neurodegeneration panel

This project involved the beta-testing of the Truseq neurodegeneration panel by Illumina; the panel is a targeted sequencing platform that includes genes involved in neurodegeneration

(including Alzheimer's disease, Parkinson's disease, and Amyotrophic Lateral Sclerosis). It targets 118 genes with over 8.7Mb of content, including exons, introns, untranslated regions (UTRs) and promotor regions. The panel defined the regions for sequencing data capture.

8.1.2 Nextera Rapid Capture Enrichment for targeted sequencing

Samples were processed first using protocols described in Illumina Nextera Rapid Caputure Enrichment Reference Guide #15037436 v01 January 2016. It was important beforehand to accurately quantify the DNA using Qubit to 5ng/µl from 10ng/µl in a two-step dilution. This is because the fragmentation step is enzymatic and particularly sensitive to the input DNA quantity. While the DNA is fragmented enzymatically through "tagmentation", adapter sequences are added to the end of each sheared DNA. Unique indexes for multiplexing, sequencing primer binding sites and sequences which bind to flow cell oligonucleotides are then added. Probes then enrich areas of interest (in this case, neurodegeneration associated regions specified by the Truseq panel), in a process called "capturing" (Head et al., 2014).

The libraries are then qualitatively and quantitatively assessed using Qubit (described previously) and Agilent bioanalyser, to ascertain whether these stages have occurred as expected. Agilent High sensitivity DNA kit with 2100 expert software was used in this stage. The Bioanalyser is a microfluidics platform which allows sizing, quantification and quality control of DNA and RNA. After loading the chip as per the instruction manual (gel-dye in 4 wells, ladder and marker in one well, sample and marker in the remaining wells), and running the software, the sample moves through microchannels and sample components are electrophoretically separated, with smaller fragments migrating faster than large ones.

Fluorescent dye molecules intercalate with DNA strands and are then detected by their fluorescence. Visual inspection of the resulting trace was used in assessing libraries.

8.1.3 Hiseq machine

Processed samples were run on the Illumina HiSeq 3000, a sequencing machine that incorporates SBS chemistry and patterned flow cell technology, at the Institute of Neurology by Debbie Hughes.

9 Bioinformatic processing of data

9.1.1 Pre-Processing and quality metrics/control

After collecting binary base call format sequencing data (.bcl) from the Hiseq 3000 machine, further processing via a bioinformatics pipeline was required before variants could be identified and downstream analyses performed.

Raw sequence data in base call format was converted to FASTQ files using bcl2fastq tool by Dr Alan Pittman. FASTQ retains the confidence scoring calculated with Bustard. This quality score (Q-score) is calculated as

$$Q = -10 \times -10 \times \log_{10} P_{Err}$$

P_{Err} is the probability of making a base call error. Base calls need a Q-score of at least 20 to be considered reliable. High quality scores are 30-40. Where data is merged from different platforms it is valuable to recalibrate Q-scores through the use of a subset of reads that map

to regions of the reference genome that contain no SNPs. When SNPs are found where no SNPs are known to be, these are used to construct a new calibration table.

Quality control then requires examination of a number of sequence reads quality metrics. Q-scores are examined on a per-base basis across all reads. Base reads at the beginning of a sequencing procedure tend to have higher Q-values than those sequenced later, although latter sequenced bases should still have a median value of at least 20. If there is a significant Q-score drop in the late phase, affected base positions may require close examination and low-quality bases should be trimmed from affected reads. Increased percentage of N calls (where no base has been read) can also be used as an indicator of base call quality. The average Q-score of each read was plotted and the distribution pattern inspected. For a successful run most reads Q-scores should be >30, with only a very small percentage of reads with an average Q-score below 20. NGS sequencing files must then be processed to filter out low-quality reads and trim off portions of reads that have low-quality base calls.

The percentage of each base across base positions is also informative: the plots of A, C, G and T should be roughly parallel to each other, and the overall percentage shown in each plot should reflect the overall frequency of each base in the target library. Plots were found not to deviate significantly from this, indicating that there were not issues in the library construction process.

Although Illumina carries out certain filtering by default, inspection of parameters may flag the need for additional filtering. Quality was assessed using the FastQC program (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). FastQC runs a series of tests on the fastq file to generate a comprehensive quality control report. FastQC assesses data quality

by evaluating: read length, per base quality score, per sequence quality score, GC content, nucleotide content, sequence duplication and overrepresented sequences.

FastQC represents the first stage in the pipeline prior to association analysis, as shown in figure 5. After producing graphs in R for various quality metrics, a pipeline developed by Dr Alan Pittman was used to perform the following stages.

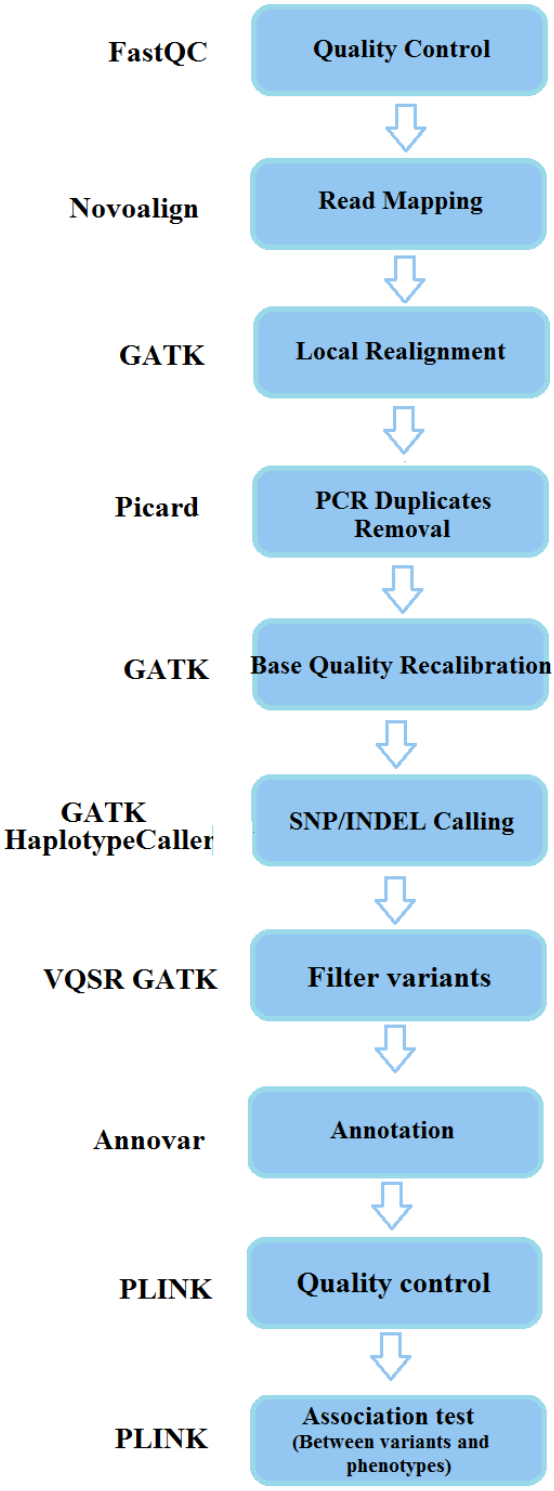


Figure 12. NGS data processing workflow, based on recommendations of best practice from GATK for processing of Germline SNPs and indels from whole genome or whole exome sequencing (June 2016)

Paired-end sequence reads were aligned with Novoalign software (www.novocraft.com) against the reference human genome (GRCh37 UCSC hg19, downloaded from UCSC genome browser). Novoalign provides higher sensitivity as compared to other aligners such as BWA and SOAP2 (Wang, 2016).

GATK RealignerTargetCreator and IndelRealigner were used in conjunction in a two-step process to overcome incorrectly mapped regions due to indels: RealignerTargetCreator determines small intervals which likely require local realignment; IndelRealigner subsequently realigns the intervals.

Duplicate reads, which are a PCR artifact, were removed using Picard software package (http://picard.sourceforge.net/), to prevent them from influencing variant calling. Markduplicates feature was used to find duplicates that arose through PCR and tag these so that they are not included in subsequent analyses. A feature which also marks optical duplicates, in addition to PCR duplicates was used. Reads were then organised in chromosomal order by Samtools.

Original base-call quality scores are recalibrated again at this stage, using GATK BaseRecalibrator which recalibrates raw quality values using a covariate based recalibration algorithm. The algorithm takes machine sequencing cycle and local sequence context (which are known to affect sequencing signal and base-call quality) as covariates. The covariation pattern is first analysed and examined (by the program) and then applied to recalibrate the data. This stage increases accuracy and cuts down the number of false positives.

GATK Haplotypecaller was used for variant calling. Haplotypecaller uses statistical modelling to model errors and biases, and sometimes to incorporate other related prior information. It then considers the linkage between nearby variants and calls SNPs and indels simultaneously, thus performing local *de novo* assembly of haplotypes.

The following programs are only mentioned briefly here, as the operation and output in the context of this project, is more relevant to the results.

9.1.2 Variant Quality Score Recalibration GATK

This program used machine learning to create models for the purpose of determining and removing variants that are likely to be false positives.

9.1.3 PLINK Quality control (QC) and Logistic regression association analyses

PLINK is a whole genome data association toolset. It was used in quality control of the data (excluding samples and variants that have potential to introduce bias), and in performing association analysis using a variety of functionalities, which were called from the command line using Linux. The VCF was prepared for PLINK by annotation with dbSNP using GATK VariantAnnotator. This function annotates variant calls based on context rather than functional annotation. Vcftools converted the VCF to PLINK format. The X and Y chromosomes were then split using PLINK, as PLINK requires data in this format. Clinical data was then added to the files. The implementation of QC is described in more detail in the results section.

Data outputted by PLINK association analysis was visualised using qqman (R package) and R Studio. Qqman is a package that is used for drawing manhattan plots and Qq plots from GWAS studies.

10 Kaplan Meier survival curve and Cox proportional hazards model

The survival package in R was used in performing Kaplan Meier survival analysis and Cox proportional hazards model. Briefly, survival analysis is useful where data takes the form of time-to-an-event (in this case, developing Parkinson's disease). Missing data (where the event had not occurred) were right censored.

11 Haplotype visualisation and analysis

11.1.1 Haploview

Haploview (Broad Institute, MA, USA) was used to generate Linkage disequilibrium heatmaps and to identify tagging SNPs, which are SNPs most adept at capturing the surrounding region. The tagging SNPs were used to generate a minimal haplotype.

11.1.2 Excel

Haplotypes were characterised in Microsoft Excel using a "back-of-the-envelope" method. This involved highlighting homozygous sites and phasing the haplotypes by visual inspection to determine the common pathogenic haplotype. Trans haplotypes were identified by subtraction of pathogenic haplotype from cells containing both alleles. Trans haplotypes were grouped into categories based on shared sites.

12 Results

12.1 G2019S carrier identification and cohort characterization

Bands in agarose gel electrophoresis were at the expected positions for the size of the fragments produced in PCR for both G2019S and rs2206543, as shown in Figure 1. 48 carriers of G2019S were identified through Sanger sequencing. When carriers were progressed to next-generation sequencing and bioinformatically analysed 41 (85.4%) were called as G2019S through the panel. Sanger sequencing, which has greater accuracy, should not yield false positives unless sample swap or contamination has occurred. 2 of the discordant samples had a positive clinical diagnostic report. The discordancy may also have arisen from inaccurate base call during NGS. Discordant samples will be Sanger sequenced again in order to establish the true allele at the G2019S locus for these samples, as the cause is currently undetermined.

Truseq data was therefore available for 41 G2019S carriers. 1 proband was excluded as the individual had left the study and requested data use be discontinued. 13 G2019S carriers were identified in exome data, and 22 were identified in PD-resequencing data, using a linux functionality to “grep” or extract the G2019S SNP from the data. Data for these carriers was then extracted from Institute of Neurology databases. Table 2. displays demographic and clinical characteristics of carriers and non-carriers from cohorts separately (and combined). Overall the greatest number of G2019S carriers for whom data was available for this project were Caucasian (n= 51), followed by Ashkenazi Jewish (n= 22), North African (n= 8), and unknown (adopted) (n = 1).

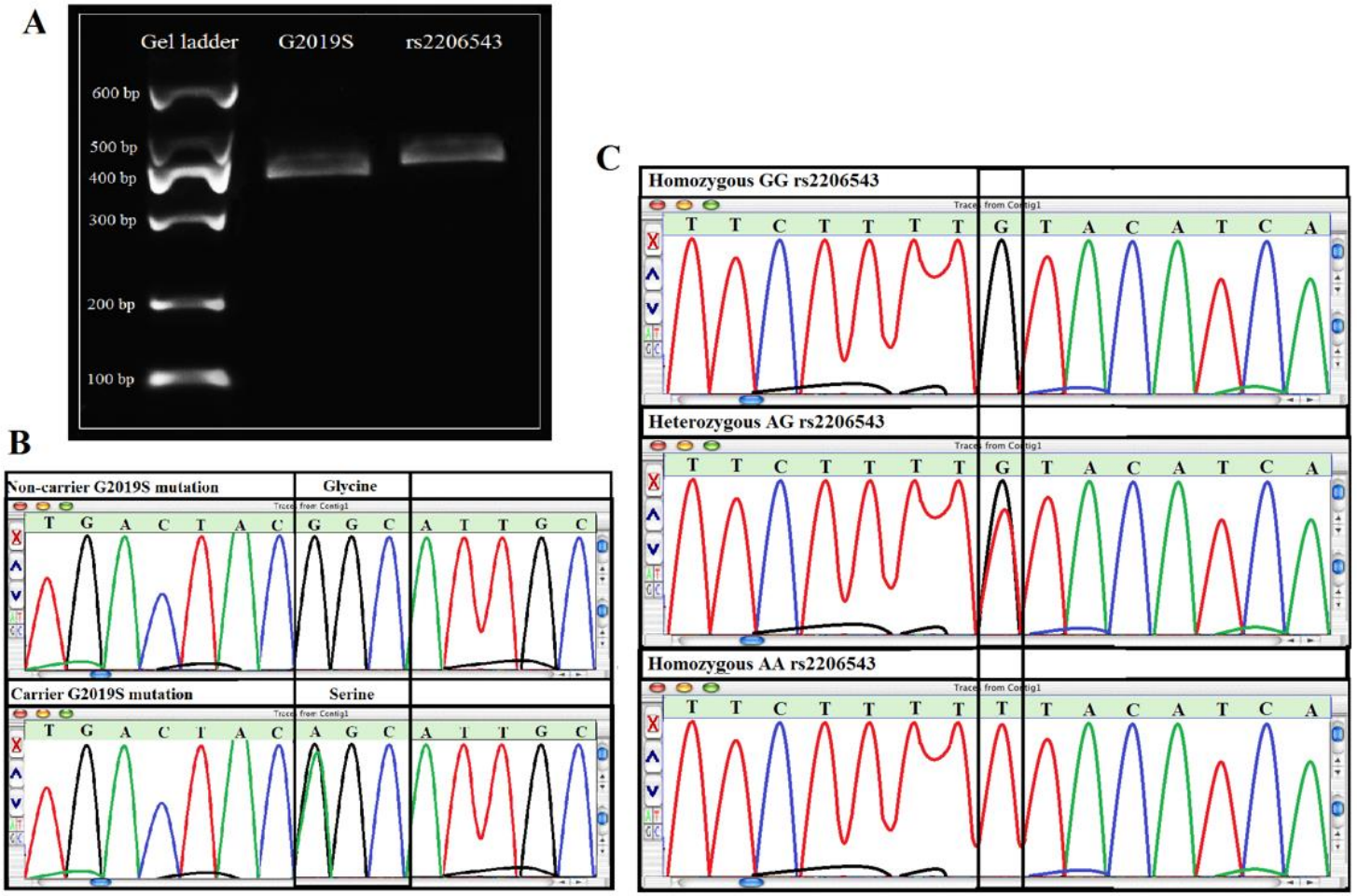


Figure 13. Sanger sequencing. A) Positions of fragments in agarose gel as expected. B) Electropherogram trace showing G2019S carrier and non-carrier. C) Electrophoregram trace showing homozygous GG, heterozygous AG, and homozygous AA at rs2206543.

Figure 14. Clinical characteristics of 158 study participants (83 G2019S, 75 non-G2019S) G2019S total includes n = 8 samples which have not been sequenced but are known to be G2019S through clinical Neurogenetics lab; these contributed AAO clinical data.

		G2019S sequenced (n = 40)	G2019S (exome) (n = 13)	G2019S (reseq) (n = 22)	Non-G2019S (n= 48, known; n = 27, unknown)
Parkinson's disease	n diagnosed	32	13	21	45
	n unaffected relative	8	0	0	3
	n unknown	0	0	1	0
Sex	n female (%)	20 (50)	3 (23.1)	10 (45.4)	22 (45.8)
	ratio female:male	(1:1)	(1:3.3)	(1:1.2)	(1:1.2)
Age	Sampling age	64	52.2	70.1	58.1
	Age at onset	53.7	44.5	60.6	45.7
Family history	n positive (%)	37 (90.2)	7 (53.8)	11 (50.0)	11 (22.9)
	n negative (%)	1 (2.5)	5 (38.5)	10 (45.5)	5 (10.4)
	n unknown (%)	2 (5)	1 (7.7)	1 (5.5)	32 (67.7)
Ethnicity	Caucasian	12	12	22	13
	Ashkenazi Jewish	20	0	0	5
	North African	7	1	0	0
	Asian	0	0	0	3
	Afro Carribean	0	0	0	2
	Turkish	0	0	0	1
	Unknown	1	0	0	24

12.2 Exploratory analysis of outcome variable

Prior to performing association analyses AAO histograms were constructed to determine the nature of the data. The bimodal distribution in A) of Figure 15 precludes use of linear regression analysis. A logistic regression analysis was considered appropriate with the data dichotomised into EOPD (57 and below), and LOPD (58 and above). The Kaplan Meier Curve median values (figure 16) were used in determining how to dichotomise the data. Cox Proportional Odds model, which does not require a normal distribution of outcome variable, was also selected as means for data analysis.

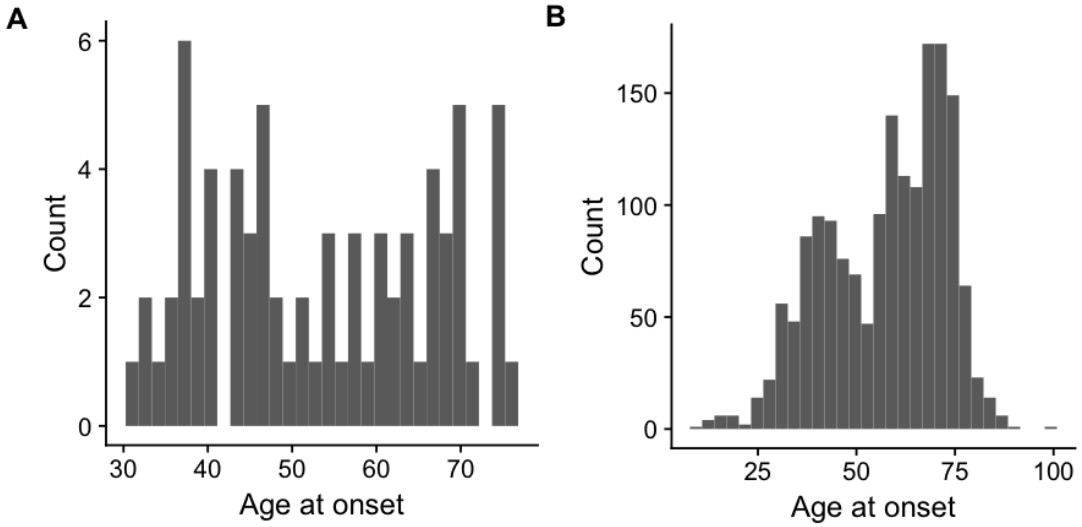


Figure 15. Age at onset histograms. A) AAO distribution for all G2019S carriers for whom data was available (n=72. C) AAO distribution for presumed idiopathic cases from resequencing data (included for comparison).

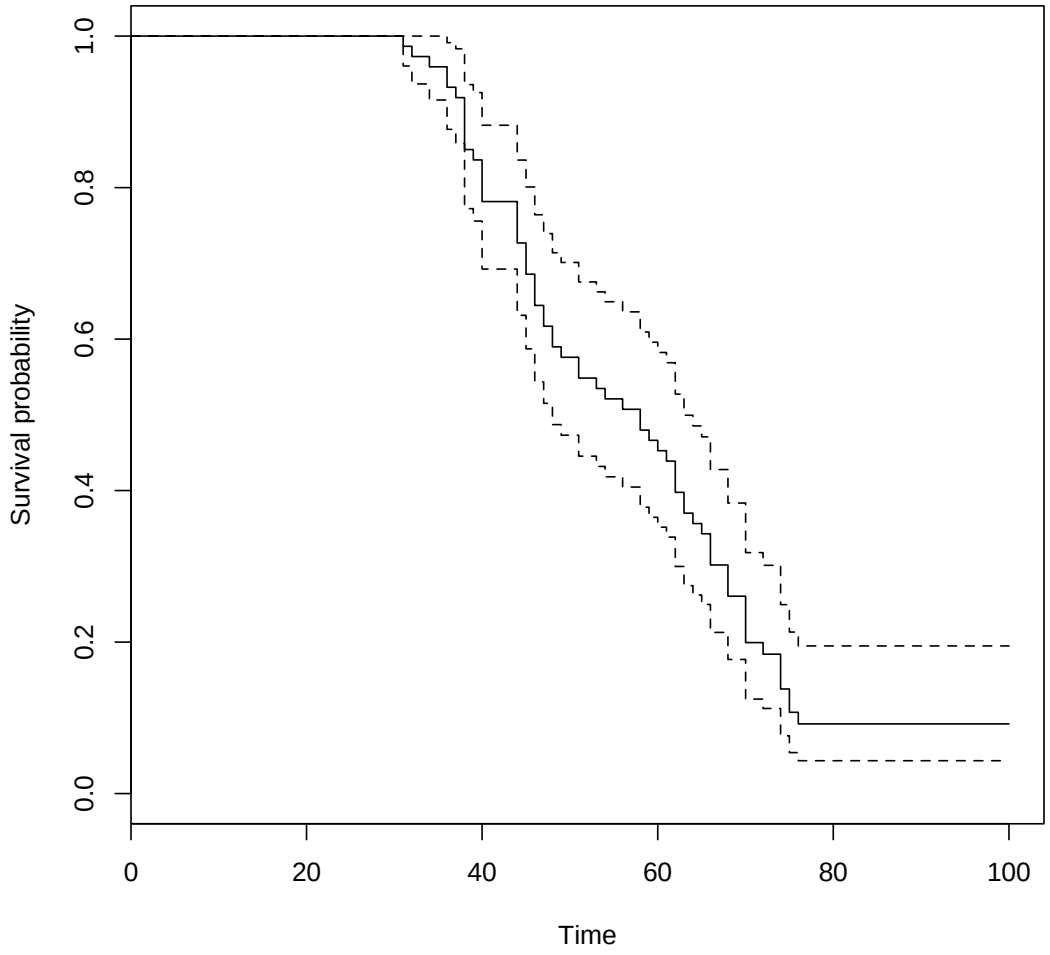


Figure 16. Kaplan Meier survival curve, with right censoring of AAO data from date of last examination.

12.3 Next-generation sequencing

12.3.1 NGS laboratory processing

After performing NGS Nextera Rapid Capture (described in section 9.1.2), the bioanalyser chip was used in checking the size distribution after fragmentation and the size distribution of the adapter ligated fragment library. Below is a sample image after

adaptor ligation; all images displayed plots in the expected range, and were progressed to the Hiseq. Following this, raw NGS data was available.

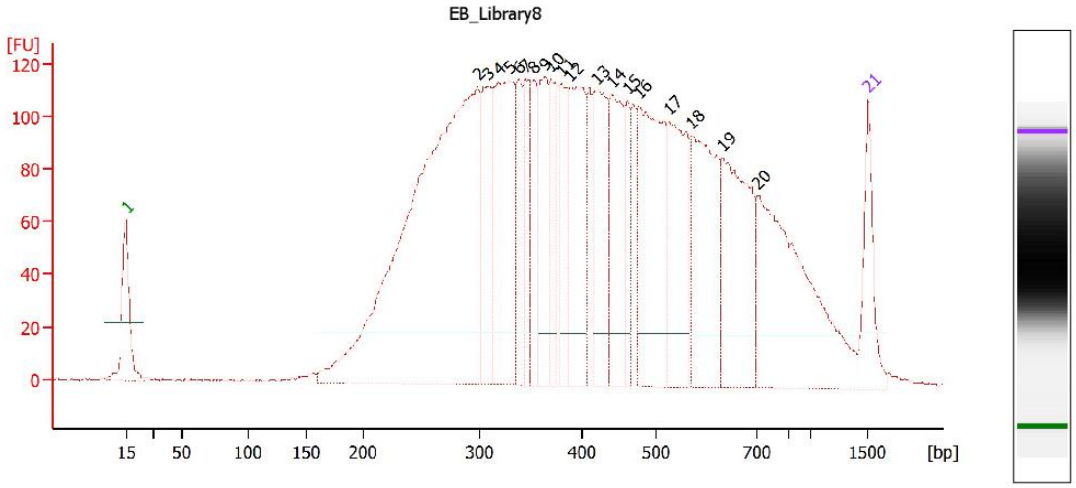


Figure 17. Bioanalyser image of library fragment sizes

12.3.2 Quality metrics and preprocessing

The early stages of NGS processing involve assessment of a number of key parameters. Coverage metrics are important as the NGS coverage level often determines whether variant discovery can be made with a certain confidence. Mean target coverage of Truseq Neurodegeneration data was high, indicating that the majority of the regions specified by the panel were targeted across samples. Another relevant metric is the percentage of the target covered by different numbers of reads, as multiplicative observations per base are required to achieve a reliable base call. If a read were to contain a 1% variant-error rate, then the combination of eight identical reads substantially increases accuracy in producing a variant call (Sims et al., 2014)

Generally, the level of coverage for SNPs that most publications require is from 10x to 30x depth of coverage, depending on the application and statistical model. The first two samples of Figure 18 were removed due to the low coverage at 10x. Low depth can mean that sequence errors are introduced and mistakenly propagated through downstream analyses as variants; this acts to misdirect conclusions. The remaining samples all had a minimum of 80% of the target region covered by at least tenfold as has been customary for exome studies. Population genetics may often use lower depth (i.e. 4X), but this is in cases where >400 samples are used and variants are called concurrently.

Figure 18 shows the number of reads that were eventually aligned to the reference contrasted with the number that were lost as duplicates or did not match the target region. Overall the Truseq panel yielded a high number of variants shown in Figure 19.

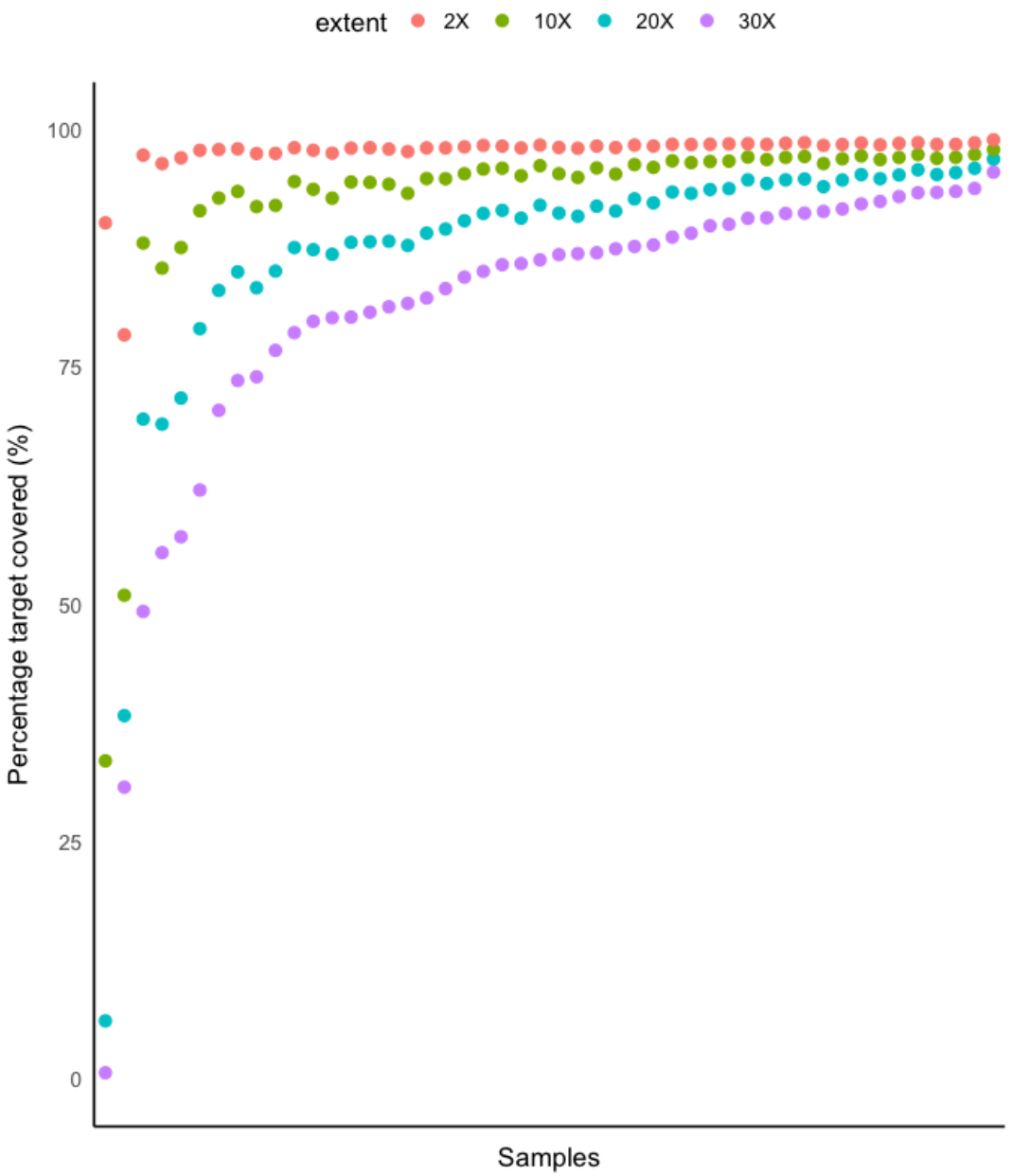
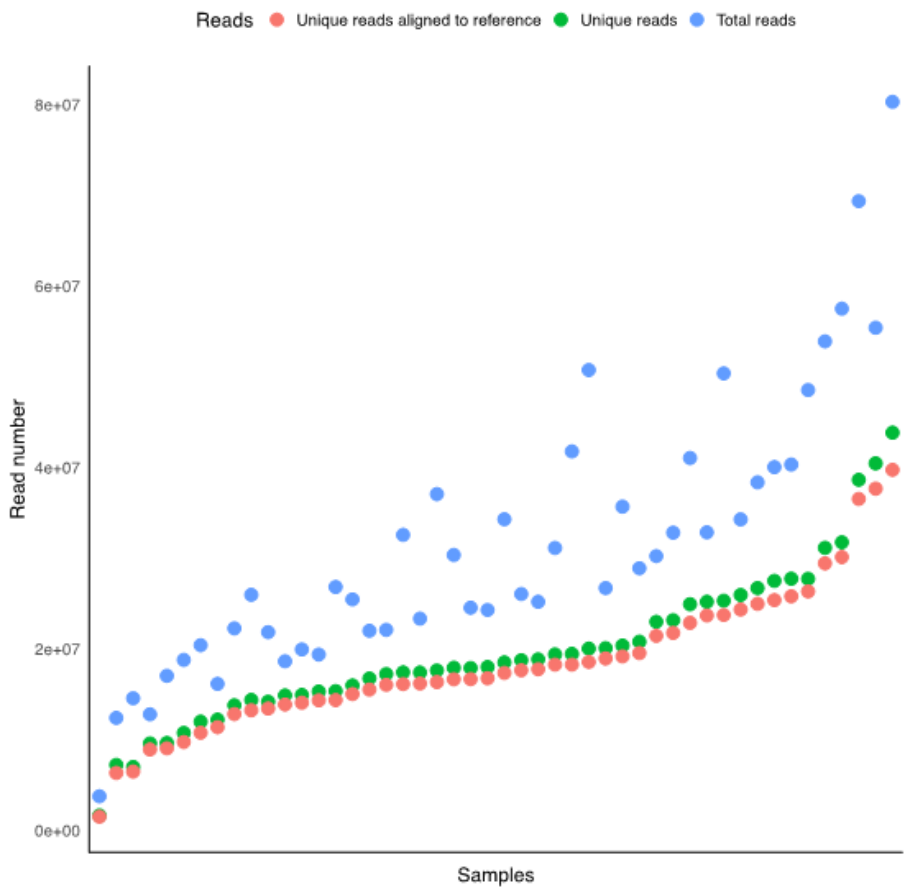


Figure 18. NGS target coverage. Percentage of the target covered by 2 reads, 10 reads, 20 reads and 30 reads.



Mean Total number of reads	31635760
Mean Non-duplicated reads	19762010
Mean Reads aligned to target	18442697
Total number of variants across samples	260700

Figure 19. Read totals per sample and mean. Total number of initial reads per sample (displayed in plot), and mean (displayed underneath plot), decreasing as duplicates are removed, and some reads are not able to be aligned to the target region of the reference. Variants are then called.

12.3.3 Variant calling and quality check

The next stage involved variant calling using GATK HaplotypeCaller.
The Variant Quality Score Recalibrator (VQSR) program then uses machine learning to learn how to determine “good” and “bad” variants in each individual dataset, and thus benefits from data volume. Output of VQSR is shown in Figure 20. It is provided with a set of robust positives, “true sites”, in this case from HapMap3 sites. It creates models for the known and novel variation in the call set to evaluate the probability each variant is real. The different 2D models it formulates (between 5-8) can be viewed by the user.

When reviewing a VQSR which displayed mapping quality on one axis it became clear that this model was flawed; high mapping quality reads were being excluded erroneously. This model was then excluded.

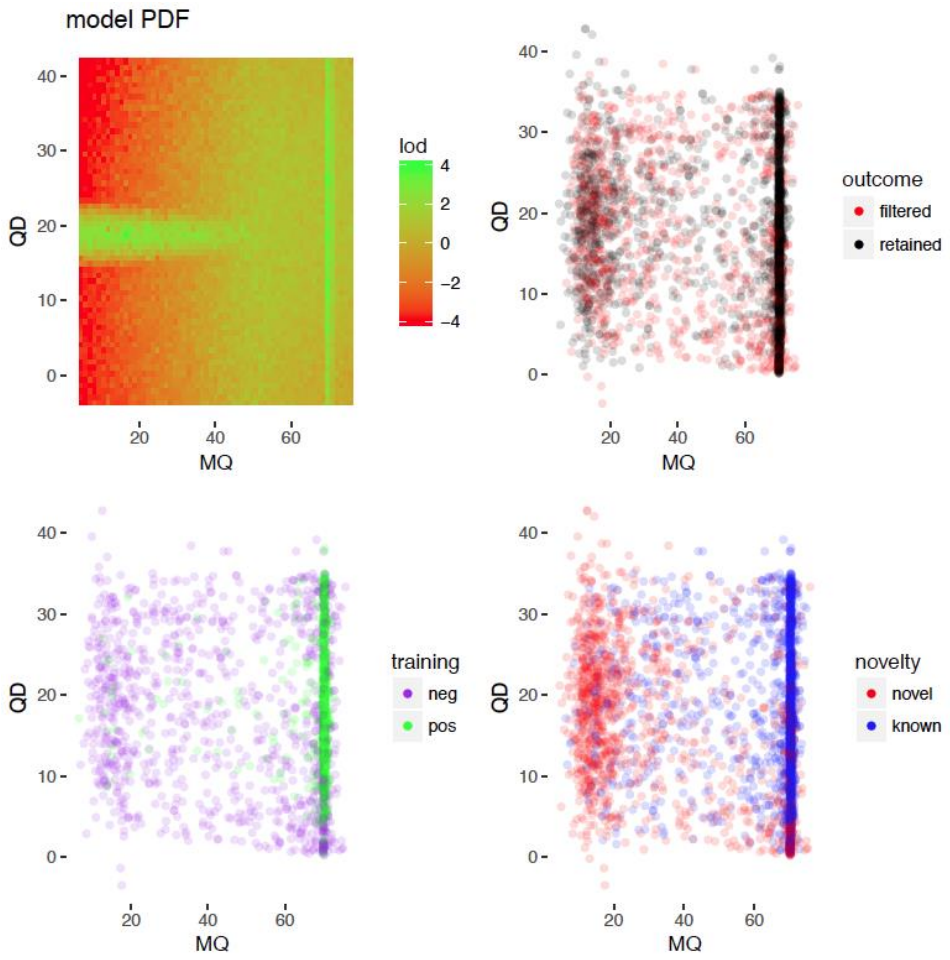


Figure 20. Model of mapping quality used by VQSR in discriminating true variants from false.

12.4 Quality control (QC)

Additional quality control stages, which were all implemented in PLINK as described in the methods, were used to exclude samples and variants that have potential to introduce bias. This procedure was performed three times: on the full Exome and Reseq cohorts of >1000 samples each separately, and on Truseq samples (which had been merged with other samples to increase data volume). Some aspects of QC

require data volume and use of 48 samples was insufficient for performing QC adequately. A number of QC measures could not be stringently applied because some of the data did not have sufficient coverage.

12.4.1 Sex check

Incongruous sex between genetic and clinical data indicates that an error has occurred in database management or laboratory processing (alternatively clinical mis-sexing of the patient could have occurred or the patient may not have a determinate sex). If the cause of the incongruity cannot be parsed, then the sample is excluded. Excluded samples due to sex QC per dataset are shown in Table 7.

There was insufficient X coverage (7 markers) in the Truseq Neurodegeneration data to rely on this parameter to genotype sex. Therefore sex was determined using y marker count. Samples with a y count ≥40 were determined as male. Samples with a y count <40 were categorised as female. All Truseq panel samples had congruent sex.

For the exome data sexes were ascertained by standard means using information from the X chromosome (see Figure 21). Differentiation of male and female samples was then determined by visual inspection of a histogram; an F value of 0.5 was specified. All G2019S exomes had congruent sex through these means.

No sex check was performed for Reseq data, due to insufficient coverage of both X and Y chromosomes.

Truseq data	Exome data	Reseq data
0	26	Not applicable

Table 7. Excluded samples due to sex check QC per cohort

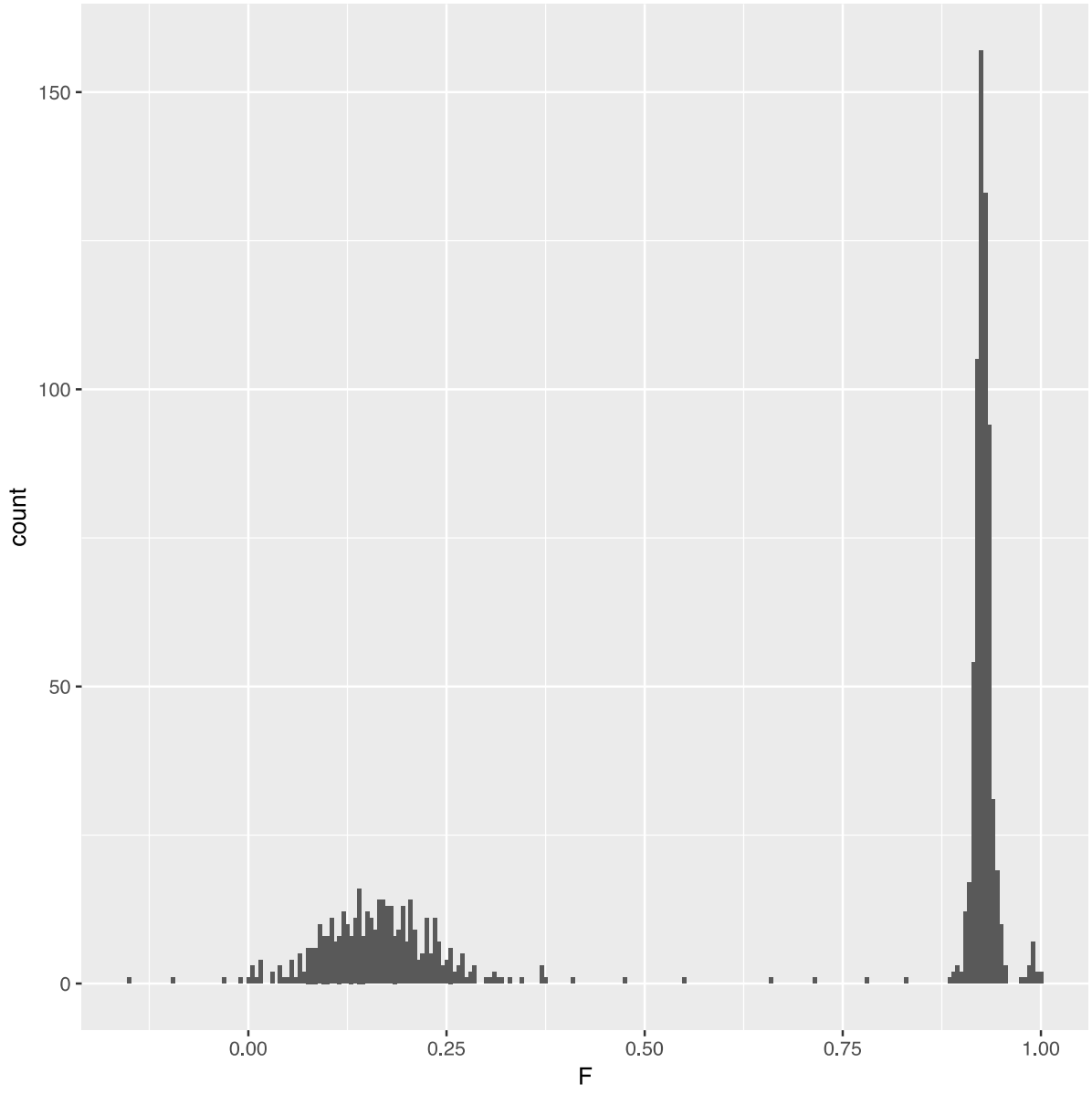


Figure 21. Sex check for exome data histogram

12.4.2 Removal of SNPs by call-rate

SNPs with a call rate less than 90% were removed, as shown in Table 8; these SNPs are likely inaccurate.

Table 8. SNPs by cohort excluded because of low call-rate.

Truseq data	Exome data	Reseq data
13035	28191	4245

12.4.3 Removal of samples by missingness

Samples were removed that have ≥20% missing SNPs, as shown in Table 9. These samples are likely poor quality. G2019S carriers were retained in each case.

Table 9. Samples per cohort excluded due to missing data

Truseq data	Exome data	Reseq data
0	15	30

12.4.4 Hardy-Weinberg Equilibrium (HWE)

Variants with a Hardy-Weinberg equilibrium exact test p-value below a threshold of 0.001 were excluded, as shown in Table 10. Hardy-Weinberg assumptions allow allele and homozygous genotype frequencies to be estimated from one generation to the next. Deviations from these estimates can indicate genotyping errors or population stratification. It is considered prudent to keep the threshold for this criteria low as

serious genotyping errors often yield extremely low p-values, whereas genuine SNP-trait associations are likely to deviate slightly from Hardy-Weinberg equilibrium.

Table 10. SNPs per cohort removed due to HWE violation

Truseq data	Exome data	Reseq data
7429	32354	333

12.4.5 Removal of low minor allele frequency (MAF) SNPs

This step was performed as the association analysis used in this study was valid only for common variation. Variants with minor allele frequency lower than 1% were excluded. Table 11 shows excluded low MAF SNPs due to this criteria.

Table 11. Low MAF SNPs removed per cohort

Truseq data	Exome data	Reseq data
100118	730974	4502

12.4.6 Identity by descent/identity by state

Identical by descent (IBD) calculations were performed. IBD refers to a matching DNA segment shared by two or more people, which was inherited from a common ancestor without any intervening recombination. Whereas Identical by state (IBS) is used to describe identical segments irrespective of descent; these segments are small and shared by many people within and between populations. Such segments do not

have genealogical relevance. Z scores were outputted by a PLINK function, which weas then assessed to identify the level of relatedness, as indicated in Table 12.

Unexpected relatedness could also be due to nonpaternity, adoption, sample mix-up, or duplicate processing of a single individual. Table 1 shows the expected Z scores for different relationships, which was used in checking relatedness of the exome data. Visualisation of IBS estimates can also uncover cryptic relatedness in the sample. If samples with cryptic relatedness are treated as independent then this will cause an increase in both false positive and negative associations.

Only the exome data was amenable to this QC measure, because only the exome data had sufficient coverage to accurately estimate IBS; IBS estimates were inflated when performed on Truseq and Reseq. These estimates were therefore excluded. G2019S carriers were unrelated in the exome cohort, and so all samples were retained.

Table 12. Probabilities that two individuals with a given relationship share 0, 1, or 2 pairs of IBD alleles

	Pr(Z0)	Pr(Z1)	PrZ2	Proportion IBD (PI_HAT)
Duplicate/identical	0	0	1	1
Parent/child	0	1	0	0.5
Full-siblings	0.25	0.5	0.25	0.5
Half-siblings	0.5	0.5	0	0.25
Grandparent/grandchild	0.5	0.5	0	0.25
Uncle/nephew	0.5	0.5	0	0.25
First cousins	0.75	0.25	0	0.125

PI_HAT is calculated as P(IBD=2) + 0.5xP(IBD=1)

12.4.7 Removal of multi-allele SNPs

Hapmap3 for CEU, CHB, JPT and YRI populations was used in the identification of multi-allelic SNPs. These were then excluded from files to leave bi-allelic sites only, as required for the association analysis used. Table 13 shows the multi-allelic sites removed.

Table 13. Multi-allelic SNPs removed per cohort

Truseq data	Exome data	Reseq data
30231	121006	15095

12.4.8 Linkage disequilibrium pruning

SNPs which were in high linkage disequilibrium were also excluded. These SNPs are surplus to requirements and would generate additional noise in the association analysis signal.

12.5 Principal components of ancestry

Principal components of ancestry (PCA) was performed in this project for two purposes: 1) To use as covariates in the association analyses (to correct for systematic ancestry differences). 2) As an additional QC measure: if samples did not have expected ethnicity then sample-swap may have occurred.

Population stratification refers to allele frequency differences between cases and controls due to systematic ancestry differences, as opposed to authentic associations; principal components of the ancestry variation were first extracted and then added as covariates in the association analyses. Figure 22 shows PCA of exome data, which was generated in PLINK using standard means. 4 main clusters with some outliers are shown. However, PLINK PCA could not be applied to Reseq and Truseq datasets because of inadequate coverage. Another technique, termed LASER, was therefore used.

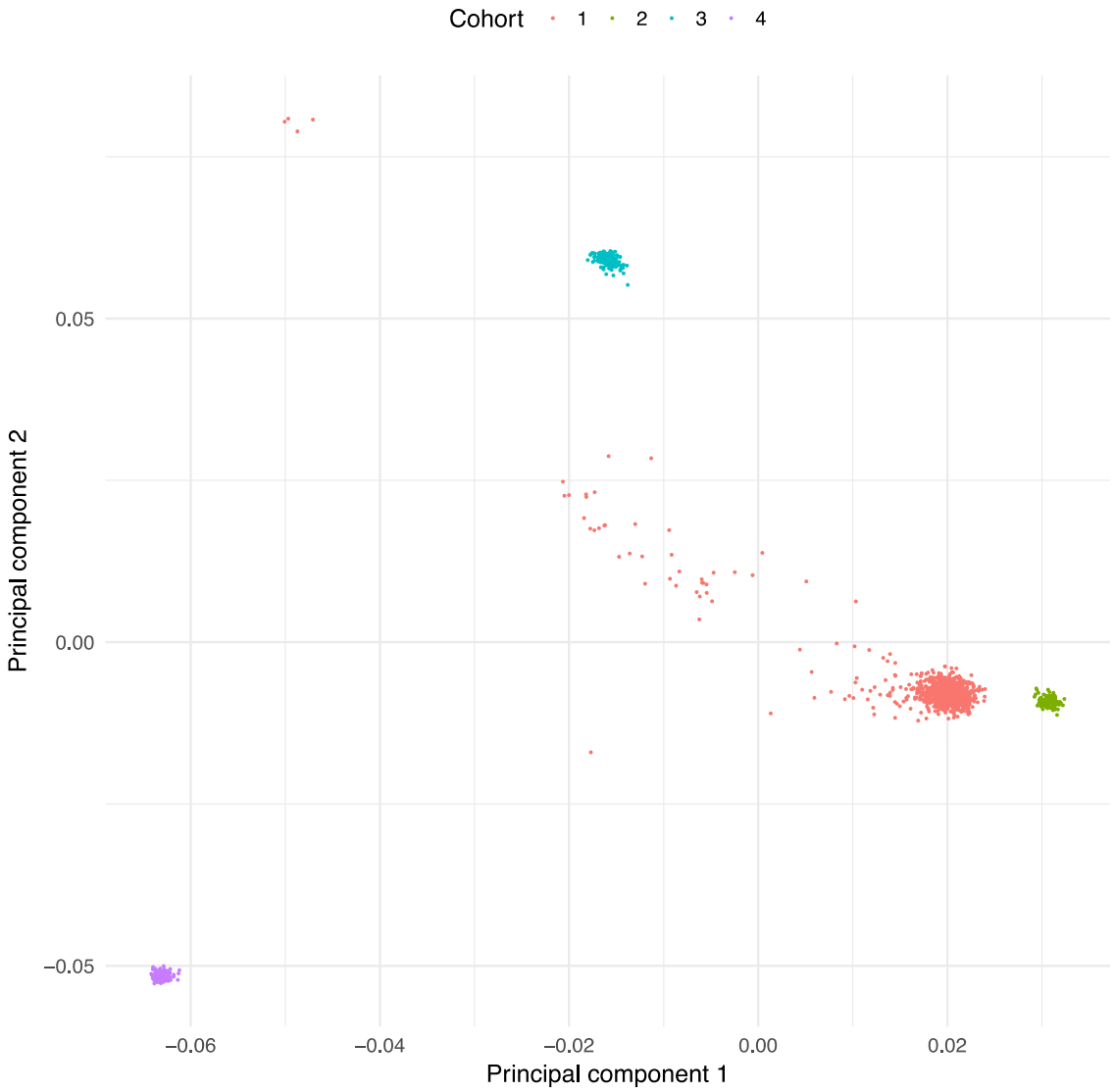


Figure 22. Principal components of ancestry analysis of exome cohort

12.5.1 LASER: principal components analysis

LASER (Locating Ancestry from SEquence Reads) software was also used to assess the principal components of ancestry (Taliun, 2017). LASER overcomes the issue of inadequate coverage by analysing off-target sequence reads that were produced as a by-product in the sequencing process. Early-stage .BAM files (including off-target information) were converted using Samtools (invoked by LASER) into Pileup files. Then pileup2seq.py is used to convert Pileup files to a SEQ file containing all study samples. LASER then uses a GENO file of references (Worldwide (imputed HGDP) reference panel was used in this project). The programme was run twice in order to assess number of useful principal components (PC):

1) With 3 principal components computed with 20 principal components for projection.
2) With 20 principal components computed with 20 components for projection.

Figure 23 shows the results of LASER PCA for 20 Principal components with 20 components for projection.

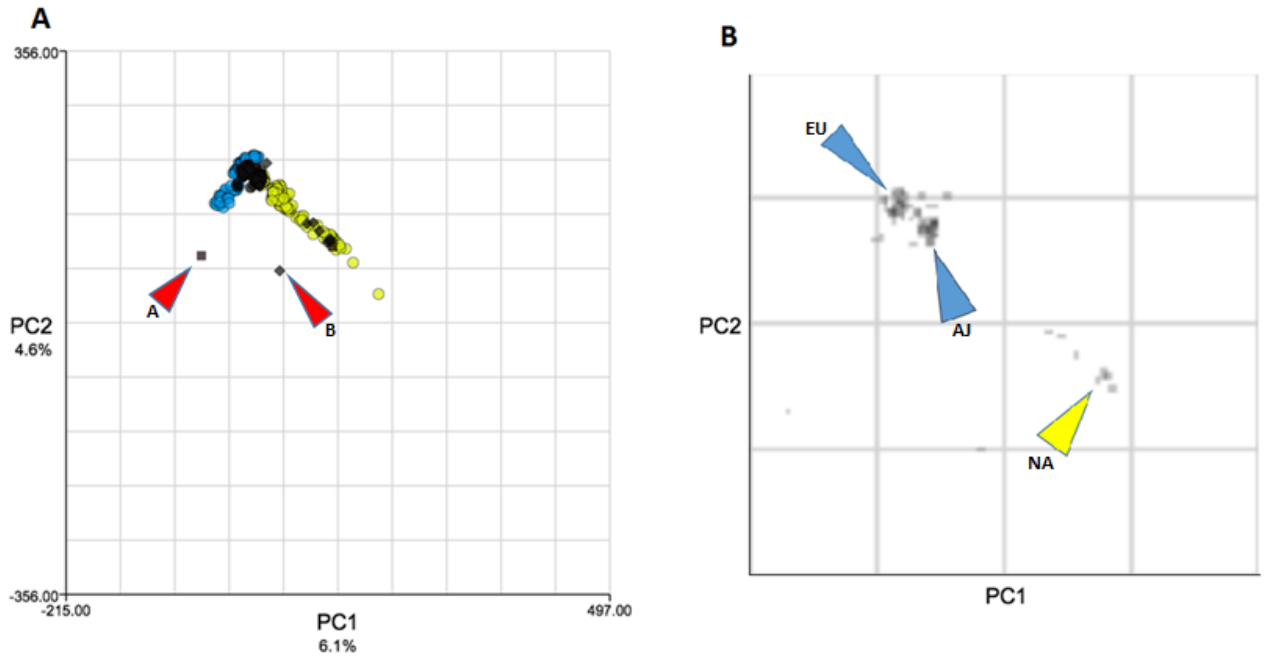


Figure 23. Principal components of ancestry plots from LASER. A) Plot overview with reference samples shown (yellow = Middle Eastern reference; Blue = European reference; Grey = Project samples.) Outliers are labelled A and B. B) Magnification of LASER plot A: North African samples lie in the Middle Eastern component of the PCA (labelled NA). European samples appear to form two clusters (European Caucasian (EU) and Ashkenazi Jewish (AJ) (lower blue arrow)), consistent with evidence from (Carmi et al., 2014)

LASER ancestry was used as an additional QC measure for Truseq samples. Quality control of samples using principal components of ethnicity showed 46/48 (95.8% (1.s.f)) of Truseq panel individuals had completely concordant ancestry between clinical report and genotyping (defined in three clusters: Middle Eastern, Ashkenazi Jewish, and European). 1 sample categorised as AJ had PC of ancestry shifted into the European cluster (PC1 = 17.2; PC2 = 193.5). This sample was retained, because in the analyses used here AJ and European were treated as a homogenous group. It was also not possible currently to determine the cause of the discrepancy out of many possibilities (increased European heritage/ non-paternity/ adoption/ database error/ laboratory sample mix-up/ contamination).

Another sample (labelled x in the above PC plot) had PC1 of 63.1 and PC2 of 66.5; the ancestry of the nearest neighbours in the reference space of this sample were from Central and Southern Asia. This sample was reported as "Caucasian" in the clinical records. This sample was excluded from the association analysis.

12.6 Data types: Truseq Neurodegeneration, Parkinson's disease exome, and Reseq cohorts

12.6.1 Data merging

The different cohorts of data were collected on different platforms and subsequently contained information that varied (with overlap), as demonstrated in Figure 24. PD exome data covered all exomes and intronic regions for 22 *LRRK2* mutation carriers. PD-reseq data for 13 individuals is targeted to Parkinson's genes and contains intronic

regions. TruSeq Neurodegeneration also covers exonic and intronic PD genes, as well as other genes implicated in neurodegeneration.

Unique VCFs were generated for each of the three datasets for only G2019S carriers. These were then converted to PLINK for ease of manipulation. An R script was used to find the intersect of the three files. It emerged that 415 variants were shared between all three datasets. Whereas 6282 variants were shared between exome and truseq datasets. There is insufficient SNP intersect between all three files to perform a targeted exome-wide association analysis. Therefore reseq data was excluded. Data from both truseq and exome cohorts were merged in VCF form.

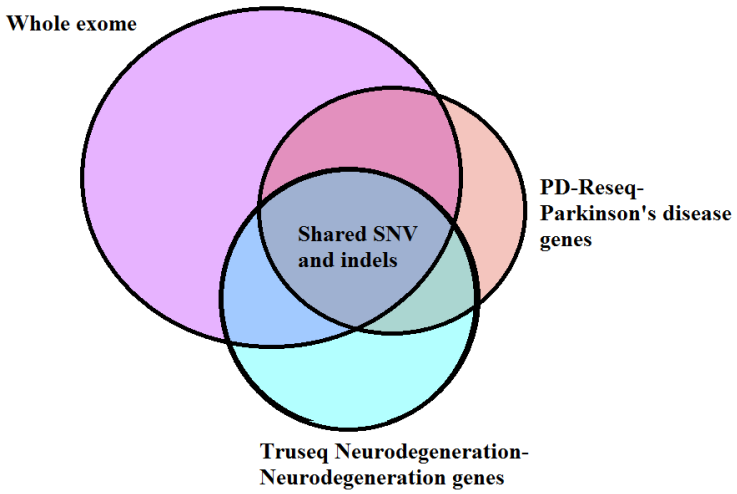


Figure 24. The divergent coverage of datasets

When merging data it was necessary to create a genomic VCF (gVCF), which allows gaps in the data to be shown as “not called”: the aim is to have every site represented in the file (for every individual), including homozygous-reference sites. The gVCF files were created and manipulated using a GATK function. The outputted files were also produced with an estimation of how closely they matched the reference. gVCF

files were then merged into a single gVCF file for all Truseq G2019S carriers. The gVCF was converted into a VCF which retained the genomic information. This process was then carried out for exome data and truseq data and then the files were merged. Data that was not shared among cohorts was culled from the merged file in PLINK, which was provided with a list of shared SNPs outputted by R.

Variants per dataset are depicted in Figure 25. This figure was generated from separate VCF files from each cohort of just G2019S, after QC steps had been completed. Many more variants exist in the exome sequencing. All three datasets had a comparatively large quantity of intronic variants.

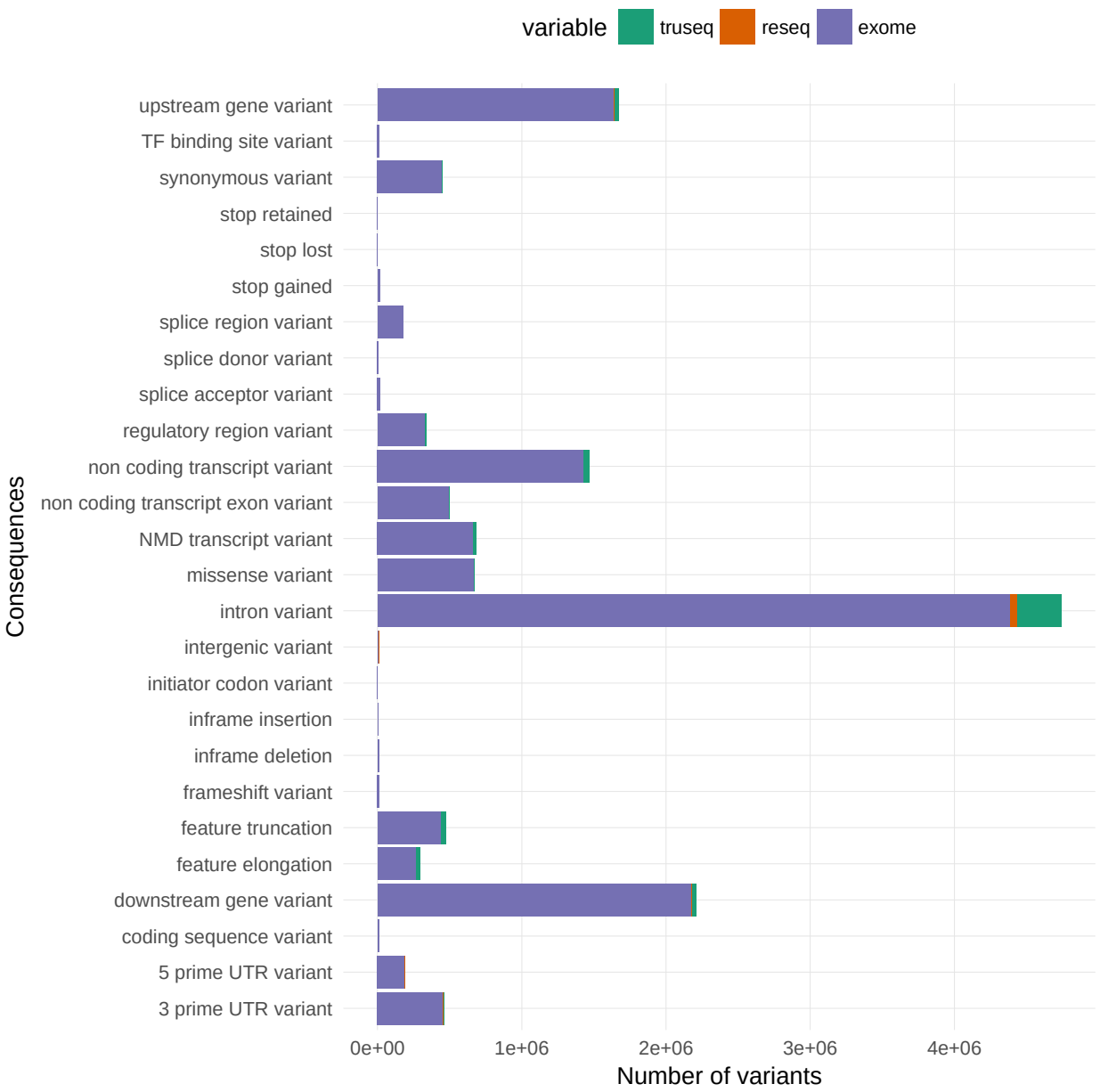


Figure 25. Types of variant per dataset

13 Haplotype analysis

After calling variants, all haplotypes were characterised as haplotype 1 (shown in Table 14).

rs2131088	40629554	A
rs2723270	40631727	T
rs10878245	40631791	C
rs7955902	40645257	A
rs10784461	40657537	G
rs10878307	40671989	A
rs41286478	40677650	A
rs4473003	40681057	T
rs7966550	40688695	T
rs11175941	40694399	G
rs11564149	40702164	C
rs11175964	40702987	G
rs11175966	40703152	C
rs41286474	40704557	T
rs33958906	40707861	C
rs35303786	40713899	T
rs11564205	40714009	A
rs7137665	40716015	T
rs17484342	40717143	G
rs2404834	40729007	C
G2019S	40734202	A
rs10784522	40740365	T
rs11176143	40742363	G
rs12426498	40761315	A
rs41286462	40761359	C
rs66737902	40761663	T

Table 14. Minimal haplotype (spanning 132kB) for all G2019S carriers used in this study (n = 75). The haplotype is consistent with haplotype 1 (European-MENA).

Figure 26 is an LD heatmap for the LRRK2 haplotype for the G2019S carriers characterised in this study. Darker squares in the LD heatmap correspond to higher R^2 values between pairs of SNPs (shown at the top of the figure). Higher R^2 values mean SNPs that are in greater linkage disequilibrium. Whereas the lighter squares indicate relative linkage equilibrium between pairs of SNPs. Four haplotype blocks are shown, which were delimited by eye.

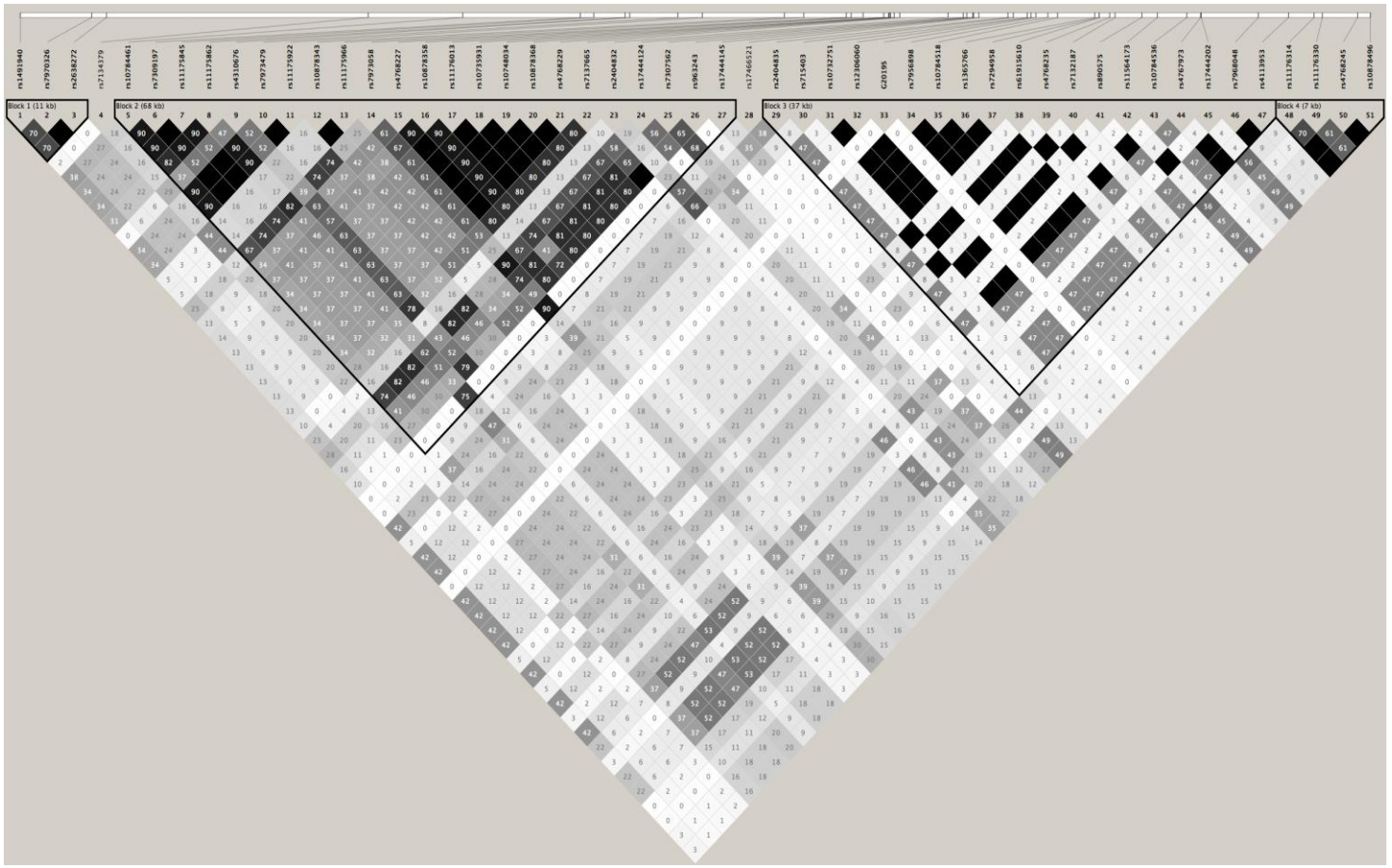


Figure 26. Linkage disequilibrium heat map of G2019S pathogenic haplotype in G2019S carriers characterised in this study

14 Association analyses

14.1 Logistic regression

Logistic regression was performed in PLINK on genome-wide data from only the exomes and truseq samples. This was because an insufficient number of SNPs were shared across all three platforms.

First degree and second degree relatives were excluded from the analysis. 1 relative from each familial set was included. Relative for inclusion from each set was determined by attempting to equalise the numbers of each participant by group in Excel (EOPD/LOPD, sex). 1 pathogenic GBA mutation carrier was excluded. 55 carriers were eligible for inclusion.

The phenotype (outcome variable) in the logistic regression was specified as either EOPD or LOPD. EOPD was defined as <57 years; LOPD was defined as ≥58 years. These categorizations were based on visual inspection of a histogram for AAO for G2019S carriers, which showed a bimodal distribution with a trough at ~57 years.

Covariates were added:

1) Clinical sex
1) 3 Principal components extracted from LASER were initially added as covariates to overcome issues of population stratification that can impact association studies. It appeared that these principal components were

inadequate in describing the data. Another logistic regression with 20 principal components and sex as covariates was subsequently performed.

Qq and Manhattan plots were visualised in R Studio using qqman package.

The manhattan plot (Figure 27) for logistic regression on 6282 genome-wide SNPs shows no significant associations; it did not reach the commonly accepted genome-wide significance criterion ($p<2.5 \times 10-8$) (Kanehisa, 2000) for gene-wide tests. The Q-Q plot for the data shows deviation from the normal. It was determined to use 20 PC of ancestry rather than 3, as this may be more adequate in describing the data. However, the Q-Q plot for this analysis still lagged below the normal line at the higher log values. Likely the size of the cohort that was amenable to association analyses ($n = 55$) was too small to detect effects, if they existed within the 6282 SNPs that were genotyped and eligible for inclusion (or further SNPs in LD with this subset).

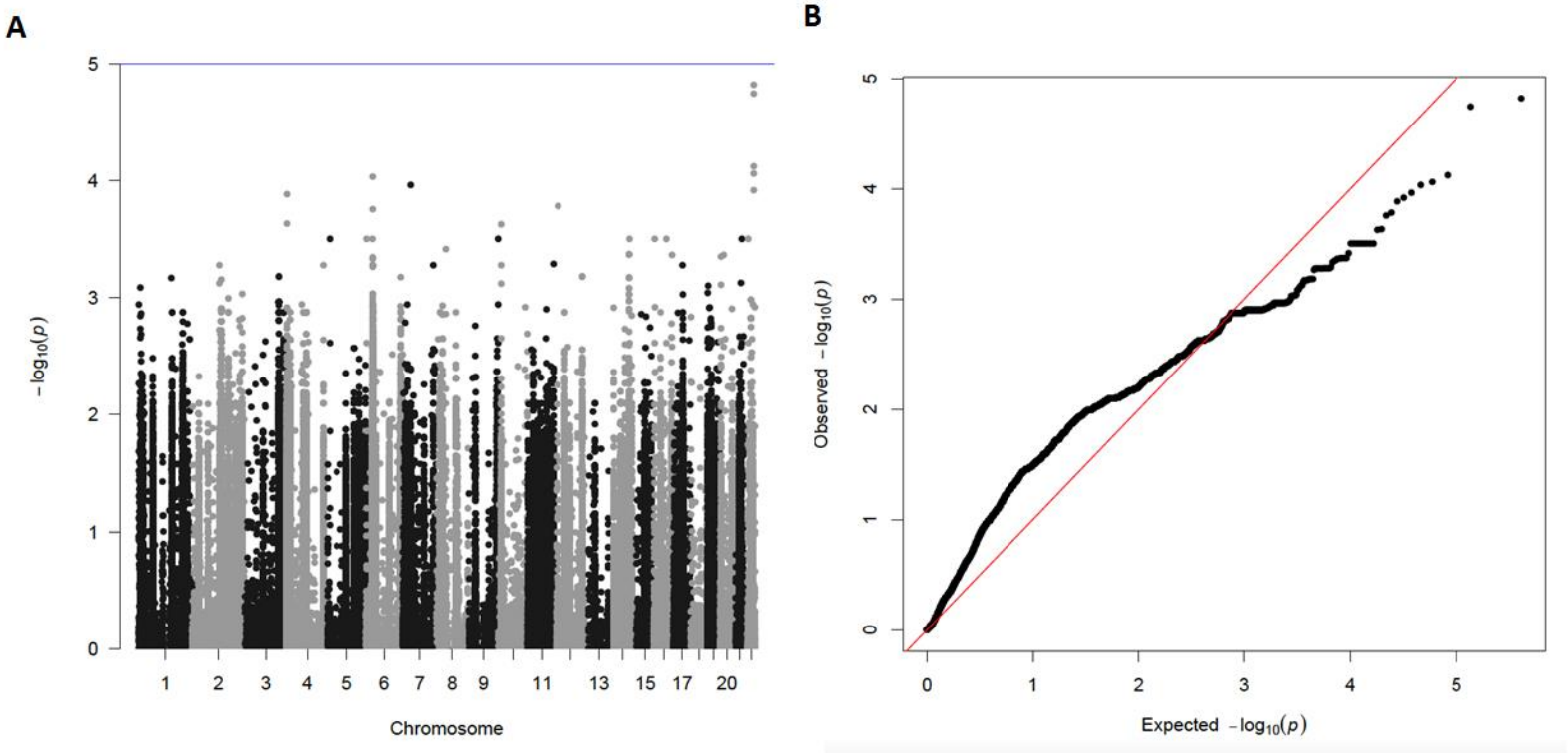


Figure 27. Logistic regression using 4 covariates (sex, 3PC of ancestry). A)

Manhattan plot. B) Associated Q-q plot

14.2 Kaplan Meier survival analysis and Cox proportional hazards model

Due to the low number of shared SNPs between exomes and Truseq samples, a Kaplan Meier survival curve and Cox proportional hazards model was applied separately in two smaller analyses only: *DNM3* and *LRRK2* Trans haplotype. Data was prepared for Kaplan Meier survival curve and Cox proportional hazards model analysis in Excel. The analyses were performed in R Studio using Survival, tidyverse and ggplot2 packages. Missing AAO data (for those that have not developed the disease) were right centred using a survival package functionality.

14.3 LRRK2

LRRK2 trans haplotype data was encoded as 0 for reference and 1 for the alternative allele for 75 G2019S carriers. 24 biallelic tag SNPs, characterised during haplotype analysis (see Table 9) were used. SNPs that did not deviate from the reference among the individuals included in analysis were removed. SNPs that were missing for certain individuals were imputed to the mean. 3 principants of ancestry and 1 prinicpal component of sex were added. 7 SNPs yielded significant p values. However, it was considered that the independence of the variables (SNPs) could have been violated by linkage disequilbrium, despite the earlier pruning. Linkage disequilbrium was assessed in LDlink for all populations, and SNPs with higher than 0.3 LD values were removed. Ideally, this would have been performed using the genetic data files used in this study, rather than with an online calculator, as patterns of linkage disequilibrium

vary among individuals. Once the data was pruned for SNPs only in relative linkage equilibrium, the p values were highly non-significant.

14.4 DNM3

38 G2019S carriers underwent Sanger sequencing for *DNM3* rs2206543 as shown in Figure 28. 2 G2019S carriers were excluded due to low remaining aliquot volumes of genomic DNA. Allele frequencies of G2019S carriers are shown in Table 6.

Archive G2019S NGS data was assessed for *DNM3* Rs2206543 SNP, both by using "grep" Linux functionality and by visual inspection of PLINK files. However the SNP was in an intronic region and was missing from exomes. It was also not in PD resequencing data. Therefore, Cox proportional hazards analysis could only be performed on these 38 carriers.

Figure 28. DNM3 rs2206543 frequencies. The ancestral allele is A.

Rs2206543 carrier status	Allele freq.	No. individuals (total n=38)
AA	0.31	12
AG	0.44	17
GG	0.24	9

An additive model was used for the SNP variants (AA encoded 0; AG encoded 1; GG encoded 2. This additive model was used because Trinh et al (2016) had indicated that risk is additive for the rs2421947 SNP, which rs2206543 is in LD with. Cox Proportional hazards model was performed in R studio, as described above. Effect size for the additive model was -0.5 (s.e. 0.6) and was not significant ($p = 0.16$).

15 Discussion

This project had two main aims: 1) To characterise the G2019S pathogenic haplotype in a new multi-ethnic cohort. 2) To find and validate genetic modifiers of AAO in G2019S Parkinson's disease. While there are established bioinformatics techniques catering to haplotype characterisation (although phase remains an issue), the second aim encompasses a field that is developing, and somewhat in its infancy: genetic modifier discovery. PD could present a particular challenge to this field, as a disease predominantly occurring in old age. The lateness of onset introduces significant scope for the impact of environmental factors on expressivity. A range of additional issues and observations related to the goal of genetic modifier discovery have arisen while implementing this project.

Overall, SNPs significantly associated with AAO in *LRRK*2 trans haplotype were not found using data from all three cohorts. The *DNM3* tag SNP rs2660543 was also not significantly associated with AAO in Truseq neurodegeneration data. There were no significant signals in the exome-wide data. Likely the cohorts were underpowered to detect associations, if they exist in these regions. The phenotypic outcome (AAO) may also not be sufficiently sensitive. All G2019S carriers shared a common pathogenic haplotype (haplotype 1).

15.1 Nature of the outcome data

Clinical data presented in this project indicated that G2019S exhibits a specific AAO distribution, although sample size was a limitation in drawing firm conclusions. The AAO distribution of the G2019S carriers appears bimodal with two similar peaks- perhaps with a negative skew on the LOPD peak (in line with clinical observations

that PD rarely begins past the ninth decade; incidence decreases after a certain age). It is possible that given a larger sample size the distribution would approach normality. AAO data from many more carriers would be required to establish this; a literature search was performed to find G2019S AAO data, however no data was found in a raw form that would allow it to be integrated.

The distribution of iPD presented in this thesis is also bimodal. Mutations in known PD genes were used as an exclusion criterion for the re-sequencing-PD gene iPD cohort. However, cases with known and unknown early-onset PD gene mutations are likely still present in the data, causing the EOPD peak. A sharp drop in incidence after age 75 is also displayed in the iPD data. With more time, we intended to remove all known EOPD gene cases from the re- sequencing data. Excluding the influence of these early-onset genetic factors, we hypothesise that onset age will be approximately normal in iPD (with some negative skew), while bimodal in G2019S disease, as described. The AAO G2019S data could indicate a genetic modifier of large effect size (because the distribution appears pronounced).

The nature of the outcome variable has implications for analyses. When performing logistic regression, data were dichotomised to either case (EOPD) or control (LOPD). As described previously, unaffected probands below the age of 58 were excluded. Whereas unaffected probands over the age of 58 were included. This meant that some individuals who reached late 80s were coded the same as those at the long tail of the LOPD distribution. This could have been a flaw in the analysis, as perhaps the data indicate that very late onset or individuals who never contract the disease have distinct genetic or environmental influences at play from those LOPD carriers who

contract disease relatively early. Decisions on where to dichotomise data were taken based on the median outputted by Kaplan Meier survival curve analysis median and on histogram data, which were constructed using the limited sample size data. Likely, a better approach would have been to perform Cox proportional hazards analysis on the targeted genome-wide data, as this has the advantage of right censoring, and does not require a specific outcome data distribution.

15.2 Cohort size

A challenge lies in acquiring a sufficient cohort of PD carriers in which to field the question of genetic modifiers. In this regard, it would be more straightforward to utilise large NGS data cohorts of iPD. However, perhaps the G2019S subgroup will be more amenable to genetic modifier identification because these patients have a more homogeneous set of genetic modifiers than iPD, and because the PD cause is better established. Although G2019S carriers may be more tractable for identifying AAO modifiers, it is more difficult to accrue sufficient sample sizes. This was likely a limitation in this study.

The Qq plot presented in the logistic regression association analysis does not follow the normal line. This can be due to biasing factors between cases and controls or can indicate insufficient quality control. However, an established protocol used previously in the laboratory for exome QC was implemented. Additionally, exclusion criteria were fairly stringent, although QC was not (because of the limitations on coverage in the targeted sequencing platforms used). Likely the cause of the lack of normality for Qq plot is the limited sample size.

There were a number of instances in which G2019S carriers were lost from analysis; some of these were unavoidable, whereas others could likely have been overcome with the use of additional techniques. 1 GBA pathogenic mutation carrier was lost, which was unpreventable as this posed a strong influence on PD development. It was not assessed whether other known Mendelian pathogenic mutations were present in the dataset, which could have been a biasing factor. GBA was specifically assessed because mutations in this gene are a common PD cause in the AJ population.

In the logistic regression analysis, only one family member per kindred was included in analysis. This resulted in unavoidable loss of data. One method that could be used to correct in part for this, is to run every possible combination of G2019S family members through separate analyses. Although this decreases power and would require corrections for multiple testing, it would also decrease the scope for bias to be introduced when group selection is performed manually, as was the case in this analysis. Group selection by equalising could also have been performed by a program designed for this purpose, however due to the small numbers of families and carriers, and because of time restrictions, this was not carried out. Relatedness can also be accounted for using linear mixed models in association analysis, as described by Eu-Ahsunthornwattana et al (2014). Otherwise kindreds could have been utilised in linkage analysis or using tests such as TRAFIC (Test for Rare-Variant Association using Family-based Internal Controls), which tests for rare variants in affected sibpairs (Risch, 2000)

Exclusion of samples based on ethnicity was also performed. 8 North African G2019S carriers from 3 kindreds were not utilised in the logistic regression analysis, as these

formed a distinct cluster in the PCA space. Case control matching of ethnicity may have been possible. Another approach is to perform independent association analyses in different ethnic groups and subsequently perform meta-analysis. For this project, this would not have been possible for North African samples as only 3 samples were eligible (because of family-based exclusion criteria). Separate analyses could have been performed on AJ and European but this is likely unnecessary as these individuals cluster closely and ancestry was corrected for. Also, this would inevitably decrease the cohort size for the individual analyses, likely to an unacceptably low level.

Although both Cox Proportional Hazards model and logistic regression allowed G2019S carriers >58 years to be included, 1 young G2019S carrier had to be excluded, because it was unknown whether this individual would develop PD early, late, or never. However, the data can be retained, and if the patient consents to re-contact, then the data later may be of use.

15.3 Platform limitations

Data used in this project were collected by piecemeal means, spanning a number of platforms. This is often necessitated in NGS for clinical research by the biological hypothesis under investigation; financial restrictions; grant specifications; limitations on available patients; availability of resources; and the prevailing mood of the field (i.e. previous underestimation of the disadvantages of restricted candidate gene data). In this project data used had been collected on three different platforms, with varying coverage. Exome data was most useful for this project as it includes genes which have not already been implicated in PD or neurodegeneration. There are advantages to

targeted approaches also, however. Targeted sequencing is generally implemented as many more samples can be sequenced over a smaller genomic region to a very good level of depth, for the same or less cost, as only a few high depth exomes or genomes. This is valuable as large sample size is such an important factor in identifying genetic variations associated with phenotype. However, care should be taken in selecting the sequencing platform for the current purpose, and to enable future re-purposing.

Different analytic techniques, quantities of data, and genetic platforms (with different coverage) are required in finding different types of variation (intronic, splice-site, SNP, CNV and rare variants). SNPs have been historically most amenable to detection, and were investigated in this study. All three platforms had a comparatively high level of intronic variation, which was not useful for our purposes. Although in future this data could be utilised. Merging datasets also pruned potentially modifier-relevant data. When merging multiple datasets, the new file becomes as limited as the least coverage dataset. This was a limitation experienced in this study with the re-sequencing dataset, which shared too small a set of shared SNPs with the other datasets. Imputation may have been of use in overcoming this (but only to an extent). It was unfortunate that a larger number (n = 22) of G2019S samples were available in the re-sequencing cohort, compared to exome data (n = 13). There is sometimes the option of sequencing samples again on another platform to increase coverage at a later stage, providing sufficient quantities of genomic DNA remain or there is a re-contact clause in the patient informed consent forms. All truseq samples were also recently sequenced on the NeuroX array, the results of which will be available in the near future. The re-sequencing cohort had also been sequenced previously on the NeuroX (Nalls et al., 2015). With collaboration (which is being recognised as integral to the

identification of rare variants and weak effect common variants in particular, and genetic variation impacting disease more broadly) this data could be used. Though this was not attempted during this project.

The intersect SNP data between Truseq cohort and exome cohorts was also likely too small (n = 6282). This could have been accommodated for to an extent by imputation of genotypes. However, the dataset would still be limited. It would also be useful to systematically review which chromosomal locations and genes the shared SNPs were contained within. Although the manhattan plot seems to indicate good coverage of all chromosomes. This presents a limitation in the interpretation of these results. The manhattan plot is not significant, but the exact coverage is not described here. Perhaps genetic modifier SNPs or those in LD with modifiers are not covered. However, knowing this information would be of somewhat limited use because the sample size is too limited to provide any evidence that these SNPs did not affect AAO.

15.4 DNM3

Because DNM3 was not included on the Truseq platform, a tag SNP was independently sequenced. Supplementing NGS with Sanger sequencing introduces additional time and cost into studies. Although Sanger is regularly utilised to confirm findings from NGS, due to its increased accuracy. The DNM3 SNP was not present in the archive data, meaning only the Truseq cohort could be utilised for this research question. This was very limiting in terms of the sample size in which DNM3 modification could be addressed. More G2019S samples need to be DNM3 sequenced in order to determine pathogenicity.

Another limitation in DNM3 analysis may have been related to the use of an additive model. Although the research by Trinh et al (2016) indicated that risk for rs2421947 (the SNP in LD with rs2206543) increases in an additive fashion, with one homozygous site causing least risk and the other homozygous site causing most risk, and the heterozygous allele causing intermediate risk, it is not known whether there is equal spacing between these risks (as is required by an additive model). Further work could be done to explore model misspecifications and make use of other models that do not impose this structure.

15.5 Quality control

Ideally sequencing data should be merged at a very early stage, and should be stringently processed concurrently. Although data in this project were processed separately, it was performed either by or under the supervision of a single bioinformatician (Dr Alan Pittman) implementing a slightly modified version of GATK best practices. Because stages were performed using the same script this minimises the scope for processing factors to affect the outcome. However, some biases could be introduced when sample sizes differ between runs of software (i.e. VQSR is impacted by the sample size in its estimations of variant quality). It may have been beneficial, by some unknown degrees, to merge the data in the first instance and perform all QC concurrently.

Although ethnic heterogeneity presented an issue for association analyses, having an ethnically heterogeneous cohort allowed for an additional quality control measure, as described by Mathias et al (2016). Depending on the heterogeneity in ethnicity in

cohorts, this can be a more informative approach than checking sex for quality control. However, there is more scope in this approach for the self-reported ethnicity not to match the PCA results, as compared to genetic sex. Exclusion based on this approach should likely be used sparingly.

15.6 Clinical data

Extremely detailed, relevant and readily available clinical data will likely be important in identifying genetic modifiers of disease. Moss et al (2017) were able to utilise TRACK-HD, a cohort with high quality phenotypic data to earmark *MSH3* as a likely disease modifying mutation in Huntington's disease (HD). Moss et al (2017) discuss that AAO is an imperfect quantitative phenotype, due to subjectivity and occasional lack of availability. It is noted that they were likely able to find MSH3 disease modifier because the effect size at this locus is large and the TRACK-HD phenotype characterisation is sensitive. Of course, comprehensive cohort characterisation, of the variety used in TRACK-HD is expensive, minimising the number of participants it could be performed upon when compared to collecting AAO data. HD is considered a disease of lesser environmental influence than PD, which could also mean that more patients would need sensitive phenotype characterisation, if a similar approach were to be used. Although expensive, this approach may substantially increase the capability to detect modifiers. G2019S carriers could also be a useful cohort in which to perform extensive cohort categorisation due to their homogeneity in disease cause and similarity to idiopathic LOPD.

During this study age-at-onset was the only readily available phenotype outcome measure. The option of using the Montreal Cognitive Assessment (MOCA) was also discussed, in addition to the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS). Although, upon review, these documents were not ubiquitously and comprehensively available in the patient clinical trial records. Patients are being re-contacted where possible to acquire this information. It is also likely that a custom-made cohort characterisation, similar to TRACK-HD, including imaging and biomarkers would be most useful.

Notably, the non-G2019S carriers that were sequenced in this project have not been characterised fully. This was due to limits on available data; these data were predominantly contained in paper records, sometimes at other locations, and could not be accessed in the timeframe of this project. This meant that the data could not be used in providing estimates of G2019S prevalence in different ethnic populations.

15.7 G2019S haplotypes

All haplotypes found in this project were consistent with haplotype 1. This is an expected outcome, based on the literature, as haplotype 1 is found in North Africans, Ashkenazi Jews and Europeans, as well as many other populations. It was considered, based on the evidence to date, that haplotype 2 may be found in a very limited number of the European individuals during the course of the project. There is limited research on the disease course of these alternative haplotype individuals. Though they may be of somewhat limited utility in understanding *LRRK2* related disease, compared to understanding whether and which variability in the trans haplotype impacts disease.

Although variability on the trans haplotype in LRRK2 haplotype 1 disease may be linked to any dysfunction or functional benefits that result from having haplotypes 2 or 3.

Another area that could have been investigated further during this project is the comparative lengths of the haplotypes. Bouhache et al (2017) found the shortest haplotype in North African Berbers and hypothesise that this is where the G2019S haplotype 1 mutation originated. Although they had access to estimates of whether carriers were Berber or Arab in ancestry, which was not the case in this project.

There is some uncertainty in determining haplotypes because of phase. It has to be deduced (either by a program or using a systematic Excel based method, as described earlier) which haplotype was present. The Excel based method is an established method in our laboratory with good differentiation.

15.8 Future directions

Overall, accruing larger sample sizes will be essential in addressing the question of genetic modifiers of G2019S disease. The current data could also be assessed in other ways: for instance, using additional techniques to analyse affected sib pairs, or in assessing intronic variation. Additionally, analysis by Lubbe et al (2016) could be replicated, assessing the influence of variants of unknown significance in ATP13A2 and other genes. Genetic WGS, WES and targeted WES data has the potential to be used and re-used. This project has introduced me to factors that must be carefully considered when designing genetic investigations.

References

Al-Mubarak, B. R., Bohlega, S. A., Alkhairallah, T. S., Magrashi, A. I., Alturki, M. I., Khalil, D. S., Alabdulaziz, B. S., Al-Shaar, H. A., Mustafa, A. E. & Alyemni, E. A. 2015. Parkinson's disease in Saudi patients: a genetic study. *PloS one,* 10**,** e0135950.

Ali, K. & Morris, H. R. 2015. Parkinson's disease: chameleons and mimics. *Pract Neurol,* 15**,** 14-25.

Ardlie, K. G., Kruglyak, L. & Seielstad, M. 2002. Patterns of linkage disequilibrium in the human genome. *Nature reviews. Genetics,* 3**,** 299.

Bar-Shira, A., Hutter, C. M., Giladi, N., Zabetian, C. P. & Orr-Urtreger, A. 2009. Ashkenazi Parkinson's disease patients with the LRRK2 G2019S mutation share a common founder dating from the second to fifth centuries. *neurogenetics,* 10**,** 355-358.

Bardien, S., Lesage, S., Brice, A. & Carr, J. 2011a. Genetic characteristics of leucine-rich repeat kinase 2 (LRRK2) associated Parkinson's disease. *Parkinsonism Relat Disord,* 17**,** 501-8.

Berwick, D. C. & Harvey, K. 2012. LRRK2 functions as a Wnt signaling scaffold, bridging cytosolic proteins and membrane-localized LRP6. *Human Molecular Genetics,* 21**,** 4966-4979.

Bonifati, V. 2012. Parkinsonism. *In:* WOOD, N. (ed.) *Neurogenetics: A Guide for Clinicians.* Cambridge: Cambridge University Press.

Bonifati, V., Rizzu, P., Van Baren, M. J., Schaap, O., Breedveld, G. J., Krieger, E., Dekker, M. C., Squitieri, F., Ibanez, P., Joosse, M., Van Dongen, J. W., Vanacore, N., Van Swieten, J. C., Brice, A., Meco, G., Van Duijn, C. M., Oostra, B. A. & Heutink, P. 2003. Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism. *Science,* 299**,** 256-9.

Botta-Orfila, T., Ezquerra, M., Pastor, P., Fernández-Santiago, R., Pont-Sunyer, C., Compta, Y., Lorenzo-Betancor, O., Samaranch, L., Martí, M. J. & Valldeoriola, F. 2012. Age at onset in LRRK2-associated PD is modified by SNCA variants. *Journal of Molecular Neuroscience,* 48**,** 245-247.

Bouhouche, A., Tibar, H., Ben El Haj, R., El Bayad, K., Razine, R., Tazrout, S., Skalli, A., Bouslam, N., Elouardi, L. & Benomar, A. 2017. LRRK2 G2019S Mutation: Prevalence and Clinical Features in Moroccans with Parkinson's Disease. *Parkinson's Disease,* 2017.

Bras, J. M., Guerreiro, R. J., Ribeiro, M. H., Januario, C., Morgadinho, A., Oliveira, C. R., Cunha, L., Hardy, J. & Singleton, A. 2005. G2019S dardarin substitution is a common cause of Parkinson's disease in a Portuguese cohort. *Movement Disorders,* 20**,** 1653-1655.

Carmi, S., Hui, K. Y., Kochav, E., Liu, X., Xue, J., Grady, F., Guha, S., Upadhyay, K., Ben-Avraham, D., Mukherjee, S., Bowen, B. M., Thomas, T., Vijai, J., Cruts, M., Froyen, G., Lambrechts, D., Plaisance, S., Van Broeckhoven, C., Van Damme, P., Van Marck, H., Barzilai, N., Darvasi, A., Offit, K., Bressman, S., Ozelius, L. J., Peter, I., Cho, J. H., Ostrer, H., Atzmon, G., Clark, L. N., Lencz, T. & Pe'er, I. 2014. Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. *Nat Commun,* 5**,** 4835.

Change, N., Mercier, G. & Lucotte, G. 2008. Genetic screening of the G2019S mutation of the LRRK2 gene in Southwest European, North African, and Sephardic Jewish subjects. *Genetic testing,* 12**,** 333-339.

Cilia, R., Siri, C., Rusconi, D., Allegra, R., Ghiglietti, A., Sacilotto, G., Zini, M., Zecchinelli, A. L., Asselta, R. & Duga, S. 2014. LRRK2 mutations in Parkinson's disease: confirmation of a gender effect in the Italian population. *Parkinsonism & related disorders,* 20**,** 911-914.

Cogo, S., Greggio, E. & Lewis, P. A. 2017. Leucine Rich Repeat Kinase 2: beyond Parkinson's and beyond kinase inhibitors. Taylor & Francis.

Cook, D. A., Kannarkat, G. T., Cintron, A. F., Butkovich, L. M., Fraser, K. B., Chang, J., Grigoryan, N., Factor, S. A., West, A. B., Boss, J. M. & Tansey, M. G. 2017. LRRK2 levels in immune cells are increased in Parkinson's disease. *NPJ Parkinsons Dis,* 3**,** 11.

Cossu, G., Van Doeselaar, M., Deriu, M., Melis, M., Molari, A., Di Fonzo, A., Oostra, B. A. & Bonifati, V. 2007. LRRK2 mutations and Parkinson's disease in Sardinia—A Mediterranean genetic isolate. *Parkinsonism & related disorders,* 13**,** 17-21.

Covy, J. P. & Giasson, B. I. 2009. Identification of compounds that inhibit the kinase activity of leucine-rich repeat kinase 2. *Biochemical and biophysical research communications,* 378**,** 473-477.

Deng, J., Lewis, P. A., Greggio, E., Sluch, E., Beilina, A. & Cookson, M. R. 2008. Structure of the ROC domain from the Parkinson's disease-associated leucine-rich repeat kinase 2 reveals a dimeric GTPase. *Proc Natl Acad Sci U S A,* 105**,** 1499-504.

Devlin, B. & Risch, N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics,* 29**,** 311-322.

Di Fonzo, A., Rohé, C. F., Ferreira, J., Chien, H. F., Vacca, L., Stocchi, F., Guedes, L., Fabrizio, E., Manfredi, M. & Vanacore, N. 2005. A frequent LRRK2 gene mutation associated with autosomal dominant Parkinson's disease. *The Lancet,* 365**,** 412-415.

Escott-Price, V., Nalls, M. A., Morris, H. R., Lubbe, S., Brice, A., Gasser, T., Heutink, P., Wood, N. W., Hardy, J., Singleton, A. B., Williams, N. M., Conso, I. P. S. D. G. & Consortium, I. 2015. Polygenic Risk of Parkinson Disease Is Correlated with Disease Age at Onset. *Annals of Neurology,* 77**,** 582-591.

Eu-Ahsunthornwattana, J., Miller, E. N., Fakiola, M., Jeronimo, S. M., Blackwell, J. M., Cordell, H. J. & 2, W. T. C. C. C. 2014. Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS genetics,* 10**,** e1004445.

Fadista, J., Manning, A. K., Florez, J. C. & Groop, L. 2016. The (in) famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics,* 24**,** 1202-1205.

Farrer, M. J., Gibson, R. & Hentati, F. 2008. The ancestry of LRRK2 Gly2019Ser parkinsonism–Authors' reply. *The Lancet Neurology,* 7**,** 770-771.

Farrer, M. J., Stone, J. T., Lin, C.-H., Dächsel, J. C., Hulihan, M. M., Haugarvoll, K., Ross, O. A. & Wu, R.-M. 2007. Lrrk2 G2385R is an ancestral risk factor for Parkinson's disease in Asia. *Parkinsonism & related disorders,* 13**,** 89-92.

Fearnley, J. M. & Lees, A. J. 1991b. Ageing and Parkinson's disease: substantia nigra regional selectivity. *Brain,* 114**,** 2283-2301.

Ferreira, J. J., Guedes, L. C., Rosa, M. M., Coelho, M., Van Doeselaar, M., Schweiger, D., Di Fonzo, A., Oostra, B. A., Sampaio, C. & Bonifati, V. 2007. High prevalence of LRRK2 mutations in familial and sporadic Parkinson's disease in Portugal. *Movement Disorders,* 22**,** 1194-1201.

Floris, G., Cannas, A., Solla, P., Murru, M. R., Tranquilli, S., Corongiu, D., Rolesu, M., Cuccu, S., Sardu, C. & Marrosu, F. 2009. Genetic analysis for five LRRK2 mutations in a Sardinian parkinsonian population: Importance of G2019S and R1441C mutations in sporadic Parkinson's disease patients. *Parkinsonism & related disorders,* 15**,** 277-280.

Franke, A., Mcgovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J. & Roberts, R. 2010. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature genetics,* 42**,** 1118-1125.

Funayama, M., Hasegawa, K., Kowa, H., Saito, M., Tsuji, S. & Obata, F. 2002. A new locus for Parkinson's disease (PARK8) maps to chromosome 12p11. 2–q13. 1. *Annals of neurology,* 51**,** 296-301.

Funayama, M., Hasegawa, K., Ohta, E., Kawashima, N., Komiyama, M., Kowa, H., Tsuji, S. & Obata, F. 2005. An LRRK2 mutation as a cause for the parkinsonism in the original PARK8 family. *Annals of neurology,* 57**,** 918-921.

Gaig, C., Ezquerra, M., Marti, M. J., Muñoz, E., Valldeoriola, F. & Tolosa, E. 2006. LRRK2 mutations in Spanish patients with Parkinson disease: frequency, clinical features, and incomplete penetrance. *Archives of neurology,* 63**,** 377-382.

Gan-Or, Z., Bar-Shira, A., Mirelman, A., Gurevich, T., Giladi, N. & Orr-Urtreger, A. 2012. The age at motor symptoms onset in LRRK2-associated Parkinson's disease is affected by a variation in the MAPT locus: a possible interaction. *Journal of Molecular Neuroscience,* 46**,** 541-544.

Gandhi, S. & Wood, N. W. 2012. The human genome project – what it really means and where next. *In:* WOOD, N. (ed.) *Neurogenetics: A Guide for Clinicians.* Cambridge: Cambridge University Press.

Gardai, S. J., Mao, W., Schule, B., Babcock, M., Schoebel, S., Lorenzana, C., Alexander, J., Kim, S., Glick, H., Hilton, K., Fitzgerald, J. K., Buttini, M., Chiou, S. S., Mcconlogue, L., Anderson, J. P., Schenk, D. B., Bard, F., Langston, J. W., Yednock, T. & Johnston, J. A. 2013. Elevated alpha-synuclein impairs innate immune cell function and provides a potential peripheral biomarker for Parkinson's disease. *PLoS One,* 8**,** e71634.

Giesert, F., Hofmann, A., Bürger, A., Zerle, J., Kloos, K., Hafen, U., Ernst, L., Zhang, J., Vogt-Weisenhorn, D. M. & Wurst, W. 2013. Expression analysis of Lrrk1, Lrrk2 and Lrrk2 splice variants in mice. *PLoS One,* 8**,** e63778.

Gilks, W. P., Abou-Sleiman, P. M., Gandhi, S., Jain, S., Singleton, A., Lees, A. J., Shaw, K., Bhatia, K. P., Bonifati, V. & Quinn, N. P. 2005. A common LRRK2 mutation in idiopathic Parkinson's disease. *The Lancet,* 365**,** 415-416.

Godena, V. K., Brookes-Hocking, N., Moller, A., Shaw, G., Oswald, M., Sancho, R. M., Miller, C. C., Whitworth, A. J. & De Vos, K. J. 2014. Increasing microtubule acetylation rescues axonal transport and locomotor deficits caused by LRRK2 Roc-COR domain mutations. *Nature communications,* 5.

Goldwurm, S., Di Fonzo, A., Simons, E., Rohe, C., Zini, M., Canesi, M., Tesei, S., Zecchinelli, A., Antonini, A. & Mariani, C. 2005. The G6055A (G2019S) mutation in LRRK2 is frequent in both early and late onset Parkinson's disease and originates from a common ancestor. *Journal of Medical Genetics,* 42**,** e65-e65.

Golub, Y., Berg, D., Calne, D. B., Pfeiffer, R. F., Uitti, R. J., Stoessl, A. J., Wszolek, Z. K., Farrer, M. J., Mueller, J. C. & Gasser, T. 2009. Genetic factors

influencing age at onset in LRRK2-linked Parkinson disease. *Parkinsonism & related disorders,* 15**,** 539-541.

González-Fernández, M. C., Lezcano, E., Ross, O. A., Gómez-Esteban, J. C., Gómez-Busto, F., Velasco, F., Alvarez-Alvarez, M., Rodríguez-Martínez, M. B., Ciordia, R. & Zarranz, J. J. 2007. Lrrk2-associated parkinsonism is a major cause of disease in Northern Spain. *Parkinsonism & related disorders,* 13**,** 509-515.

Gorostidi, A., Ruiz-Martinez, J., De Munain, A. L., Alzualde, A. & Massó, J. M. 2009. LRRK2 G2019S and R1441G mutations associated with Parkinson's disease are common in the Basque Country, but relative prevalence is determined by ethnicity. *Neurogenetics,* 10**,** 157.

Hala, K., Vilhelmova, M., Hartmanova, I. & Pink, W. 1983. Chronic parkinsonism in humans due to product of meperidine-analog synthesis. *Science,* 219**,** 979-980.

Hashad, D. I., Abou-Zeid, A. A., Achmawy, G. A., Allah, H. M. S. & Saad, M. A. 2011. G2019S mutation of the leucine-rich repeat kinase 2 gene in a cohort of Egyptian patients with Parkinson's disease. *Genetic testing and molecular biomarkers,* 15**,** 861-866.

Hassin-Baer, S., Laitman, Y., Azizi, E., Molchadski, I., Galore-Haskel, G., Barak, F., Cohen, O. S. & Friedman, E. 2009. The leucine rich repeat kinase 2 (LRRK2) G2019S substitution mutation. *Journal of neurology,* 256**,** 483-487.

Healy, D. G., Falchi, M., O'sullivan, S. S., Bonifati, V., Durr, A., Bressman, S., Brice, A., Aasly, J., Zabetian, C. P., Goldwurm, S., Ferreira, J. J., Tolosa, E., Kay, D. M., Klein, C., Williams, D. R., Marras, C., Lang, A. E., Wszolek, Z. K., Berciano, J., Schapira, A. H., Lynch, T., Bhatia, K. P., Gasser, T., Lees, A. J., Wood, N. W. & International, L. C. 2008. Phenotype, genotype, and worldwide genetic penetrance of LRRK2-associated Parkinson's disease: a case-control study. *Lancet Neurol,* 7**,** 583-90.

Hedrich, K., Winkler, S., Hagenah, J., Kabakci, K., Kasten, M., Schwinger, E., Volkmann, J., Pramstaller, P. P., Kostic, V. & Vieregge, P. 2006. Recurrent LRRK2 (Park8) mutations in early-onset Parkinson's disease. *Movement disorders,* 21**,** 1506-1510.

Hentati, F., Trinh, J., Thompson, C., Nosova, E., Farrer, M. J. & Aasly, J. O. 2014. LRRK2 parkinsonism in Tunisia and Norway: a comparative analysis of disease penetrance. *Neurology,* 83**,** 568-569.

Hernan, M. A., Takkouche, B., Caamano-Isorna, F. & Gestal-Otero, J. J. 2002. A meta-analysis of coffee drinking, cigarette smoking, and the risk of Parkinson's disease. *Ann Neurol,* 52**,** 276-84.

Ho, C. C., Rideout, H. J., Ribe, E., Troy, C. M. & Dauer, W. T. 2009. The Parkinson disease protein leucine-rich repeat kinase 2 transduces death signals via Fas-associated protein with death domain and caspase-8 in a cellular model of neurodegeneration. *J Neurosci,* 29**,** 1011-6.

Ho, D. H., Jang, J., Joe, E.-H., Son, I., Seo, H. & Seol, W. 2016. G2385r and I2020T mutations increase LRRK2 GTPase activity. *BioMed research international,* 2016.

Huang, Y., Halliday, G. M., Vandebona, H., Mellick, G. D., Mastaglia, F., Stevens, J., Kwok, J., Garlepp, M., Silburn, P. A. & Horne, M. K. 2007. Prevalence and clinical features of common LRRK2 mutations in Australians with Parkinson's disease. *Movement disorders,* 22**,** 982-989.

Illarioshkin, S., Shadrina, M., Slominsky, P., Bespalova, E., Zagorovskaya, T., Bagyeva, G. K., Markova, E., Limborska, S. & Ivanova-Smolenskaya, I. 2007.

A common leucine-rich repeat kinase 2 gene mutation in familial and sporadic Parkinson's disease in Russia. *European journal of neurology,* 14**,** 413-417.

Infante, J., Rodríguez, E., Combarros, O., Mateo, I., Fontalba, A., Pascual, J., Oterino, A., Polo, J. M., Leno, C. & Berciano, J. 2006. LRRK2 G2019S is a common mutation in Spanish patients with late-onset Parkinson's disease. *Neuroscience letters,* 395**,** 224-226.

Kalinderi, K., Fidani, L., Bostantjopoulou, S., Katsarou, Z. & Kotsis, A. 2007. The G2019S LRRK2 mutation is uncommon amongst Greek patients with sporadic Parkinson's disease. *European journal of neurology,* 14**,** 1088-1090.

Keller, M. F., Saad, M., Bras, J., Bettella, F., Nicolaou, N., Simón-Sánchez, J., Mittag, F., Büchel, F., Sharma, M. & Gibbs, J. R. 2012. Using genome-wide complex trait analysis to quantify 'missing heritability'in Parkinson's disease. *Human molecular genetics,* 21**,** 4996-5009.

Khan, N. L., Jain, S., Lynch, J. M., Pavese, N., Abou-Sleiman, P., Holton, J. L., Healy, D. G., Gilks, W. P., Sweeney, M. G. & Ganguly, M. 2005. Mutations in the gene LRRK2 encoding dardarin (PARK8) cause familial Parkinson's disease: clinical, pathological, olfactory and functional imaging and genetic data. *Brain,* 128**,** 2786-2796.

Kim, W.-G., Mohney, R. P., Wilson, B., Jeohn, G.-H., Liu, B. & Hong, J.-S. 2000. Regional difference in susceptibility to lipopolysaccharide-induced neurotoxicity in the rat brain: role of microglia. *Journal of Neuroscience,* 20**,** 6309-6316.

Kitada, T., Asakawa, S., Hattori, N., Matsumine, H., Yamamura, Y., Minoshima, S., Yokochi, M., Mizuno, Y. & Shimizu, N. 1998. Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. *Nature,* 392**,** 605-8.

Latourelle, J. C., Hendricks, A. E., Pankratz, N., Wilk, J. B., Halter, C., Nichols, W. C., Gusella, J. F., Destefano, A. L., Myers, R. H. & Foroud, T. 2011. Genomewide linkage study of modifiers of LRRK2-related Parkinson's disease. *Movement Disorders,* 26**,** 2039-2044.

Lee, A. J., Wang, Y., Alcalay, R. N., Mejia-Santana, H., Saunders-Pullman, R., Bressman, S., Corvol, J. C., Brice, A., Lesage, S., Mangone, G., Tolosa, E., Pont-Sunyer, C., Vilas, D., Schule, B., Kausar, F., Foroud, T., Berg, D., Brockmann, K., Goldwurm, S., Siri, C., Asselta, R., Ruiz-Martinez, J., Mondragon, E., Marras, C., Ghate, T., Giladi, N., Mirelman, A., Marder, K. & Michael, J. F. L. C. C. 2017a. Penetrance estimate of LRRK2 p.G2019S mutation in individuals of non-Ashkenazi Jewish ancestry. *Mov Disord*.

Lee, B. D., Dawson, V. L. & Dawson, T. M. 2012. Leucine-rich repeat kinase 2 (LRRK2) as a potential therapeutic target in Parkinson's disease. *Trends in pharmacological sciences,* 33**,** 365-373.

Lee, H., James, W. S. & Cowley, S. A. 2017b. LRRK2 in peripheral and central nervous system innate immunity: its link to Parkinson9s disease. *Biochemical Society Transactions,* 45**,** 131-139.

Lesage, S., Dürr, A., Tazir, M., Lohmann, E., Leutenegger, A.-L., Janin, S., Pollak, P. & Brice, A. 2006. LRRK2 G2019S as a cause of Parkinson's disease in North African Arabs. *New England Journal of Medicine,* 354**,** 422-423.

Lesage, S., Ibanez, P., Lohmann, E., Pollak, P., Tison, F., Tazir, M., Leutenegger, A. L., Guimaraes, J., Bonnet, A. M. & Agid, Y. 2005. G2019S LRRK2 mutation in French and North African families with Parkinson's disease. *Annals of neurology,* 58**,** 784-787.

Lesage, S., Patin, E., Condroyer, C., Leutenegger, A.-L., Lohmann, E., Giladi, N., Bar-Shira, A., Belarbi, S., Hecham, N. & Pollak, P. 2010. Parkinson's disease-related LRRK2 G2019S mutation results from independent mutational events in humans. *Human molecular genetics,* 19**,** 1998-2004.

Lewis, P. A. & Manzoni, C. 2012. LRRK2 and human disease: a complicated question or a question of complexes? *Sci Signal,* 5**,** pe2.

Lill, C. M. 2016. Genetics of Parkinson's disease. *Mol Cell Probes,* 30**,** 386-396.

Liu, M., Bender, S. A., Cuny, G. D., Sherman, W., Glicksman, M. & Ray, S. S. 2013. Type II kinase inhibitors show an unexpected inhibition mode against Parkinson's disease-linked LRRK2 mutant G2019S. *Biochemistry,* 52**,** 1725-1736.

Liu, Z., Lee, J., Krummey, S., Lu, W., Cai, H. & Lenardo, M. J. 2011. The kinase LRRK2 is a regulator of the transcription factor NFAT that modulates the severity of inflammatory bowel disease. *Nature immunology,* 12**,** 1063-1070.

Lubbe, S. J., Escott-Price, V., Gibbs, J. R., Nalls, M. A., Bras, J., Price, T. R., Nicolas, A., Jansen, I. E., Mok, K. Y., Pittman, A. M., Tomkins, J. E., Lewis, P. A., Noyce, A. J., Lesage, S., Sharma, M., Schiff, E. R., Levine, A. P., Brice, A., Gasser, T., Hardy, J., Heutink, P., Wood, N., Singleton, A. B., Williams, N. M., Morris, H. R. & Genomics, I. P. S. D. 2016. Additional rare variant analysis in Parkinson's disease cases with and without known pathogenic mutations: evidence for oligogenic inheritance. *Human Molecular Genetics,* 25**,** 5483-5489.

Macleod, D. A., Rhinn, H., Kuwahara, T., Zolin, A., Di Paolo, G., Mccabe, B. D., Marder, K. S., Honig, L. S., Clark, L. N., Small, S. A. & Abeliovich, A. 2013. RAB7L1 interacts with LRRK2 to modify intraneuronal protein sorting and Parkinson's disease risk. *Neuron,* 77**,** 425-39.

Manzoni, C., Denny, P., Lovering, R. C. & Lewis, P. A. 2015. Computational analysis of the LRRK2 interactome. *PeerJ,* 3**,** e778.

Manzoni, C., Mamais, A., Dihanich, S., Mcgoldrick, P., Devine, M. J., Zerle, J., Kara, E., Taanman, J. W., Healy, D. G., Marti-Masso, J. F., Schapira, A. H., Plun-Favreau, H., Tooze, S., Hardy, J., Bandopadhyay, R. & Lewis, P. A. 2013. Pathogenic Parkinson's disease mutations across the functional domains of LRRK2 alter the autophagic/lysosomal response to starvation. *Biochem Biophys Res Commun,* 441**,** 862-6.

Marder, K., Wang, Y., Alcalay, R. N., Mejia-Santana, H., Tang, M.-X., Lee, A., Raymond, D., Mirelman, A., Saunders-Pullman, R. & Clark, L. 2015. Age-specific penetrance of LRRK2 G2019S in the Michael J. Fox Ashkenazi Jewish LRRK2 Consortium. *Neurology,* 85**,** 89-95.

Marongiu, R., Ghezzi, D., Ialongo, T., Soleti, F., Elia, A., Cavone, S., Albanese, A., Altavista, M. C., Barone, P. & Brusa, L. 2006. Frequency and phenotypes of LRRK2 G2019S mutation in Italian patients with Parkinson's disease. *Movement disorders,* 21**,** 1232-1235.

Mata, I., Ross, O. A., Kachergus, J., Huerta, C., Ribacoba, R., Moris, G., Blazquez, M., Guisasola, L., Salvador, C. & Martinez, C. 2006. LRRK2 mutations are a common cause of Parkinson's disease in Spain. *European journal of neurology,* 13**,** 391-394.

Mata, I. F., Cosentino, C., Marca, V., Torres, L., Mazzetti, P., Ortega, O., Raggio, V., Aljanati, R., Buzó, R. & Yearout, D. 2009. LRRK2 mutations in patients with Parkinson's disease from Peru and Uruguay. *Parkinsonism & related disorders,* 15**,** 370-373.

Meixner, A., Boldt, K., Van Troys, M., Askenazi, M., Gloeckner, C. J., Bauer, M., Marto, J. A., Ampe, C., Kinkl, N. & Ueffing, M. 2011. A QUICK Screen for Lrrk2 Interaction Partners - Leucine-rich Repeat Kinase 2 is Involved in Actin Cytoskeleton Dynamics. *Molecular & Cellular Proteomics,* 10.

Monfrini, E. & Di Fonzo, A. 2017. Leucine-Rich Repeat Kinase (LRRK2) Genetics and Parkinson's Disease. *Leucine-Rich Repeat Kinase 2 (LRRK2).* Springer.

Moss, D. J. H., Pardiñas, A. F., Langbehn, D., Lo, K., Leavitt, B. R., Roos, R., Durr, A., Mead, S., Coleman, A. & Santos, R. D. 2017. Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. *The Lancet Neurology*.

Motulsky, A. G. 1995. Jewish diseases and origins. *Nature genetics,* 9**,** 99-101.

Nalls, M. A., Bras, J., Hernandez, D. G., Keller, M. F., Majounie, E., Renton, A. E., Saad, M., Jansen, I., Guerreiro, R. & Lubbe, S. 2015. NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases. *Neurobiology of aging,* 36**,** 1605. e7-1605. e12.

Nalls, M. A., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., Saad, M., Destefano, A. L., Kara, E., Bras, J., Sharma, M., Schulte, C., Keller, M. F., Arepalli, S., Letson, C., Edsall, C., Stefansson, H., Liu, X., Pliner, H., Lee, J. H., Cheng, R., International Parkinson's Disease Genomics, C., Parkinson's Study Group Parkinson's Research: The Organized, G. I., Andme, Genepd, Neurogenetics Research, C., Hussman Institute of Human, G., Ashkenazi Jewish Dataset, I., Cohorts For, H., Aging Research in Genetic, E., North American Brain Expression, C., United Kingdom Brain Expression, C., Greek Parkinson's Disease, C., Alzheimer Genetic Analysis, G., Ikram, M. A., Ioannidis, J. P., Hadjigeorgiou, G. M., Bis, J. C., Martinez, M., Perlmutter, J. S., Goate, A., Marder, K., Fiske, B., Sutherland, M., Xiromerisiou, G., Myers, R. H., Clark, L. N., Stefansson, K., Hardy, J. A., Heutink, P., Chen, H., Wood, N. W., Houlden, H., Payami, H., Brice, A., Scott, W. K., Gasser, T., Bertram, L., Eriksson, N., Foroud, T. & Singleton, A. B. 2014. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat Genet,* 46**,** 989-93.

Nichols, W. C., Pankratz, N., Hernandez, D., Paisán-Ruíz, C., Jain, S., Halter, C. A., Michaels, V. E., Reed, T., Rudolph, A. & Shults, C. W. 2005. Genetic screening for a single common LRRK2 mutation in familial Parkinson's disease. *The Lancet,* 365**,** 410-412.

Nixon-Abell, J., Berwick, D. C., Grannó, S., Spain, V. A., Blackstone, C. & Harvey, K. 2016. Protective LRRK2 R1398H variant enhances GTPase and Wnt signaling activity. *Frontiers in molecular neuroscience,* 9.

Noyce, A. J., Kia, D. A., Hemani, G., Nicolas, A., Price, T. R., De Pablo-Fernandez, E., Haycock, P. C., Lewis, P. A., Foltynie, T., Davey Smith, G., International Parkinson Disease Genomics, C., Schrag, A., Lees, A. J., Hardy, J., Singleton, A., Nalls, M. A., Pearce, N., Lawlor, D. A. & Wood, N. W. 2017. Estimating the causal influence of body mass index on risk of Parkinson disease: A Mendelian randomisation study. *PLoS Med,* 14**,** e1002314.

Orr-Urtreger, A., Shifrin, C., Rozovski, U., Rosner, S., Bercovich, D., Gurevich, T., Yagev-More, H., Bar-Shira, A. & Giladi, N. 2007. The LRRK2 G2019S mutation in Ashkenazi Jews with Parkinson disease Is there a gender effect? *Neurology,* 69**,** 1595-1602.

Ozelius, L. J., Senthil, G., Saunders-Pullman, R., Ohmann, E., Deligtisch, A., Tagliati, M., Hunt, A. L., Klein, C., Henick, B. & Hailpern, S. M. 2006. LRRK2

G2019S as a cause of Parkinson's disease in Ashkenazi Jews. *New England Journal of Medicine,* 354**,** 424-425.

Paisan-Ruiz, C., Jain, S., Evans, E. W., Gilks, W. P., Simon, J., Van Der Brug, M., De Munain, A. L., Aparicio, S., Gil, A. M., Khan, N., Johnson, J., Martinez, J. R., Nicholl, D., Carrera, I. M., Pena, A. S., De Silva, R., Lees, A., Marti-Masso, J. F., Perez-Tur, J., Wood, N. W. & Singleton, A. B. 2004. Cloning of the gene containing mutations that cause PARK8-linked Parkinson's disease. *Neuron,* 44**,** 595-600.

Paisán-RuíZ, C., Jain, S., Evans, E. W., Gilks, W. P., Simón, J., Van Der Brug, M., De Munain, A. L., Aparicio, S., Gil, A. M. N. & Khan, N. 2004. Cloning of the gene containing mutations that cause PARK8-linked Parkinson's disease. *Neuron,* 44**,** 595-600.

Papapetropoulos, S., Adi, N., Shehadeh, L., Bishopric, N., Singer, C., Argyriou, A. A. & Chroni, E. 2008. Is the G2019S LRRK2 mutation common in all southern European populations? *Journal of Clinical Neuroscience,* 15**,** 1027-1030.

Papapetropoulos, S., Argyriou, A., Mitsi, G. & Chroni, E. 2007. The G2019S LRRK2 mutation is uncommon amongst Greek patients with familial Parkinson's disease. *European journal of neurology,* 14.

Pellicano, C., Benincasa, D., Pisani, V., Buttarelli, F. R., Giovannelli, M. & Pontieri, F. E. 2007. Prodromal non-motor symptoms of Parkinson's disease. *Neuropsychiatric disease and treatment,* 3**,** 145.

Perez-Pastene, C., Cobb, S. A., Díaz-Grez, F., Hulihan, M. M., Miranda, M., Venegas, P., Godoy, O. T., Kachergus, J. M., Ross, O. A. & Layson, L. 2007. Lrrk2 mutations in South America: A study of Chilean Parkinson's disease. *Neuroscience letters,* 422**,** 193-197.

Pirkevi, C., Lesage, S., Condroyer, C., Tomiyama, H., Hattori, N., Ertan, S., Brice, A. & Başak, A. 2009. A LRRK2 G2019S mutation carrier from Turkey shares the Japanese haplotype. *neurogenetics,* 10**,** 271-273.

Polymeropoulos, M. H., Lavedan, C., Leroy, E., Ide, S. E., Dehejia, A., Dutra, A., Pike, B., Root, H., Rubenstein, J., Boyer, R., Stenroos, E. S., Chandrasekharappa, S., Athanassiadou, A., Papapetropoulos, T., Johnson, W. G., Lazzarini, A. M., Duvoisin, R. C., Di Iorio, G., Golbe, L. I. & Nussbaum, R. L. 1997. Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science,* 276**,** 2045-7.

Ramirez, A., Heimbach, A., Gruendemann, J., Stiller, B., Hampshire, D., Cid, L. P., Goebel, I., Mubaidin, A. F., Wriekat, A. L., Roeper, J., Al-Din, A., Hillmer, A. M., Karsak, M., Liss, B., Woods, C. G., Behrens, M. I. & Kubisch, C. 2006. Hereditary parkinsonism with dementia is caused by mutations in ATP13A2, encoding a lysosomal type 5 P-type ATPase. *Nature Genetics,* 38**,** 1184-1191.

Rideout, H. J. 2017. Leucine-Rich Repeat Kinase 2 (LRRK2). Springer.

Risch, N. J. 2000. Searching for genetic determinants in the new millennium. *Nature,* 405**,** 847.

Savchenko, V., Mckanna, J., Nikonenko, I. & Skibo, G. 2000. Microglia and astrocytes in the adult rat brain: comparative immunocytochemical analysis demonstrates the efficacy of lipocortin 1 immunoreactivity. *Neuroscience,* 96**,** 195-203.

Schapansky, J., Nardozzi, J. D., Felizia, F. & Lavoie, M. J. 2014. Membrane recruitment of endogenous LRRK2 precedes its potent regulation of autophagy. *Hum Mol Genet,* 23**,** 4201-14.

Shan, Y., Seeliger, M. A., Eastwood, M. P., Frank, F., Xu, H., Jensen, M. Ø., Dror, R. O., Kuriyan, J. & Shaw, D. E. 2009. A conserved protonation-dependent switch controls drug binding in the Abl kinase. *Proceedings of the National Academy of Sciences,* 106**,** 139-144.

Sheerin, U.-M., Houlden, H. & Wood, N. W. 2014. Advances in the Genetics of Parkinson's Disease: A Guide for the Clinician
. *Movement Disorders Clinical Practice*

Shin, N., Jeong, H., Kwon, J., Heo, H. Y., Kwon, J. J., Yun, H. J., Kim, C. H., Han, B. S., Tong, Y., Shen, J., Hatano, T., Hattori, N., Kim, K. S., Chang, S. & Seol, W. 2008. LRRK2 regulates synaptic vesicle endocytosis. *Exp Cell Res,* 314**,** 2055-65.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet,* 15**,** 121-32.

Spanaki, C., Latsoudis, H. & Plaitakis, A. 2006. LRRK2 mutations on Crete: R1441H associated with PD evolving to PSP. *Neurology,* 67**,** 1518-1519.

Spillantini, M. G., Crowther, R. A., Jakes, R., Hasegawa, M. & Goedert, M. 1998. alpha-Synuclein in filamentous inclusions of Lewy bodies from Parkinson's disease and dementia with lewy bodies. *Proc Natl Acad Sci U S A,* 95**,** 6469-73.

Steger, M., Tonelli, F., Ito, G., Davies, P., Trost, M., Vetter, M., Wachter, S., Lorentzen, E., Duddy, G. & Wilson, S. 2016. Phosphoproteomics reveals that Parkinson's disease kinase LRRK2 regulates a subset of Rab GTPases. *Elife,* 5**,** e12813.

Taliun, D., Chothani, S. P., Schönherr, S., Forer, L., Boehnke, M., Abecasis, G. R. & Wang, C. 2017. LASER server: ancestry tracing with genotypes or sequence reads. *Bioinformatics***,** btx075.

Taymans, J.-M. 2017. Regulation of LRRK2 by Phosphatases. *Leucine-Rich Repeat Kinase 2 (LRRK2).* Springer.

Tomiyama, H., Li, Y., Funayama, M., Hasegawa, K., Yoshino, H., Kubo, S. I., Sato, K., Hattori, T., Lu, C. S. & Inzelberg, R. 2006. Clinicogenetic study of mutations in LRRK2 exon 41 in Parkinson's disease patients from 18 countries. *Movement disorders,* 21**,** 1102-1108.

Trabzuni, D., Ryten, M., Emmett, W., Ramasamy, A., Lackner, K. J., Zeller, T., Walker, R., Smith, C., Lewis, P. A. & Mamais, A. 2013. Fine-mapping, gene expression and splicing analysis of the disease associated LRRK2 locus. *PloS one,* 8**,** e70724.

Trinh, J., Gustavsson, E. K., Vilarino-Guell, C., Bortnick, S., Latourelle, J., Mckenzie, M. B., Tu, C. S., Nosova, E., Khinda, J., Milnerwood, A., Lesage, S., Brice, A., Tazir, M., Aasly, J. O., Parkkinen, L., Haytural, H., Foroud, T., Myers, R. H., Sassi, S. B., Hentati, E., Nabli, F., Farhat, E., Amouri, R., Hentati, F. & Farrer, M. J. 2016. DNM3 and genetic modifiers of age of onset in LRRK2 Gly2019Ser parkinsonism: a genome-wide linkage and association study. *Lancet Neurol,* 15**,** 1248-1256.

Valente, E. M., Abou-Sleiman, P. M., Caputo, V., Muqit, M. M., Harvey, K., Gispert, S., Ali, Z., Del Turco, D., Bentivoglio, A. R., Healy, D. G., Albanese, A., Nussbaum, R., Gonzalez-Maldonado, R., Deller, T., Salvi, S., Cortelli, P., Gilks, W. P., Latchman, D. S., Harvey, R. J., Dallapiccola, B., Auburger, G. & Wood, N. W. 2004. Hereditary early-onset Parkinson's disease caused by mutations in PINK1. *Science,* 304**,** 1158-60.

Wakabayashi, K., Tanji, K., Mori, F. & Takahashi, H. 2007. The Lewy body in Parkinson's disease: molecules implicated in the formation and degradation of alpha-synuclein aggregates. *Neuropathology,* 27**,** 494-506.

Wang, X. L., Yan, M. H., Fujioka, H., Liu, J., Wilson-Delfosse, A., Chen, S. G., Perry, G., Casadesus, G. & Zhu, X. W. 2012. LRRK2 regulates mitochondrial dynamics and function through direct interaction with DLP1. *Human Molecular Genetics,* 21**,** 1931-1944.

Williams-Gray, C., Goris, A., Foltynie, T., Brown, J., Maranian, M., Walton, A., Compston, D., Sawcer, S. & Barker, R. 2006. Prevalence of the LRRK2 G2019S mutation in a UK community based idiopathic Parkinson's disease cohort. *Journal of Neurology, Neurosurgery & Psychiatry,* 77**,** 665-667.

Wooten, G., Currie, L., Bovbjerg, V., Lee, J. & Patrie, J. 2004. Are men at greater risk for Parkinson's disease than women? *Journal of Neurology, Neurosurgery & Psychiatry,* 75**,** 637-639.

Zabetian, C., Samii, A., Mosley, A., Roberts, J., Leis, B., Yearout, D., Raskind, W. & Griffith, A. 2005. A clinic-based study of the LRRK2 gene in Parkinson disease yields new mutations. *Neurology,* 65**,** 741-744.

Zabetian, C. P., Hutter, C. M., Yearout, D., Lopez, A. N., Factor, S. A., Griffith, A., Leis, B. C., Bird, T. D., Nutt, J. G. & Higgins, D. S. 2006. LRRK2 G2019S in families with Parkinson disease who originated from Europe and the Middle East: evidence of two distinct founding events beginning two millennia ago. *The American Journal of Human Genetics,* 79**,** 752-758.

Zhang, F.-R., Huang, W., Chen, S.-M., Sun, L.-D., Liu, H., Li, Y., Cui, Y., Yan, X.-X., Yang, H.-T. & Yang, R.-D. 2009. Genomewide association study of leprosy. *New England Journal of Medicine,* 361**,** 2609-2618.

Zimprich, A., Biskup, S., Leitner, P., Lichtner, P., Farrer, M., Lincoln, S., Kachergus, J., Hulihan, M., Uitti, R. J. & Calne, D. B. 2004a. Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. *Neuron,* 44**,** 601-607.

Appendix

Quality control script

```
#Variant annotation

GATK_VariantAnnotator="/data/kronos/General_Software/jre1.7.0_67/bin/java -jar
/data/kronos/NGS_Software/GATK_v3_5/GenomeAnalysisTK.jar -R
/data/kronos/NGS_Reference/fasta/human_g1k_v37.fasta -T VariantAnnotator"

#Add dbSNP annotation to VCF file in GATK

#$GATK_VariantAnnotator -V $WorkingDirectory/$inputVCF --dbsnp
/data/kronos/NGS_Reference/GATK_refFiles/common_all.vcf -o $WorkingDirectory/db_$inputVCF

#mv db_$inputVCF $studyNAME.vcf

#Make a binary (bed) file for PLINK input:
plink --vcf $studyNAME.vcf --double-id --make-bed --out $studyNAME

#Update individual sex information:
plink --bfile $WorkingDirectory/$studyNAME --update-sex $WorkingDirectory/$sexINFO --make-bed
--out $WorkingDirectory/${studyNAMEeu
_updated_sex

#Split XY chromosomes
plink --bfile $WorkingDirectory/${studyNAME}_updated_sex --double-id --split-x hg19 --make-bed -
-out $WorkingDirectory/${studyNAME}_Xsplit

#Sex Check and removal of ambiguous
plink --bfile $WorkingDirectory/${studyNAME}_Xsplit --check-sex 0.2 0.7 --out
$WorkingDirectory/${studyNAME}
cat $WorkingDirectory/${studyNAME}.sexcheck | awk '($3=="0" || $5=="OK"){print $1"\t"$2}' >
Sex_samples_to_keep
plink --bfile $WorkingDirectory/${studyNAME}_Xsplit --keep Sex_samples_to_keep --make-bed --
out $WorkingDirectory/${studyNAME}_Xsplit_SexPruned

#remove SNPs that have a call rate less than 90%
plink --bfile $WorkingDirectory/${studyNAME}_Xsplit_SexPruned --missing --out
$WorkingDirectory/${studyNAME}_missingness_before_bad_SNP_removal
plink --bfile $WorkingDirectory/${studyNAME}_Xsplit_SexPruned --geno 0.1 --make-bed --out
$WorkingDirectory/${studyNAME}_Xsplit_SexPruned_90pc_call_rate

#Removal of samples with high missingness
plink --bfile $WorkingDirectory/${studyNAME}_Xsplit_SexPruned_90pc_call_rate --missing --out
$WorkingDirectory/${studyNAME}_missingness_after_bad_SNP_removal
plink --bfile $WorkingDirectory/${studyNAME}_Xsplit_SexPruned_90pc_call_rate --mind 0.2 --
make-bed --out
$WorkingDirectory/${studyNAME}_Xsplit_SexPruned_90pc_call_rate_80pc_sample_call_rate
plink --bfile
$WorkingDirectory/${studyNAME}_Xsplit_SexPruned_90pc_call_rate_80pc_sample_call_rate --
missing --out
$WorkingDirectory/${studyNAME}_missingness_after_bad_SNP_removal_and_bad_sample_removal

#Hardy-Weinberg Equilibrium
plink --bfile
$WorkingDirectory/${studyNAME}_Xsplit_SexPruned_90pc_call_rate_80pc_sample_call_rate --
hardy --out
```

```
$WorkingDirectory/${studyNAME}_missingness_after_bad_SNP_removal_and_bad_sample_removal
_HWE

#Filter:
plink --bfile
$WorkingDirectory/${studyNAME}_Xsplit_SexPruned_90pc_call_rate_80pc_sample_call_rate --hwe
0.001 --make-bed --out
$WorkingDirectory/${studyNAME}_Xsplit_SexPruned_90pc_call_rate_80pc_sample_call_rate_HWE
_filtered

#Het outlier check and removal
plink --bfile
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC/${studyNAME}_post_vari
antANDsampleGenotypeQC --het --out
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_het
awk ' $6 >= 0.2 || $6 <= -0.2 '
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_het.het >
$WorkingDirectory/het_samples_to_remove.txt
plink --bfile
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC/${studyNAME}_post_vari
antANDsampleGenotypeQC --remove $WorkingDirectory/het_samples_to_remove.txt --make-bed --
out $WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails

#Remove low MAF SNPs

plink --bfile $WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails
--maf 0.01 --make-bed --out
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails1

#Prune SNPs in LD
plink --bfile
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails1 --indep-
pairwise 50 5 0.5 --out
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails_LD_Prune
d
#Use only rs numbers here
awk '$1 ~ /^rs/'
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails_LD_Prune
d.prune.in >
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails_LD_Prune
d.prune2.in
plink --bfile $WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails
--extract
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails_LD_Prune
d.prune2.in --make-bed --out
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails_LD_Prune
d
#IBD calculation:
plink --bfile
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails_LD_Prune
d --genome --min 0.1 --out
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails_LD_Prune
d

#IBS Calculation
plink --bfile
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails_LD_Prune
d --cluster --neighbour 1 5 --out
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails_LD_Prune
d_IBS
```

```
#----> Neighborhood Z scores (visually Inspect to remove outliers that slip off at the end(low) )

#PCA visualisation

plink --bfile
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails_LD_Prune
d --extract /data/kronos/NGS_Reference/HapMap_Refernce/hapmap3r2_CEU.CHB.JPT.YRI.no-at-cg-
snps.txt --make-bed --out
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails_LD_Prune
d.hapmap-snps

#Discover and remove multi-allelic SNPs
plink --bfile
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails_LD_Prune
d.hapmap-snps --bmerge
/data/kronos/NGS_Reference/HapMap_Refernce/hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-
snps.bed
/data/kronos/NGS_Reference/HapMap_Refernce/hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-
snps.bim
/data/kronos/NGS_Reference/HapMap_Refernce/hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-
snps.fam
plink --bfile
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails_LD_Prune
d.hapmap-snps --exclude $WorkingDirectory/plink.missnp --make-bed --out
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails_LD_Prune
d.hapmap-snps2
plink --bfile
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_less_het_fails_LD_Prune
d.hapmap-snps2 --bmerge
/data/kronos/NGS_Reference/HapMap_Refernce/hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-
snps.bed
/data/kronos/NGS_Reference/HapMap_Refernce/hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-
snps.bim
/data/kronos/NGS_Reference/HapMap_Refernce/hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-
snps.fam --out $WorkingDirectory/${studyNAME}_merged_with_HapMap_for_PCA

#Remove warning SNPs
#grep 'Warning: Multiple' $WorkingDirectory/${studyNAME}_merged_with_HapMap_for_PCA.log
> $WorkingDirectory/removal1.txt
#awk '{print $7}' $WorkingDirectory/removal1.txt > $WorkingDirectory/removal2.txt
#sed -i 's/^.//' $WorkingDirectory/removal2.txt
#sed -i 's/..$//' $WorkingDirectory/removal2.txt

#And additional warning SNPs
#plink --bfile $WorkingDirectory/${studyNAME}_merged_with_HapMap_for_PCA --exclude
$WorkingDirectory/removal2.txt --make-bed --out
$WorkingDirectory/${studyNAME}_merged_with_HapMap_for_PCA_clean

#Prepare for PCA

#GCTA
gcta --bfile $WorkingDirectory/${studyNAME}_merged_with_HapMap_for_PCA --make-grm --
autosome --thread-num 10 --out $WorkingDirectory/${studyNAME}_matrix
gcta --grm $WorkingDirectory/${studyNAME}_matrix --pca 4

#Select samples for association analysis
plink --bfile
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_PopulationStratification/$
{studyNAME}_post_variantANDsampleGenotypeQC_PopulationStratification --make-bed --remove
$SAMPLEEXLUSION --out
```

$WorkingDirectory/${studyNAME}_Analysis_Ready_Variants/${studyNAME}_Analysis_Ready_Va
riants

cp $WorkingDirectory/*.log $WorkingDirectory/${studyNAME}_analysis_log_files
cp $WorkingDirectory/*.eigenval
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_PopulationStratification/
cp $WorkingDirectory/*.eigenvec
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_PopulationStratification/

cp $WorkingDirectory/*.id
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_PopulationStratification/
cp $WorkingDirectory/*.bin
$WorkingDirectory/${studyNAME}_post_variantANDsampleGenotypeQC_PopulationStratification/

Primer sequences

SNP	Left primer	Right primer
G2019S	tgggtctttgcctgagataga	tgactcttctgaactcacatctg
Rs2206543	CTCTAATCGCTCCCTTTCCCT	GGCATGCCTAAAGACCTAAGG