Perturbed Molecular Pathways in Parkinson's Disease

Stephanie Millin

Christ Church College

MRC Functional Genomics Unit

Department of Physiology, Anatomy and Genetics

University of Oxford

A thesis submitted in partial fulfilment of the requirements of

Doctor of Philosophy

Hilary / Trinity term 2017

Abstract

Parkinson's Disease (PD) is the most common movement disorder and second most common neurodegenerative disorder, affecting 1 in every 100 people over the age of 60. It is a heterogeneous disorder whose pathology and causes remain incompletely understood. Identification of genetic risk factors can provide valuable understanding of the disease process and pave the way for the development of novel treatment.

Firstly, eQTLs were identified that affected the expression of functionally related PD-linked gene pairs and were within PD associated genomic regions. This was achieved by integrating multiple data sources into a network tailored to PD, then interrogating this in tandem with genome-wide association study and eQTL data. Four eQTLs were identified, two affecting LRRK2. The genotype conferring greatest additive increase in LRRK2 expression was significantly over-represented among two independent case populations but not among controls.

Secondly, Copy Number Variants were classified by their functional annotations to identify common molecular pathways on which PD-linked variation converged. Seven pathways were enriched among PD patients, two of which remained so after independently significant variation within PARK2 was removed. However this was not replicated in an independent cohort.

Thirdly genome-wide association studies were carried out first comparing PD case and control and second comparing phenotypic subtypes among PD cases. Enrichment analysis identified two pathways significantly

associated with disease onset and implicated a subset of one with a specific phenotypic subgroup.

Finally, continuous phenotypic variation was analysed. Phenotypic axes were identified each representing multiple co-varying phenotypes. Genome-wide genetic analyses of these identified 10 genomic regions significantly affecting the severity of specific measured phenotypes.

This work implicates genetic variation in mediating both PD onset and phenotypic progression and yields insight into the common molecular pathways that may be involved. A novel method of quantifying patient phenotype was also developed that should facilitate future analysis.

Word count: 48,214

Acknowledgements

Firstly I'd like to thank my parents, sister and family for all their support and for always believing in me. You gave me the confidence to begin my DPhil and the motivation to finish. It's thanks to you that I am where I am today.

I'd also like to thank Sam McGrail, for always being there for me and for all the help and encouragement along the way. You push me to be the best that I can be and for that I am hugely grateful.

Thank you to Caleb for your unending support and encouragement and for always having confidence in me. Also for sharing with me your love and passion for science, which is a constant inspiration and bought out the best of my abilities.

Thank you to the entire Webber group and everyone within the OPDC, all of whom have helped hugely throughout my DPhil.

Thanks to Chris Ponting, for making time in your busy schedule to give me guidance and advice whenever it was needed.

Finally, I am very grateful for the funding provided by the Medical Research Council, and for the funding and support provided by Christ Church College, and would like to thank both accordingly.

Table of Contents

Chapter 1: Literature Review 11

1.1) Introduction 11

1.2) Pathology underlying PD 13

1.2.1) Oxidative stress 13

1.2.2) Mitochondrial dysfunction 14

1.2.3) Alpha-synuclein 16

1.2.4) Altered protein handling 18

1.2.5) Autophagy 19

1.2.6) Inflammation 21

1.3) Genetic causes of PD 22

1.3.1) SNCA 22

1.3.2) LRRK2 (PARK8) 24

1.3.3) Parkin (PARK2) 26

1.3.4) PINK1 (PARK6) 27

1.3.5) GBA 28

1.3.6) DJ-1 (PARK7) 29

1.3.7) Polygenic risk scoring 30

1.4) Environmental causes and risk factors 32

1.4.1) 1-methyl-1-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) 32

1.4.2) Rotenone and paraquat 32

1.4.3) Smoking 33

1.4.4) Caffeine 33

1.4.5) Non-steroidal anti-inflammatory drugs (NSAIDs) 34

1.5) Current treatments **35**

1.5.1) Levodopa 35

1.5.2) MAO-B inhibitors 36

1.5.3) Dopamine agonists 36

1.5.4) Anticholinergics 37

1.5.5) Deep Brain Stimulation (DBS) 37

1.5.6) Cell transplantation 38

1.6) Phenotypic aspects **39**

1.7) Thesis structure **42**

Chapter 2: Gene network analysis **44**

2.1) Introduction **44**

2.2) Methods **45**

2.2.1) Creating a gene network 45

2.2.2) Testing the network 50

2.2.3) EQTL analysis using the network 52

2.2.4) Modelling allelic co-inheritance 53

2.3) Results **55**

2.3.1) Network testing 55

2.3.2) EQTL analysis 56

2.4) Discussion **65**

2.4.1) Overview 65

2.4.2) Gene networks 65

2.4.3) EQTL analysis 66

2.4.4) Co-inheritance of LRRK2 variants 69

2.4.5) Conclusion 71

Chapter 3: CNV analysis **72**

3.1) Introduction **72**

3.2) Method **73**

3.2.1) Data 73

3.2.2) CNV calling and Quality control 73

3.2.3) Statistical analysis 78

3.2.4) Analysis of human PD phenotypic subgroups 80

3.2.5) Network analysis 80

3.3) Results **82**

3.3.1) Batch effects 84

3.3.2) Enrichment analysis 86

3.3.3) Pathway analysis 88

3.3.4) Phenotypic analysis in the OPDC dataset 93

3.3.5) Network analysis 95

3.4) Discussion **98**

3.4.1) Overall enrichment analysis 98

3.4.2) Pathway approaches 99

3.4.3) Network analysis 103

3.4.4) Phenotype analysis 105

3.4.5) Study limitations 106

3.4.6) Conclusion 108

Chapter 4: Binary phenotypes in Parkinson's Disease **110**

4.1) Introduction **110**

4.2) Methods **111**

4.2.1) Datasets 111

4.2.2) Quality control 112
4.2.3) Genotype imputation 113
4.2.4) Case-control GWAS 114
4.2.5) Discrete phenotype GWAS 114
4.2.6) Enrichment analysis 116
4.3) Results 117
4.3.1) Case-control analysis 120
4.3.2) Phenotypic analysis 126
4.4) Discussion 132
4.4.1) Imputation 132
4.4.2) Case-control analysis 132
4.4.3) K-means cluster analysis 135
4.4.4) Conclusion 139
Chapter 5: Phenotype-Genotype Analysis of the OPDC Discovery Cohort 141
5.1) Introduction 141
5.2) Methods 142
5.2.1) Data 142
5.2.2) Phenotype imputation and generation of latent components 142
5.2.3) Quantitative trait GWAS 151
5.2.4) Phenotype regression model 152
5.2.5) Polygenic risk scoring 154
5.3) Results 155
5.3.1) Phenotypic imputation 155
5.3.2) Generation of phenotype axes 159

5.3.3) Quantitative trait analysis of phenotypic axis scores 166

5.3.4) Elastic net regression 172

5.4) Discussion 186

5.4.1) Heritability estimates 191

5.4.2) Continuous phenotypes 194

5.4.3) Genetic models of phenotype severity 200

5.4.4) Study limitations 204

5.4.5) Conclusion 205

6) Concluding remarks 207

7) Bibliography 214

Chapter 1: Literature Review

1.1) Introduction

Parkinson's Disease (PD) is the most common movement disorder and second most common neurodegenerative disorder, affecting one in every 500 people and costing the UK £2billion every year [1]. Four motor symptoms are central to clinical diagnosis: tremor, rigidity, bradykinesia (slowness of movement) and postural instability. These are considered sufficient to differentiate PD from other disorders, but are frequently accompanied by a range of comorbid motor and non-motor symptoms.

Aging is the main risk factor for PD. Onset generally begins around 60 years of age with mild symptoms and affects roughly 1% of individuals. However frequency increases with aging and by the age of 80 prevalence rises to 4%. Disease progression causes symptoms to worsen and consequently aging is also a risk factor for more severe disease.

PD onset is caused by the degeneration and loss of dopaminergic neurons in the substantia nigra. This structure forms part of the basal ganglia in the brain along with the cortex, striatum, thalamus, globus pallidus and subthalamic nucleus. Its main role in healthy individuals is motor control. Neurons produce dopamine which is carried along axons to the striatum, forming the nigrostriatal pathway. This allows excitation of thalamocortical pathways and transmission of signals to the cerebral cortex, resulting in voluntary movement.

Neuronal loss occurs naturally with aging but is accelerated in PD. By the time clinical signs are recognised around 50% of nigral neurons and 80% of striatal dopamine are lost [2]. Estimates suggest that accelerated decline may begin up to 10 years previous to disease onset [3].

This process is often accompanied by the development of Lewy bodies: neuronal inclusions consisting of alpha-synuclein and other misfolded and mutated proteins. These aggregate in the cytoplasm and neurites, forming large deposits that damage the neuron. As disease progresses and more Lewy bodies accumulate this eventually leads to cell death, further accelerating neuronal decline.

The resulting dopamine deficiency reduces excitatory drive in the striatum, causing irregular stimulation of nerve cells. This disrupts voluntary motor control resulting in tremor, bradykinesia and rigidity. As degeneration of the dopaminergic nigrostriatal pathway progresses and dopamine depletion worsens, the severity of these and other motor symptoms increases.

The substantia nigra is also involved in learning, reward and temporal processing and consequently dopaminergic decline can result in cognitive impairment, novelty-seeking behaviour and sleep disturbances respectively. Although not globally observed these non-motor phenotypes are relatively common comorbidities of PD. Dementia affects 20-40% of individuals and sleep disturbance is reported in over 80% [4, 5]. Furthermore REM-sleep behaviour disorder (RBD) commonly precedes PD onset with over half of diagnosed individuals developing Parkinsonism within 10 years, possibly representing an early autonomic manifestation of neuronal loss [6].

1.2) Pathology underlying PD

The course of events leading to PD onset is incompletely understood. A number of altered cellular processes appear to be central to the development of PD pathology, however their disruption can arise through different mechanisms and few apply universally to the whole PD population. Several processes are inter-linked, so additive effects of perturbations within multiple pathways may accumulate to accelerate the rate of neuronal decline beyond that which the body can cope with during the normal aging process. It is therefore likely that multiple primary causes of disease pathology differ between individuals. The main factors thought to cause neuronal decline and the production of Lewy bodies are discussed here.

1.2.1) Oxidative stress

Oxidative stress results from the increased production of highly reactive free radicals, in particular reactive oxygen species (ROS). Although difficult to measure directly due to their reactivity there is much evidence to support their role in PD pathology. The destruction of nigral neurons by 6-OHDA and paraquat, toxins that mimic the PD phenotype in humans, occurs through autoxidation, NADPH oxidase activity, and lipid peroxidation, processes which all produce free radicals [7-10]. In post-mortem PD brain increased iron levels compared to controls and a shift in iron(II):iron(III) ratios toward the more oxidised ion are consistent with increased ROS burden [11, 12]. Finally, altered concentrations and activity of antioxidant enzymes occurs in PD substantia nigra. Decreased levels of reduced glutathione and increased levels of oxidated glutathione, in addition to increased superoxide dismutase activity, indicate a localised cellular

attempt to compensate for an additional free radical burden [13, 14]. As a result oxidative damage is caused to DNA, lipids and proteins in the substantia nigra, accumulating to eventually cause cell death [15-17].

A number of mechanisms responsible for the increase in ROS production have been proposed. The most probable cause seems to be mitochondrial dysfunction and alteration of the respiratory pathway. However other possibilities include iron accumulation, calcium channel activity, proteolysis, alpha-synuclein aggregation and the presence of mutant protein such as DJ-1 [18]. Multiple underlying causes could contribute to overall free radical accumulation within an individual.

1.2.2) Mitochondrial dysfunction

Mitochondria are responsible for cellular energy production and consequently are integral to cellular infrastructure. The energy demands of neurons are particularly high so any limitation of mitochondrial activity is likely to affect neuronal function and viability first. Increased fission, altered morphology and functional inhibition have been observed in both familial and sporadic PD [19-23]. These phenotypes reduce the ability of the mitochondria to meet the energy demands of the cell, increasing the likelihood of oxidative stress and cell death.

Complex I is the first enzyme of the mitochondrial electron transfer chain. It oxidises NADH and transfers electrons to either Ubiquinone or Coenzyme Q. Specific inhibition of complex I is observed in the substantia nigra of PD patients and in cell models of alpha-synuclein and mutant parkin expression [24-27]. Complex I blockade is also associated with the toxicity of rotenone and MPTP (1-

methyl-4-phenyl-1,2,3,6-tetrahydropyridine), chemical compounds that cause Parkinsonism in humans [28-30]. This results in leaked electrons which reduce oxygen and form ROS, damaging numerous cellular components [15-17, 31].

Mitophagy is the process by which damaged mitochondria are removed from the cell and is therefore an important factor in maintaining mitochondrial integrity. Perturbations in this pathway are known to play a key role in PD onset, particularly for familial forms linked to PINK1 and Parkin which are involved in the targeting of mitochondria for mitophagy. PINK1 localises to the outer mitochondrial membrane where sufficient membrane potential ensures it is continuously imported and degraded. However the membranes of damaged mitochondria become depolarised, resulting in a build-up of PINK1 on the outside. This accumulation recruits Parkin and activates its E3 ubiquitin ligase activity, causing it to build ubiquitin chains that flag the cell for degradation [32-35]. Loss of function of PINK1 or Parkin therefore impairs mitophagy and is also linked to increased mitochondrial fission [36]. The simultaneous increase in fragmented mitochondria and decrease in clearance capability causes an accumulation of dysfunctional organelles. These are unable to fulfil the bioenergetic demands of the cell, causing cell stress and death.

Loss of mitochondrial protective mechanisms is also linked to PD. DJ-1 is a neuroprotective protein that self-oxidises in response to oxidative stress and can then act as a scavenger protein [37]. However this function is lost in the PD-linked DJ-1 mutant. This reduces mitochondrial protection against ROS, causing an increase in mitochondrial dysfunction that doubles PD risk [37-39]. This link

between disease risk and mitochondrial vulnerability further highlights the importance of mitochondria in PD pathology.

Overall a large number of known PD risk variants and PD-linked toxins impart effects through perturbation of mitochondrial processes. Additionally mitochondria are likely to be the main cause of increased ROS production. Consequently there is very strong evidence for mitochondrial involvement in PD pathology.

1.2.3) Alpha-synuclein

Alpha-synuclein is a known pathological hallmark of PD and plays a central role in disease onset. It is a natively unfolded protein generally present in presynaptic terminals. However under certain circumstances it can adopt beta-sheet structures, forming fibrils that are the main constituent of Lewy bodies and Lewy neurites. These aggregates reflect aberrant localization of alpha-synuclein and have been found to degrade mitochondrial and nuclear DNA [40, 41]. This conformation can also migrate between cells, spreading to healthy neurons and acting as a template for the misfolding of native alpha-synuclein [42, 43]. This has led to the hypothesis that alpha-synuclein may impart a prion-like mode of propagation.

Several post-translational modifications affect the readiness of alpha-synuclein to adopt beta-sheet conformation. Oxidation of alpha-synuclein promotes its aggregation into fibrils and large multi-protein inclusions [44-47]. Nitrosylation of monomeric and dimeric alpha-synuclein also promotes fibril formation and nitrated alpha-synuclein is found in Lewy bodies [48, 49]. Both nitrosylation and oxidation occur as a result of free radical production,

suggesting a convergent link between alpha-synuclein and mitochondrial dysfunction in the development of Lewy body pathology.

Aggregated alpha-synuclein is thought to affect a wide variety of cellular processes. The first is that its relocation away from preynaptic terminals causes alterations to vesicle exocytosis. Over-expression and mutant alpha-synuclein studies show changes in the frequency and number of exocytotic events per stimulus [50]. Abnormal neurotransmitter release is likely to exacerbate the problems already caused by neurotransmitter deficiency, contributing to the development of motor symptoms.

Alpha-synuclein toxicity may also be mediated through interactions with endoplasmic reticulum (ER)-golgi pathways. Trafficking between these organelles, including that of the dopamine transporter (DAT), is blocked in the presence of both unfolded and aggregated alpha-synuclein [51-53]. Consequently ER stress is increased and fragmentation of golgi is observed that correlates with the presence of prefibrillar aggregates [52, 54]. Furthermore DAT is the main mechanism of dopamine clearance from the synapse. Trafficking problems can therefore result in prolonged dopamine stimulation of the post-synaptic neuron, in extreme cases causing dystonia. By stressing organelles and limiting dopamine reuptake alpha-synuclein can cause both the cell degeneration and dopaminergic dysfunction phenotypes associated with PD through this pathway.

Components of the protein degradation pathway are co-located with alpha-synuclein in Lewy bodies, indicating a role of this pathway also in alpha-synuclein-linked disease pathology [55]. Oligomers of alpha-synuclein are

normally degraded by the 26S proteasome [56]. However PD-associated alpha-synuclein mutations inhibit this function by binding to the proteasome and acting as an uptake inhibitor of other proteins [56-58]. The resulting accumulation of alpha-synuclein is likely to promote its aggregation and further inhibit proteasomal function, accelerating the disease process.

Finally, both wild-type and mutant alpha-synuclein interact with the mitochondrial respiratory chain and influence the susceptibility of the cell to oxidative stress. Wild type alpha-synuclein promotes mitochondrial fission and mutant alpha-synuclein causes over-active mitophagy, both resulting in a deficiency of functional mitochondria [59-61]. Complex I activity is also down-regulated by alpha-synuclein and even more so by mutant forms [27]. Consequently the cell experiences bioenergetic defects and increased ROS production arising from a direct link between alpha-synuclein and mitochondrial dysfunction.

Alpha-synuclein therefore has the capacity to disturb a wide range of cellular functions. Cell models have demonstrated that endogenous production of dopamine is required for alpha-synuclein toxicity, explaining the selectivity of cell death [62]. Consequently degeneration is limited to dopamine-producing neurons whereas non-dopamine-producing cells are spared.

1.2.4) Altered protein handling

The ubiquitin-proteasome system (UPS) is primarily responsible for the breakdown of short-lived proteins in the cytosol and nucleus. This pathway adds chains of ubiquitin molecules to damaged proteins, which are then targeted for degradation by the 26S proteasome. Interest in this pathway originally stemmed

from the role of the ubiquitin-protein ligase Parkin and the ubiquitin hydrolase UCH-L1 in this system, as pathogenic loss of function mutations in these genes is linked to disease onset [63, 64]. Experimental evidence now indicates a wider deficiency of this pathway also in sporadic PD.

In the substantia nigra of PD patients genes coding for ubiquitin-proteasome system proteins are down-regulated [65-67]. Alpha-subunits of the 26S proteasome are lost and levels of the PA700 proteasome activator are reduced [68]. Hydrolysing activities of the 26S proteasome are therefore compromised, leading to the impairment of a wide range of proteins specifically in the substantia nigra of PD patients [66, 69].

The action of proteasomal inhibitors in cell and animal models supports a primary role of the UPS in PD onset. These compounds selectively destroy dopaminergic neurons, resulting in an increase of ubiquitinated proteins, DNA fragmentation, labile iron and ROS [70, 71]. In mice these effects are observed preferentially in the substantia nigra and also cause the development of protein aggregates and motor impairments comparable to that observed in human PD [72]. Alterations in the UPS therefore occur selectively in nigral dopaminergic neurons of familial and sporadic PD patients and in cell and mouse models. The resulting impairment of protein regulation may be sufficient to independently cause the cellular and motor phenotypes and Lewy body pathology associated with PD.

1.2.5) Autophagy

Autophagy is the process by which long-lived cellular components such as organelles and proteins are degraded by lysosomes. This occurs in one of three

ways. Macroautophagy, of which mitophagy is a specific branch, involves the formation of autophagolysosomes (autophagic vacuoles): double membrane structures fused with the lysosomal membrane that engulf entire portions of cytoplasm. Microautophagy is the process by which cytoplasmic materials are taken in directly by the lysosome by membrane invagination. Finally, chaperone-mediated autophagy involves the transport of misfolded proteins directly across the lysosomal membrane.

Wild type alpha-synuclein is degraded by chaperone-mediated autophagy [73]. However in the substantia nigra and amygdala of PD patients the expression of chaperone-mediated autophagy proteins is significantly reduced, decreasing the metabolism and increasing the half-life of alpha-synuclein [74]. Mutation of the alpha-synuclein protein further prevents its degradation, resulting in the formation of toxic aggregates [73, 75]. The degree to which different alpha-synuclein mutants impede this pathway is correlated with their degree of toxicity, indicating a direct link between autophagy and PD risk [76] .

Lysosomes are a key component in chaperone-mediated autophagy, consequently their dysfunction can greatly increase PD risk. ATP13A2 is implicated in familial PD and encodes a lysosomal membrane protein. Its mutation affects a number of lysosomal attributes, increasing PD risk via abnormal alpha-synuclein handling [77, 78]. GBA mutations cause deficiency of the lysosomal enzyme GCase and increase PD risk over 5-fold [79]. This occurs via a positive feedback loop, whereby impaired lysosomal function causes an increase in alpha-synuclein abundance, which in turn inhibits GCase trafficking and further impairs alpha-synuclein degradation [80]. This ultimately results in

alpha-synuclein accumulation and the associated Lewy body pathology [81, 82]. GCase deficiency is also observed in the substantia nigra of sporadic PD patients, indicating this aetiology is not limited to familial PD cases [82].

Evidence suggests that macroautophagy may attempt to provide a coping mechanism to substitute for the PD-linked reduction in chaperone-mediated autophagy. An excess of autophagic vacuoles is observed in post-mortem PD patient substantia nigra and alpha-synuclein mutant cell models [57, 83]. Autophagy-related proteins are also found in Lewy bodies [74]. This indicates up-regulation of macroautophagy, likely in response to elevated alpha-synuclein aggregation. However it can also cause autophagic cell death through excess turnover of proteins and organelles, so may further exacerbate neurodegeneration [76]. Several autophagic pathways are therefore differentially altered in both sporadic and familial PD, converging on the mediation of alpha-synuclein toxicity and consequently Lewy body production.

1.2.6) Inflammation

GWAS have consistently linked the Human Leukocyte Antigen (HLA) region to PD onset, implicating components of the immune system in disease pathology [84]. Activated microglia are present in the substantia nigra of PD patients, indicating stimulation of the immune response specifically in this brain region [85]. These cells mediate inflammation and oxidative stress and can also directly induce cell death [86].

Changes in the concentration of several cytokines have been observed in brain, cerebrospinal fluid and serum of PD patients, indicating consistent up-regulation of the immune system throughout disease course [87-91]. This may

also reflect the severity of patient phenotype as serum levels of the cytokine RANTES correlate with UPDRS scores [92, 93]. However genes involved in neuroimmune signalling are not consistently over-expressed in PD patient brain compared to control, indicating that this may only be a subsidiary mechanism of disease onset [65-67, 94].

Over-activation of the neuroimmune system is a relatively consistent finding of genetic and pathological PD analyses, which can cause several cellular phenotypes associated with neurodegeneration. However it is unlikely to be the primary cause of PD onset for the majority of patients. Instead it is likely only to enhance the loss of dopaminergic neurons conferred by another mechanism [18].

1.3) Genetic causes of PD

Heritability of PD is estimated at up to 27%, yet only 5-10% patients have a known genetic cause. Familial linkage and genome-wide association studies have identified around 30 genetic variants linked to PD, but largely these modify risk rather than causing certain disease onset. In most cases familial and sporadic PD are indistinguishable, however some genetic variants are associated with characteristic phenotypes and progression. Importantly though the discovery of PD-linked genes has greatly aided understanding of the disease process.

1.3.1) SNCA

Alpha-synuclein is a 140 amino acid protein encoded by the SNCA gene that lies at the centre of PD pathology. It is abundant in the brain with lower

expression in the heart and muscles. Within the brain it is found largely in the pre-synaptic terminals of neurons, where it is involved in neurotransmitter release and vesicle turnover [95-97]. It is also present on the inner membrane of neuronal mitochondria, where it has a dose-dependent inhibitory effect on complex I activity [98, 99]. The amount of alpha-synuclein in mitochondria varies by brain region, but is highest among regions affected in PD including the olfactory bulb, hippocampus, striatum and thalamus [99].

There exist both risk and causal SNPs within the SNCA gene. Risk variants are relatively common and increase disease risk approximately 1.34-fold [100]. In contrast causal variants are rare, explaining just 2.5% of cases with no family history [101, 102]. Generally they consist of missense mutations that cause alterations to the normal protein structure. The abnormal alpha-synuclein is more susceptible to forming aggregates, causing widespread Lewy body pathology that affects the substantia nigra, locus ceruleus, cerebral cortex and hypothalamus [103]. This results in early-onset PD with fast progression, often accompanied by dementia and cognitive decline [104, 105].

Multiplications of the entire SNCA gene also cause PD in a dose-dependent manner [106, 107]. Duplications result in PD that presents almost identically to idiopathic PD with late age of onset and slow disease progression [108]. Triplication carriers have double the amount of alpha-synuclein present in blood and significantly increased amount in the brain [109]. Similarly to missense mutations this results in an early onset, rapidly progressing form of PD with poor long-term prognosis [110]. Penetrance of these SNCA mutations is high, with estimates up to 85% [101].

1.3.2) LRRK2 (PARK8)

Leucine-rich repeat kinase 2 is a large protein encoded by the LRRK2 gene. It is a complex protein of the ROCO superfamily which contains several domains and is consequently involved in multiple interactions. It is found mainly in the cytosol or mitochondrial outer membrane and is highly involved in the promotion of mitochondrial fission [111-114].

Several LRRK2 interactions may involve other PD-linked proteins. In disease-affected brain regions LRRK2 and alpha-synuclein co-localise in neurons and Lewy bodies. Furthermore the level of LRRK2 protein correlates with the increase in alpha-synuclein aggregation [115], possibly due phosphorylation of alpha-synuclein by LRRK2 [116]. Cell models also show evidence of an interaction between LRRK2 and Parkin [117]. This is replicated in fly models that also suggest an interaction between LRRK2 and PINK1 [118]. LRRK2 may therefore interact with the protein degradation pathway and directly with alpha-synuclein to mediate the formation of protein aggregates.

Point mutations in LRRK2 are associated with late-onset autosomal dominant and sporadic PD. Over 50 different missense and nonsense mutations have been observed. Of these R1141C, R1441G, R1441H, Y1699C, G2019S and I2020T are pathogenic, all of which are highly penetrant.

G2019S is the most common LRRK2 mutation and is involved in 2% of familial and 0.2% of sporadic cases in the UK [119]. Its frequency is highly population specific and it can explain up to 40% of PD cases in certain communities [120, 121]. This and the less common I2020T mutation both increase the kinase activity of the LRRK2 protein [122, 123]. Cell models show

that this results in decreased mitochondrial fusion and increased fission [111]. Fragmented mitochondria produce more ROS and less ATP, contributing towards elevated cell stress [124-126].

The R1441C/G/H and Y1699C mutations lie in the ROC and COR protein domains respectively. All of these variants weaken ROC-COR dimerization, reducing GTPase activity [127]. The mechanism by which this causes neurodegeneration is unknown as protein kinase activity is unchanged [128-131]. Post-translational modifications such as phosphorylation are altered which may disrupt multiple interactions with other proteins [131]. The autophagy and vesicle trafficking pathways have also been proposed to mediate toxicity [132]. The effects of these mutations may therefore be multi-faceted, however further study is required to pinpoint which alterations contribute directly to neurodegeneration.

Despite differences in molecular mode of action, clinical phenotypes of LRRK2 mutants are all largely indistinguishable from sporadic PD [133-135]. Mid- to late-onset disease is characterised by bradykinseia, rigidity, tremor, good levodopa response and low risk of dementia [135, 136]. Lewy body pathology is variable and segregates with particular mutations. It is frequently observed in carriers of the G2019S mutation and is correlated with non-motor phenotypes including cognitive impairment, anxiety and orthostatic hypotension [137, 138]. In contrast it is often absent in R1441C/G/H, Y1669C and I2020T patients [138-141].

1.3.3) Parkin (PARK2)

The PARK2 gene encodes Parkin, a 465 amino acid protein involved in the E3 ubiquitin ligase complex. This protein is involved in a number of interactions. Most prominently it mediates ubiquitin-dependent degradation in combination with PINK1. However it also directly promotes the degradation of Synphillin-1 and aminoacetyl-tRNA cofactor p38, both of which are found in Lewy bodies and whose over-expression is sufficient to cause protein aggregation [142-147]. Additionally, Parkin is able to attenuate the neurotoxicity of ataxin-2 and Parkin-associated endothelin receptor-like receptor [148, 149]. Consequently functional Parkin protects the cell from numerous endogenous sources of toxicity and protein aggregation. Correspondingly its' overexpression is neuroprotective and can reduce LRRK2- and alpha-synuclein-mediated neurodegeneration, whereas loss of function is linked to protein aggregation and the acceleration of neuronal decline [150-154].

Mutations in this gene cause autosomal recessive PD and are also thought to play a role in sporadic cases. Pathogenic Parkin mutations include single base pair substitutions, splice site mutations and small and large deletions, all of which are thought to cause loss of protein function. Studies have found missense mutations to cause decreased E3 ligase activity resulting in aberrant ubiquitination and impairment of proteasomal degradation [155, 156]. Solubility and aggregation properties of Parkin are also altered [157, 158]. These findings are consistent with PD-linked Parkin deletions spanning several exons, and indicate that diverse mutations within this gene are likely cause PD through similar cellular effects [159].

Familial Parkin mutations are associated with juvenile- and early-onset PD. Mean age of onset is 36 years for homozygous or compound heterozygous mutations and 50 years for heterozygous mutation [160, 161]. Patients generally present with dyskinesia in the lower body, motor fluctuations with symmetric onset and frequent dystonia [162-164]. Disease progression is slow with low risk of dementia [164, 165]. Phenotype onset therefore shows many similarities to sporadic PD but with earlier onset. Lewy body pathology is inconsistent as Parkin mutation carriers display both diffuse and absent Lewy bodies [166-170].

1.3.4) PINK1 (PARK6)

PINK1 encodes phosphatase and tensin homologue (PTEN)-induced putative kinase 1, a ubiquitously expressed serine/threonine-protein kinase found in the cytosol and on the mitochondrial outer membrane. In combination with Parkin it targets dysfunctional mitochondria for degradation, thereby ensuring sufficient energy production and minimal generation of ROS. It also facilitates the creation of vesicles that can separate ROS and transport them to lysosomes for degradation, providing additional protection from oxidative stress [171].

Consequently wild type PINK1 confers neuroprotective properties by reducing overall apoptotic activity and protecting against multiple endogenous and external sources of oxidative damage [172-174]. Correspondingly loss of protein function is associated with oxidative stress and neurodegeneration [172-174].

Over 60 mutations have been observed in this gene, the majority of which are missense or nonsense mutations. Two-thirds cause loss-of-function in the kinase domain, which inhibits the ability of PINK1 to recruit Parkin and induce

mitophagy [105, 175]. Similar effects are observed for mutations in the transmembrane domain [175]. Compromised activity of PINK1 within the autophagy pathway is therefore central in conferring elevated PD risk for the majority of pathogenic mutations.

This results in early-onset, autosomal recessive PD. Clinical phenotype is similar to that observed in Parkin-linked and sporadic PD however with slightly increased risk of psychiatric phenotypes and gait disturbance [176, 177]. Post-mortem studies of PINK1-linked disease are limited but provide evidence of both extensive and absent Lewy body pathology [176, 178].

1.3.5) GBA

The GBA gene encodes glucocerebrosidase (GCase), a lysosomal enzyme that catalyses the breakdown of glucocerebroside within the autophagy pathway. Homozygous GBA mutation causes Gaucher's Disease, a lysosomal storage disorder. Although Parkinsonism affects only a minority of Gaucher's Disease patients it is frequently observed among unaffected relatives. This led to the study of GBA in the context of PD, and heterozygous GBA mutation is now widely accepted as a significant risk factor for PD onset.

Around 300 mutations have been described in this gene within several different haplotypes [179]. These are associated with a 5-fold increase in the risk of autosomal recessive PD and could explain 7% of PD cases [79]. Reduced GCase activity is observed in both GBA-linked and sporadic PD patients compared to controls [180]. This impairs lysosomal protein degradation, enhancing the exosomal release and deposition of alpha-synuclein [181, 182]. Increased alpha-synuclein levels inhibit GCase trafficking, which further reduces lysosomal

function. This results in a positive feedback loop in which alpha-synuclein can rapidly accumulate [80]. Consequently reduced GCase activity is associated with extensive Lewy body pathology and GCase protein is found within the aggregates of both GBA-linked and sporadic PD patients [183, 184].

Disease onset is 1.7-6 years earlier in patients carrying a GBA mutation and is associated with rapid disease progression [185, 186]. Dementia risk is high and hallucinations are frequently reported [81]. Lewy body pathology is prevalent and aggregates contain a higher proportion of GCase than those in sporadic patients [81, 184].

1.3.6) DJ-1 (PARK7)

This gene encodes the protein deglycase DJ-1, a member of the peptidase C56 family. It is ubiquitously expressed and generally present as a dimer in the cytoplasm, nucleus and mitochondria. It participates in a number of processes including transcriptional regulation, antioxidative stress reaction and regulation of chaperones, proteases and mitochondria.

Several of these functions protect against neurodegeneration. Under oxidative conditions DJ-1 prevents the aggregation of alpha-synuclein and is able to self-oxidise to protect the cell against ROS [187-192]. It also sequesters Daxx in the nucleus so that it cannot activate apoptotic pathways [193]. Consequently this protein protects cells from alpha-synuclein aggregation, oxidative stress and cell death.

Pathogenic variants in this gene are relatively rare, representing just 1-2% of early onset cases [194]. Several mutations have been identified which

result in loss of functional protein [195-197]. In neuronal cell models the corresponding deficit in cellular protection results in mitochondrial dysfunction and increased susceptibility to oxidative stress and neurotoxic compounds [198].

Phenotypically patients present similarly to those with mutations in PINK1 and Parkin with early onset, slow progression and development of psychiatric phenotypes [199, 200]. Only one post-mortem study has been performed on the brain of a DJ-1-linked PD patient. This found diffuse Lewy body pathology, however more studies are required before conclusions can be drawn more generally for DJ-1-linked disease [201].

1.3.7) Polygenic risk scoring

Polygenic risk scoring is the process by which disease-associated genetic variants are weighted and summed to provide an additive measure of total disease risk for an individual. It has provided a clinically useful method of stratifying high-risk patients for a number of disorders including diabetes and coeliac disease [202, 203]. A similarly useful scoring system could be invaluable for PD as neuronal loss begins many years before symptoms develop, so identifying high-risk individuals would allow the targeting of neuroprotective treatments before the bulk of this loss occurs.

Several attempts have been made to create a risk score for PD. The one that best discriminated between PD case and control was developed by Nalls *et al*. [204]. Their method combined olfactory function, family history, age, gender and a polygenic risk score calculated from the most highly associated 28 SNPs in the most recent PD GWAS. This achieved a high AUC value of 0.92. However the genetic risk score explained just 13.6% of variance, indicating a small

contribution of this factor to total disease risk. Hall *et al.* created several risk scores derived from up to 33 pathogenic PD variants and family history only [205]. Among these the best predictive model achieved an AUC value of just 0.73, reiterating that additive effects of known PD risk variants only is insufficient to provide meaningful separation between PD case and healthy control.

In general polygenic risk scores improve as the number of SNPs used to calculate them increases. Only one study has examined the use of more lenient P value thresholds in PD. Escott-Price *et al.* created polygenic scores using all SNPs with a P value less than 0.5 [206]. These scores were significantly higher among PD cases than controls. However this was tested using logistic regression, which provided no information on how well separated the two groups were. The sensitivity and specificity of this score for identifying high-risk individuals therefore remains unknown. This study also reported a significant correlation between polygenic score and age of onset using several different P value thresholds. However the correlation R^2 values were less than 0.002 for all models, so the relationship between additive genetic risk and age of onset may not be as clear as the P values suggest. Therefore despite a known genetic element in PD onset polygenic risk scores have thus far failed to provide an effective means of identifying high-risk individuals. This is likely to be due to the genetic component being relatively moderate and the influence of several environmental factors in moderating PD risk.

1.4) Environmental causes and risk factors

In addition to genetic variants a number of environmental factors have been linked to PD risk. Some induce Parkinsonism, such as rotenone and MPTP, and have further aided understanding of the disease process. However some factors impart a dose-dependent reduction in PD risk and the mechanisms behind these are less clear. In addition to complex genetic underpinnings a multitude of opposing environmental factors therefore also modifies PD risk.

1.4.1) 1-methyl-1-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP)

MPTP is a pyridine analogue which is converted to MPP+ by MAO-B in the brain, creating ROS as a by-product [207]. MPP+ then selectively affects nigral dopaminergic neurons by inhibiting the oxidation of NADH, reducing mitochondrial oxidative capacity and promoting additional ROS formation [208, 209]. Oxidative stress is increased from two facets, resulting in extensive damage to cellular components and ultimately causing cell death. Pathologically this inhibition of complex I and accumulation of ROS is very similar to true PD except that Lewy bodies are not formed [210].

1.4.2) Rotenone and paraquat

Rotenone and Paraquat are compounds that have been used as pesticides and herbicides since the 1950s. Regular exposure to either is associated with a 2.5-fold increase in PD risk, although age of onset in Paraquat-linked disease is around 5 years younger [211]. The toxicity of both compounds is mediated through oxidative stress, however via different mechanisms. Rotenone inhibits complex I activity by preventing oxidation of NADH [30, 212]. Paraquat induces NADPH oxidase, directly stimulating the production of superoxide [213]. Both

result in a toxic increase in ROS production that is specific to dopaminergic neurons. Lewy bodies are also formed in PD caused by these compounds [214-216].

1.4.3) Smoking

Smoking cigarettes is associated with a time- and dose-dependent protective effect against developing PD. Ever having smoked conveys a 50% reduction in disease risk, rising to a 70% reduction for recent smokers [217, 218]. Furthermore there is a strong inverse correlation between disease risk and pack-years smoked [217]. It has been proposed that this is caused by inhibitory effects of nicotine on MAO-B, an enzyme that catabolises dopamine [219]. This would reduce disease risk by preventing dopamine degradation. However this has not been confirmed so the mechanism by which smoking reduces PD risk remains unknown.

After diagnosis there is no significant effect of smoking on rate of progression, mood or development of cognitive impairment [220]. Mortality rate is also unaffected [220]. The protective effects of smoking therefore seem confined to disease onset and do not modify disease progression.

1.4.4) Caffeine

Caffeine consumption causes a dose-dependent reduction in PD risk [221, 222]. This is also independently observed for coffee consumption, although this is likely to be attributable to caffeine content rather than other nutrients [221, 222]. When adjusted for age and smoking PD risk may be up to 5 times higher among individuals who do not drink coffee than those who consume more than 28oz per day [221, 223]. This is likely to be due to blockade of the adenosine A_{2A}

receptor, which alters the release of inflammatory cytokines and thereby attenuates nigrostriatal dopaminergic neuron loss [224, 225].

Evidence suggests that caffeine may also improve motor symptoms during disease course. In a randomised, placebo-controlled trial UPDRS motor scores significantly improved after 6 weeks of 200-400mg daily caffeine consumption [226]. Despite this the effects of caffeine on disease progression have not been extensively tested. However caffeine is an A_{2A} receptor antagonist and alternative compounds with this property have been explored as possible therapeutics. Although they demonstrate no consistent effects on motor symptoms, antiparkinsonian response to levodopa is improved and time spent in the "OFF" state is significantly reduced [227-230]. This supports a beneficial effect of A_{2A} receptor antagonists in mediating the response of motor phenotypes to medication. Caffeine therefore has the potential to provide some symptomatic relief in PD patients and would benefit from further study.

1.4.5) Non-steroidal anti-inflammatory drugs (NSAIDs)

Inflammation is widely observed in post-mortem PD brain, so the frequent use of NSAIDs, such as aspirin and ibuprofen, was proposed to protect against neurodegeneration. Correspondingly risk of PD is almost halved among individuals reporting regular use of these compounds [231]. Cell and animal models suggest that this may be due to inhibition of ROS production and destabilization of alpha-synuclein fibrils, although the exact mechanism remains unknown [232-234]. However a more recent study demonstrated that the link between PD risk and NSAIDs was only significant when confounding factors

were unaccounted for [235]. Consequently the link between disease risk and NSAIDs remains unclear.

1.5) Current treatments

1.5.1) Levodopa

Dopamine is unable to cross the blood-brain barrier so cannot be used as a therapeutic agent. However the molecule from which it is manufactured, DOPA, is actively transported across by the LAT-1 transporter [236]. Levodopa is the pharmacological equivalent of this compound and is considered the best PD treatment at present. Neurons in the brain are able to convert it into dopamine which can be stored by the cells until it is required for voluntary movement. Although fewer neurons remain able to complete this process the increased substrate availability means that global supplies of dopamine can largely be sustained at a normal level.

Levodopa is generally well tolerated and provides good control of motor phenotypes for several years. However this treatment is symptomatic only and neurons continue to degenerate. As disease progresses higher dosages of Levodopa are required to compensate for the fewer neurons remaining to synthesize it. After 4-6 years this often results in problems with motor fluctuations, dyskinesias and other side-effects [237].

A large proportion of the Levodopa dose is degraded by peripheral mechanisms before it reaches the brain. The COMT enzyme metabolises Levodopa to 3-0-methyldopa, a form unusable by neurons. Dopa decarboxylases

convert Levodopa into dopamine in the off-target peripheral nervous system, causing nausea and vomiting. Consequently Levodopa is often administered in combination with COMT inhibitors or Dopa Decarboxylase Inhibitors (DDCIs). These compounds prevent the breakdown of Levodopa before it reaches the brain, prolonging its peripheral half-life. This means that total dose can be reduced, minimising side-effects and motor fluctuations. However even in combination with COMT inhibitors or DDCIs long-term Levodopa use is associated with a number of undesirable side-effects and complications.

1.5.2) MAO-B inhibitors

Once dopamine has crossed the synapse and transmitted the nerve signal most molecules are repackaged into vesicles or transported back to the presynaptic terminal. Any remaining in the synaptic cleft are then degraded by the monoamine oxidase (MAO) enzymes. MAO-B inhibitors bind to the MAO-B enzyme preventing dopamine degradation. This increases the pool of available dopamine, reducing the demand for synthesis of new molecules and increasing the likelihood of neurotransmitter binding to the post-synaptic membrane. MAO-B inhibitors can be used in isolation as a treatment in early disease, but are more commonly used to supplement levodopa treatment as disease progresses.

1.5.3) Dopamine agonists

Dopamine agonists such as Apomorphine are compounds that mimic the effect of dopamine by binding to dopamine receptors, stimulating the same downstream response. Although not as effective as Levodopa for treating motor dysfunction they are associated with fewer motor fluctuations and also improve non-motor symptoms. They are often used to treat early-onset patients to delay

Levodopa treatment and minimise the motor fluctuations and dyskinesias associated with long-term use. In advanced disease Levodopa and dopamine agonists can be used in combination to reduce side-effects over treatment with Levodopa alone.

1.5.4) Anticholinergics

Acetylcholine is a neurotransmitter that works in partnership with dopamine to produce smooth muscle control. When dopamine levels are reduced in PD the amount of acetylcholine remains the same, resulting in altered proportions of the two neurotransmitters. This causes loss of control over fine voluntary movements which manifests in tremor phenotypes. Anticholinergics inhibit acetylcholine, restoring the balance of neurotransmitters and reducing tremor. However they treat fewer symptoms than other medication, and this coupled with undesirable side effects means that they are no longer commonly used.

1.5.5) Deep Brain Stimulation (DBS)

DBS is a surgical procedure whereby electrodes are implanted into the subthalamic nucleus, thalamus or the globus pallidus. These deliver high frequency electrical stimulation to the surrounding regions, blocking the native abnormal electrical signals that otherwise cause tremor, dyskinesia and other motor problems. It is generally only performed in more severe cases as although reversible it is an invasive procedure with associated surgical risk. Generally it is also most effective in these individuals. The degree of improvement varies significantly among patients however most continue to rely on Levodopa or other medication. Nevertheless dosage can usually be considerably reduced.

Accordingly fewer dyskinesias and other side effects are experienced and quality of life is drastically improved in the majority of patients.

1.5.6) Cell transplantation

Foetal ventral mesencephalic tissue is enriched in dopaminergic neuroblasts and has therefore been used to create grafts that replenish the dopaminergic cells lost in PD. This treatment is still in experimental stages and consequently grafts are unstandardized and highly variable. Despite this the majority of studies have reported increased fluorodopa uptake in the putamen that indicates some restoration of native function [238-241].

Responses to this treatment are mostly positive but highly variable. The majority of patients experience improvements in motor symptoms that are maintained up to 18 years post-surgery [238, 240-242]. Levodopa dose can be reduced in most cases and in some can be stopped altogether [240-242]. However a minority of cases have reported worsening of symptoms [239, 240, 242]. Furthermore graft-induced dyskinesias are a relatively common side effect, which although minor in most patients can occasionally constitute severe problems [239, 243]. In general improvements are less drastic in double-blind sham-controlled studies than open-label trials, perhaps indicating a placebo effect associated with this treatment [239]. Additionally post-mortem studies have shown that some grafted tissues also develop Lewy bodies, indicating that the grafts will also eventually succumb to PD pathology [244-246].

Foetal grafts may therefore provide long-term relief of motor symptoms by repopulating the striatum with dopaminergic neurons. This can provide a degree of symptomatic relief unattainable with medication alone. However high

variability and the difficulty in obtaining foetal cells makes this treatment not viable on a large scale at present. It is hoped that developments in stem cell technology may in future overcome these limitations.

1.6) Phenotypic aspects

PD is characterised by a high degree of heterogeneity. At onset patients can present with any combination of motor, neuropsychiatric and autonomic phenotypes of differing severity. Disease progression is also highly variable in rate of decline and the development of additional symptoms such as dementia. Many studies have sought to quantify this by determining PD subtypes. These define groups of patients by common sets of co-occurring phenotypes in an attempt to further the understanding of the disease process.

The first study to explore phenotypic subgroups was performed by Jankovic *et al.* in 1990 [247]. Their analysis of 800 patients provided the tremor-dominant (TD) and postural-instability gait disorder (PIGD) phenotypes still used widely today. Subgroup can be inferred from the UPDRS questionnaire and is therefore easily obtainable.

Since that time a number of further studies have been carried out to explore phenotypic heterogeneity in PD. Generally around four phenotypic groups are identified, several of which are common between multiple different analyses. Tremor-dominant disease is identified as a subtype by almost every study [247-251]. Patients in this group experience severe tremor but otherwise relatively benign symptoms with low risk of cognitive decline. Between 12 and

21% of the total PD population are assigned to this subtype [248, 249]. Another group of younger-onset patients is also frequently identified [249, 251, 252]. These individuals begin displaying symptoms in their 50s and experience slow disease progression with mild motor and non-motor symptoms.

The remaining subsets are more variable but still largely comparable. A subgroup characterised by severe motor and non-motor symptoms is frequently identified, and often an opposing subset of individuals with mild symptoms also. Many studies have identified a "non-tremor" phenotypic group defined by severe non-tremor motor symptoms, freezing of gait and moderate cognitive decline. Other phenotypes associated with this group vary but have included depression, daytime sleepiness, night-time sleep disturbance, hypokinesia and levodopa complications. Additional phenotypic groups have been identified in isolated studies, but are not widely replicated and are therefore unlikely to reflect disease stratification in the general population.

Although some variability remains the core definitions of several phenotypic subgroups therefore seem to be consistent throughout most studies. This provides strong evidence for their biological relevance and universal application to the whole PD population. The possibility of diverse molecular aetiologies corresponding to particular phenotypic subgroups should therefore be considered.

The majority of these studies used some form of variable selection to create a distance matrix between individuals, followed by k-means clustering. Although the resulting phenotypic groups provide a convenient system of partitioning the PD population, the method has a number of shortfalls. Variable

selection methods quantify how much variance each phenotype explains, however there is no robust method of defining a threshold for this measure above which a phenotype contributes to the distance matrix. Consequently the definition of which phenotypes are important and which are irrelevant can be somewhat arbitrary.

K-means clustering requires the number of phenotypic groups to be specified before the groups themselves are defined, which can bias results toward preconceived ideas. Lewis *et al.* and Lin-Yang *et al.* explored how specifying different numbers of subgroups within this method affected the content of the phenotypic groups [250, 251]. Their findings show that increasing the number of groups generally results in the splitting of one existing group into two. This indicates that these methods are robust as groups maintain their overall structure but are split on finer details. However there is no established best method for defining the optimal number of clusters, so the most relevant solution remains unknown.

Within the phenotypic space defined by the distance matrix, K-means and many other clustering methods can only identify spherical clusters. As spheres do not tessellate this results in areas of phenotypic space, and consequently some patients, not assigned to any cluster. This indicates that the clusters may not fully recapitulate the underlying biology and results in some patients being lost from downstream analyses.

The use of discrete groups also limits the comparison of phenotypes to binary tests of absent and present or mild and severe, and consequently does not reflect the continual nature of phenotypic variability. Few studies have

investigated links between continuous phenotypic measurements and those that have were limited to hypothesis-led tests. Analysis of continuous traits has greater statistical power, so methods such as Principle Component Analysis (PCA) that identify representative quantitative variables could be a valuable avenue for discovery in PD.

The development of similar phenotypes among patients with pathogenic mutations indicates a genetic component influencing disease progression. Despite this analysis of genotype-phenotype links in PD has been limited. Few studies have investigated these effects and those that have were confined to candidate genes, overlooking the potentially large effect of millions of other variants. Links between genotype and phenotype could be invaluable for understanding diverse molecular mechanisms and for developing and targeting more specific symptomatic treatments. Consequently there is a need for hypothesis-free, genome-wide genetic analysis of phenotypic variation to further the understanding of the disease process.

1.7) Thesis structure

The genetic underpinnings of disease can provide valuable insight into molecular method of onset. From this understanding novel treatments can be developed. This study employed network- and pathway-based approaches in the analysis of SNP and CNV data to identify altered cellular processes on which pathogenic variants converge in PD onset. Phenotypic patterns within PD were quantified using both discrete patient groups and continuous measures of

severity. Patient genotypes were interrogated to elucidate genetic variants influencing disease course.

Chapter 2: Gene network analysis

2.1) Introduction

There now exists a wealth of data sources that describe the role and function of genes and their products, such as protein-protein interactions, co-expression analyses and functional annotations. However each data type measures only a few characteristics of a gene and for only a fraction of the genes in the genome. In this analysis multiple data sources were integrated in order to increase coverage and fully represent gene features in a disease-specific phenotypic linkage network (PLN), documenting the interactions and functional similarities of several thousand genes.

This network was then interrogated using expression quantitative trait loci (eQTL) information. EQTLs are individual SNPs that influence the expression of genes either physically close (cis-eQTLs) or distant (trans-eQTLs) to their genomic location. Gene pairs linked both by regulatory variant and in the PLN were identified, and from this four PD-relevant eQTLs were elucidated. Two were cis-eQTLs of LRRK2. The genotype resulting in the greatest additive increase in LRRK2 expression was significantly over-represented among PD cases. This implicates increased LRRK2 gene expression in PD onset as a product of novel additive SNP variation.

2.2) Methods

2.2.1) Creating a gene network

This work builds on the method created by Frank Honti during his DPhil, and extensions to this method suggested by Julia Steinberg. The procedure developed by Honti *et al.* was modified to create a PD-specific network [253]. The first step was to define a "gold standard" measure of relatedness between gene pairs by which the accuracy of other data sources could be evaluated. Semantic similarity (SS) is a measure of distance based on meaning or content, and was used here to describe the degree of similarity of phenotype annotations between gene pairs. Phenotype annotations were taken from the Human Phenotype Ontology and Mouse Genome Informatics databases [254, 255]. Across the genome three times more mouse genes are annotated with phenotypes than are human genes, and consequently coverage of the semantic similarity measure is much greater. Values of semantic similarity are highly correlated between mouse and human phenotypes [253]. Mouse phenotypes therefore provide greater coverage whilst accurately representing human gene function and were therefore well suited to providing the required benchmark.

To create a PD-specific network only those phenotypes relevant to PD were used as defined by clinicians (see Tables 2.1 a and b for full lists). For these particular phenotypes there existed over twice the number of genes annotated with human phenotypes than mouse. To increase coverage a combined scoring system was developed that incorporated the similarity of both mouse and human PD-relevant phenotype annotations using the following equation:

$$\text{Benchmark score} = \begin{cases} \dfrac{(\text{Mouse SS} + \text{Human SS})}{2} & \text{If both measures exist} \\ \text{Semantic similarity score} & \text{If one measure exists} \end{cases}$$

Where all semantic similarities were normalised to a 0-1 scale

For each dataset gene pairs were extracted and ordered according to their benchmark score. They were then assigned to bins of 100 pairs, for which the mean benchmark and dataset values were calculated. These were then compared across all bins and the strength of correlation between them used to evaluate the usefulness of the dataset.

Regression up to a third degree polynomial was performed on all scores above the median value of the benchmark (see Figure 2.2 for example). This threshold eliminated noise from observations between unrelated genes. The dataset value for each gene pair was then individually rescaled according to this regression up to a maximum limit, at which point a maximal value was assigned.

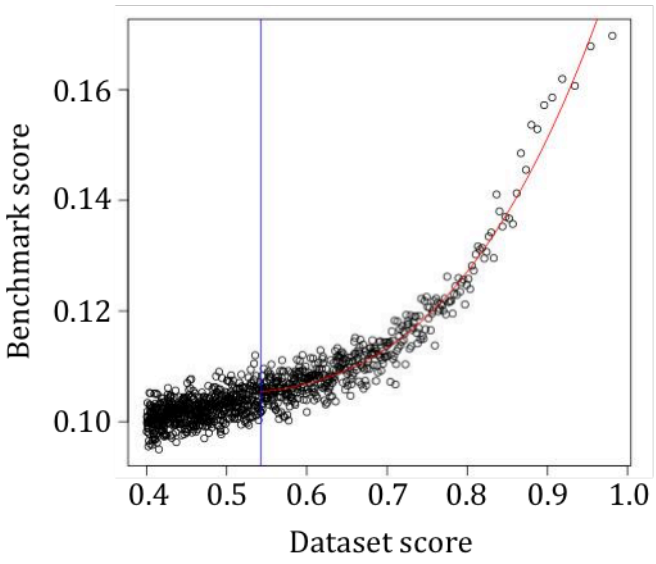


Figure 2.2: Rescoring values for each dataset against the benchmark. For all values above the median (blue line), regression analysis was performed (red line), to normalise dataset values according to the benchmark.

Table 2.1a: list of mouse PD-relevant phenotypes used to generate the semantic similarity benchmark

Mouse phenotype	MP number
Abnormal autophagy	MP:0008260
Abnormal intestinal transit time	MP:0006001
Abnormal mitochondrial ATP synthesis coupled electron transport	MP:0010956
Abnormal neurotransmitter uptake	MP:0003444
Abnormal synaptic dopamine release	MP:0010149
Abnormal synaptic vesicle clustering	MP:0004793
Abnormal synaptic vesicle recycling	MP:0004770
Abnormal vesicle-mediated transport	MP:0008546
Akinesia	MP:0002690
Alpha-synuclein inclusion body	MP:0008493
Amyloid beta deposits	MP:0003329
Anhedonia	MP:0009710
Anosmia	MP:0004512
Ataxia	MP:0001393
Behavioral despair	MP:0002573
Bradykinesia	MP:0005156
Decreased dopamine level	MP:0005643
Decreased dopaminergic neuron number	MP:0011448
Decreased vertical activity	MP:0002757
Dystonia	MP:0005323
Hunched posture	MP:0001505
Hypoactivity	MP:0001402
Impaired balance	MP:0001525
Impaired coordination	MP:0001405
Impaired limb coordination	MP:0001524
Impaired olfaction	MP:0008544
Impaired righting response	MP:0001523
Increased dopamine level	MP:0001906
Induced hyperactivity	MP:0008911
Loss of dopaminergic neurons	MP:0003244
Neuronal cytoplasmic inclusions	MP:0011975
Oxidative stress	MP:0003674
Short stride length	MP:0001407

Table 2.1b: list of human PD-relevant phenotypes used to generate the semantic similarity benchmark

Human Phenotype	HP number
Abnormality of extrapyramidal motor function	HP:0000006
Adult onset	HP:0002172
Aggressive behavior	HP:0002172, HP:0002312
Akinesia	HP:0004409
Anarthria	HP:0000725
Anosmia	HP:0002063
Anxiety	HP:0002172
Apathy	HP:0001300
Autosomal dominant inheritance	HP:0000738, HP:0002548
Blepharospasm	HP:0002396
Bradykinesia	HP:0000006, HP:0000751, HP:0001288, HP:0002071
Clumsiness	HP:0002172
Cogwheel rigidity	HP:0003812
Constipation	HP:0002063
Dementia	HP:0001300, HP:0002322, HP:0007311
Depression	HP:0000739, HP:0001260, HP:0002172
Diffuse brain atrophy	HP:0006892
Dysarthria	HP:0001332, HP:0002360, HP:0002529
Dysautonomia	HP:0001278, HP:0002322
Dysphagia	HP:0001332
Dystonia	HP:0001257, HP:0002063, HP:0002548, HP:0100315
Frontotemporal cerebral atrophy	HP:0003236, HP:0100315
Gait disturbance	HP:0003581
Hallucinations	HP:0000741, HP:0001300, HP:0001824, HP:0011960
Hypokinesia	HP:0002304, HP:0002375
Hyposmia	HP:0000738
Hypotension	HP:0002067
Insidious onset	HP:0001300, HP:0003596
Insomnia	HP:0001337
Lewy bodies	HP:0002360, HP:0003587
Mask-like facies	HP:0001621, HP:0002425
Mental deterioration	HP:0002283
Middle age onset	HP:0002459
Neuronal loss in central nervous system	HP:0002015
Orthostatic hypotension	HP:0000738
Paranoia	HP:0000726
Parkinsonism	HP:0000458, HP:0000716, HP:0001268, HP:0002063,

	HP:0002459, HP:0002615, HP:0003812
Parkinsonism with favorable response to dopaminergic medication	HP:0000298, HP:0002067, HP:0002172
Personality changes	HP:0000718, HP:0000738
Phenotypic variability	HP:0001337, HP:0002548
Postural instability	HP:0000012, HP:0001260, HP:0001260, HP:0001332, HP:0002063, HP:0011960
Postural tremor	HP:0000725, HP:0001300
Progressive disorder	HP:0000716
Psychotic episodes	HP:0000643, HP:0000726
Resting tremor	HP:0002019, HP:0002067
Rigidity	HP:0000718, HP:0001300, HP:0001337, HP:0003676, HP:0100785
Short stepped shuffling gait	HP:0001337, HP:0003587
Sleep disturbance	HP:0000298
Substantia nigra gliosis	HP:0000012, HP:0001332
Tremor	HP:0000716, HP:0000751, HP:0002067, HP:0002375
Urinary urgency	HP:0001300, HP:0002174
Weak voice	HP:0000726, HP:0007311
Weight loss	HP:0011999

For each unique gene pair the final integrated score (WS) was calculated by ranking rescaled dataset scores in order of their value, then scaling them accordingly so that the most informative scores contributed most to the final total. These were then summed to provide an overall measure of gene functional similarity. This method of weighting was proposed by Lee *et al.* [256] and is summarised as:

$$WS = L_0 + \sum_{i=1}^{n} \frac{L_i}{D \times i}$$

Where L_0 = *largest link weight*

L_i = *remaining link weights indexed by size*

i = *index of the link*

D = *free parameter*

Where the free parameter, D, is optimised to provide the best correlation with the original benchmark. The resulting PD-specific PLN shows increased accuracy and coverage above that conferred by any individual data type [253].

2.2.2) Testing the network

To test the functionality of the network the strength of links between genes previously implicated in Genome Wide Association Studies (GWAS) was tested (downloaded from www.PDgene.org; December 2013). This was carried

out empirically by permuting random subsets of genes to create a null distribution. A number of confounding factors were accounted for.

Longer genes harbour a greater quantity of random mutations [253]. This can cause a high degree of linkage within a PLN, resulting in a systematic correlation between gene length and the number of links to other genes. The significance of linkage between sets containing longer genes could therefore be inflated if length is not adequately accounted for. Consequently, for each test gene permutations were carried out selecting only from those genes most similar in length.

Several of the input datasets, and therefore also the final network, were affected by study bias. Those genes that are interesting to researchers, often those implicated in disease, are more highly studied and were consequently more linked as a product of their reputation rather than their function. This was more prevalent for annotation-based datasets than those based on untargeted, high-throughput assays. Again this could inflate P values especially in relation to the study of the disease [257]. This was accounted for by matching each test gene to those with a similar degree of connectivity in addition to similar length, then permuting only from that pool.

Some noise was still likely to remain in the PLN despite efforts to minimise it. For this reason only the most reliable links were used during analysis. The strongest 500,000 links were used for testing in the general network, and this was scaled accordingly for the smaller PD-specific and non-PD networks. This filtered out most of those that were misleading or irrelevant.

In order to demonstrate that making a network PD-specific improved performance, two further networks were constructed for comparison. Using the same methodology a general network was created that included genes annotated with any phenotype. A "non-PD" network was also produced that was comprised of only those genes for which none of their annotations fell within the previously defined PD-relevant set. The strength of connectivity between known PD risk genes was explored for all three networks.

2.2.3) EQTL analysis using the network

EQTLs were identified by Viola Volpato. 11,555,102 genotyped and imputed variants were downloaded from the GTEx dataset (release version 6), in addition to RPKM values for 449 healthy individuals and 17 tissues. GTEx data was not included in the network. For each tissue type protein coding genes were selected and retained if greater than 0.1 RPKM was observed in at least 10 individuals. Quantile normalisation was then performed.

eQTL analysis was performed using the Matrix eQTL R package. Linear regression was carried out on gene-SNP pairs, using three genotyping principle components (provided by GTEx), gender and genotyping array as covariates. P-values were calculated using the t-statistic. False discovery rate was estimated using the Benjamini-Hochberg procedure.

Cis-eQTLs are located within 1mb of the transcription start sight of the gene whose expression they affect, and could therefore be reliably identified. Trans-eQTLs were more difficult to detect as they can lie anywhere in the genome. Complex patterns of genome-wide interactions, in addition to the need to account for millions of tests, meant that reliable identification of these

variants was low in statistical power. To maximise the probability of only including genuine eQTLs, and to enrich for SNPs involved in interacting pathways, only those with an FDR-corrected P value of less than 0.2 for at least one cis-gene and at least one trans-gene were included in the analysis.

Next, cis- and trans- gene pairs were identified that were both under the control of the same eQTL and highly linked in the PD-specific network. In defining linkage both direct links and indirect links via one or more intermediate genes were considered. To investigate indirect links the R package 'tnet' was used to calculate the shortest path between each gene pair [258]. Significance of the link was evaluated empirically. The trans-gene was randomly permuted, and the shortest distance between that and the cis-gene formed a null distribution. Gene pairs with P value less than 0.1 were defined as significantly linked.

PD-associated SNPs were identified from the most recent meta-analysis that compared 13,000 cases and 95,000 controls [100]. Two sets of variants were defined: genome-wide significant and those with association P value less than 0.001. GWAS intervals were defined containing all variants in high linkage disequilibrium (D'>0.8) with the most strongly associated SNP. This ensured the identification of all loci the lead SNP could be tagging. EQTLs that were located within these regions and regulated a gene pair that was highly linked in the network were investigated further.

2.2.4) Modelling allelic co-inheritance

Co-inheritance of eQTL alleles was examined in two cohorts. The first used was the OPDC Discovery cohort described in detail in Chapter 3. This provided 843 cases and 270 controls. The second cohort consisted of publically

available data for 642 PD case individuals from the Autopsy-Confirmed Parkinson Disease GWAS Consortium (APDGC) (dbGaP accession phs000394.v1.p1). Participants were required to have an antemortem clinical diagnosis of Parkinsonism without prominent dementia. Postmortem examination must have confirmed the presence of Lewy bodies and moderate-to-severe substantia nigra neuronal loss with minimal Alzheimer-like pathology. Genotype data was generated using the Illumina HumanOmni1 Quad v1 SNP array.

PLINK was used to carry out data quality control for the APDGC cohort identically to that carried out for the OPDC cohort, which is described in detail in Chapter 3 [259]. In brief, individuals were excluded if missing data rate was high or genotypic and phenotypic sex was discordant. SNPs were excluded based on criteria for missing data, minor allele frequency and Hardy-Weinberg equilibrium. Finally individuals of divergent ancestry were excluded using EIGENSTRAT [260]. This ensured that only high quality data and genetically comparable individuals were used in analysis.

VCFtools was used to format data for upload to the Michigan imputation server [261, 262]. This performed both phasing and imputation using Eagle and Minimac3 respectively, employing the 1000Genomes Project (phase 3, release 5) as the reference panel [262-264]. These methods are discussed in detail in Chapter 3. The returned dataset consisted of several million additional variants, whose genotypes were inferred for each individual by combining observed data with linkage disequilibrium (LD) patterns calculated from the reference dataset.

None of the eQTLs significantly diverged from Hardy-Weinberg equilibrium and none were in LD with each other. They were therefore modelled as independent variables. Co-inheritance probabilities were calculated using binomial distributions that conditioned on the allele frequencies of each variant individually. Family members were excluded from this analysis in order to prevent bias towards inherited genotypes.

2.3) Results

2.3.1) Network testing

By combining human and mouse semantic similarity, 1,203,499 gene pairs were assigned a benchmark score, almost three times more than using mouse phenotypes in isolation. The final disease-specific network consisted of 130,626 links between 2055 genes. A list of datasets included in the final network can be found in Table 2.3.

Utility of the PD-specific network was explored by examining how closely

Table 2.3: six datasets provided useful information as measured against the semantic similarity benchmark and were included in the final PD-specific network

Dataset	Data type	Reference
Gene ontology - biological process	Functional annotation	[265]
Gene ontology - cellular component	Functional annotation	[265]
Gene ontology - molecular function	Functional annotation	[265]
GNF2	Co-expression	[266]
InterPro	Protein families, domains and active sites	[267]
String	Protein-protein interactions	[268]

linked genes implicated in PD GWAS were relative to randomly permutated gene sets. PD-specific, non-PD and general networks were tested to provide comparison. Only the PD-specific network demonstrated clustering of these genes that was significantly greater than would be expected by chance (Figure 2.4). No significant linkage was observed between them in the general network, despite including the same information. This shows that in the PD-specific network many of the irrelevant links were removed and the relative strength of PD-relevant links was increased. Furthermore there was no significant clustering of these genes in the non-PD network, demonstrating that it was the particular choice of phenotypes, rather than interrogating a smaller network generally, that resulted in the greater specificity (Figure 2.4).

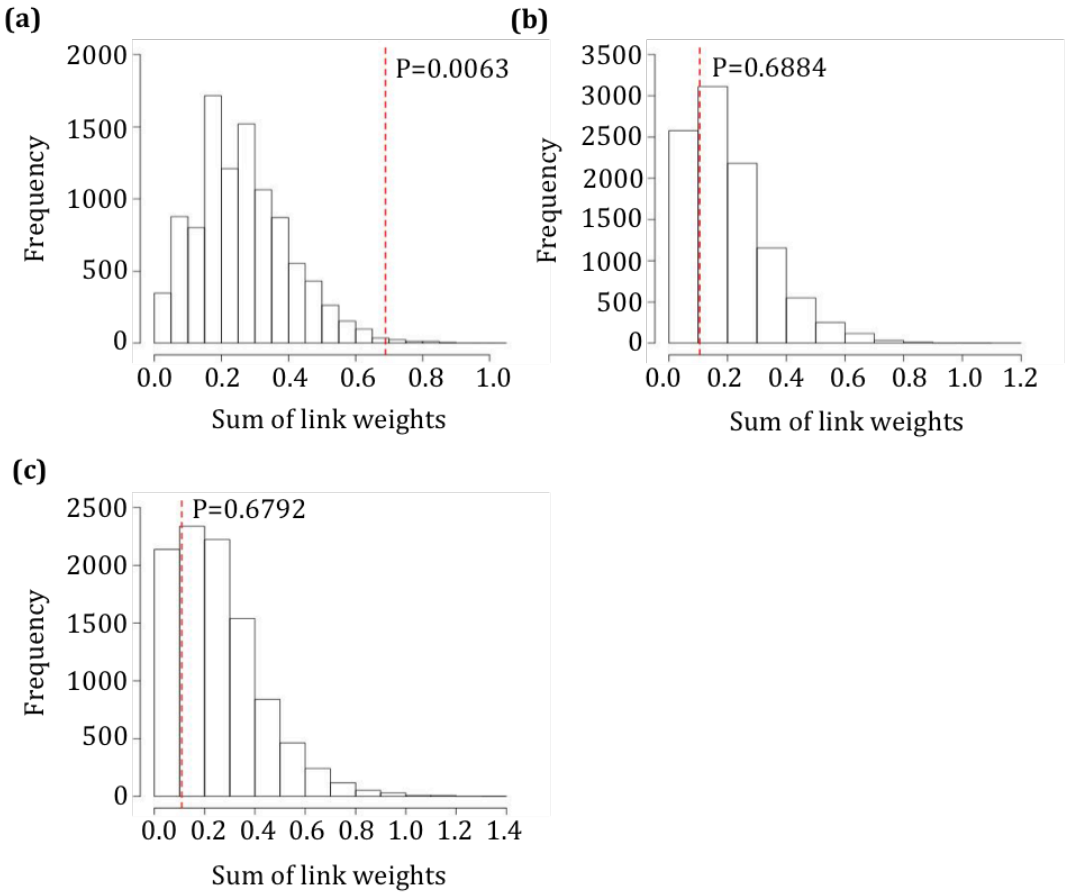
2.3.2) EQTL analysis

The PD-specific network was interrogated to determine disease-relevant, highly linked gene pairs. EQTLs influencing the expression of both members of the connected pair, and located in linkage with a GWAS interval, were identified as potential modifiers of disease risk. In this way four PD-relevant eQTLs were identified across three tissue types (Table 2.5).

Table 2.5: tissue type, genes and effect sizes for the four PD-relevant eQTLs identified

SNP	**Tissue**	**Cis-gene**		**Trans-gene**	
		Gene name	Effect size	Gene name	Effect size
rs34179446	Skeletal muscle	PDE10A	-0.723216	NPY2R	-0.208838
rs12095503	Skeletal muscle	FMO3	-0.576312	ARSA	-0.450782
rs7303059	Subcutaneous adipose	LRRK2	-0.285162	CISD1	-0.522734
rs1472118	Pancreas	LRRK2	0.348326	PPARGC1A	-0.629243

Figure 2.4: Strength of clustering between genes implicated in PD GWAS compared to random gene sets in different networks. The frequency histogram represents connectivity between 10,000 permuted gene sets and the red line shows connectivity between PD risk genes. The PD-specific PLN (a) showed significantly stronger links between PD risk genes than expected by chance (P=0.0063). This was not observed in either the non-PD PLN (b) or the general PLN (c) (P=0.688 and P=0.679 respectively).



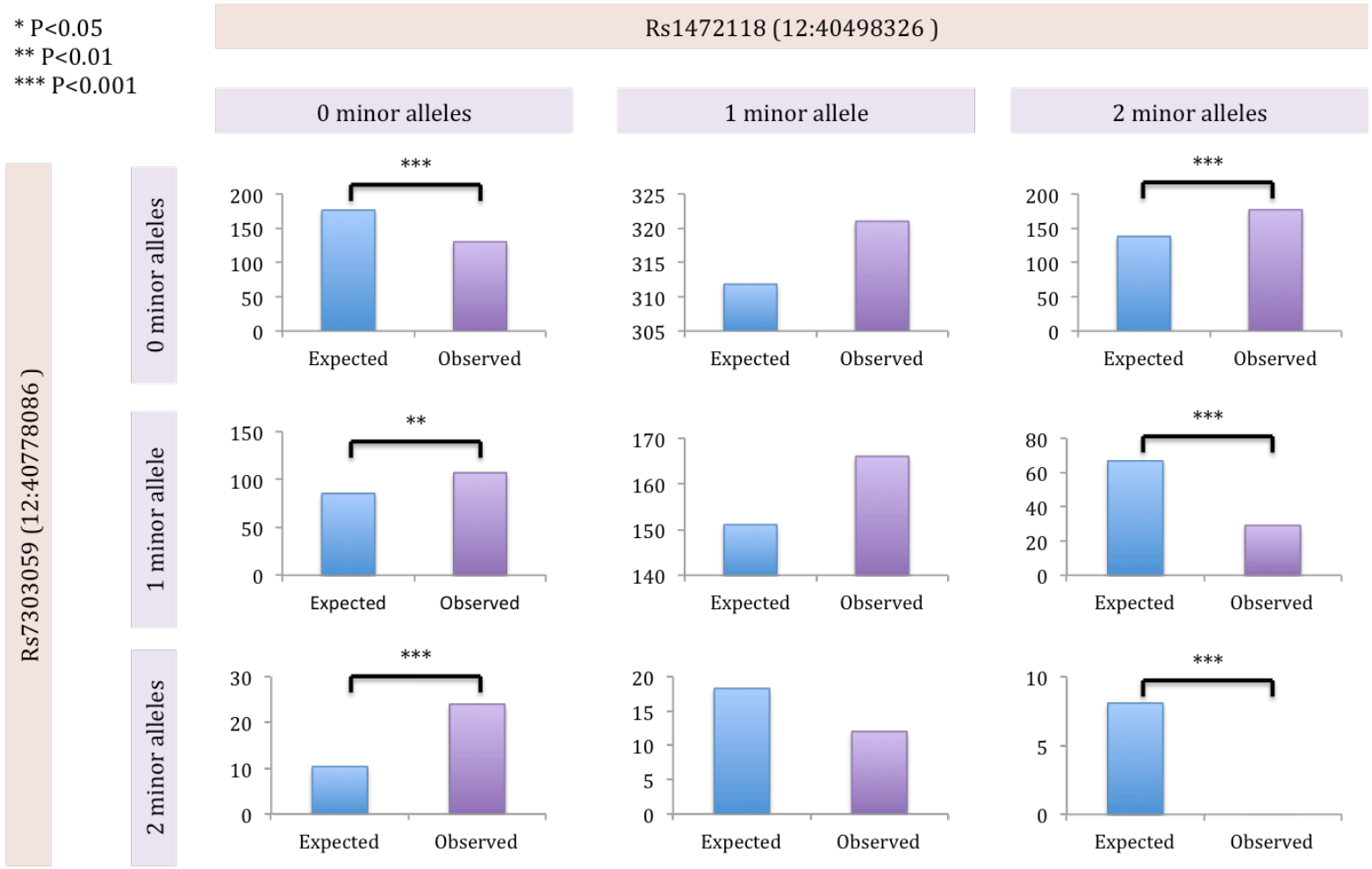
Rs34179446 and rs12095503 were both identified in skeletal muscle. Rs34179446 affects the expression of PDE10A in cis and NPY2R in trans. The minor allele of this variant is associated with a reduction in expression of both genes, which is greatest for PDE10A. Rs12095503 influences FMO3 in cis and ASRA in trans. This minor allele causes a decrease in expression of both genes of similar magnitude.

The two remaining variants were both cis-eQTLs of LRRK2, a gene implicated in autosomal dominant PD. Neither variant was in LD with any known pathogenic mutations and each corresponded to a different GWAS interval. One was in linkage with an associated SNP lying just before the LRRK2 transcription start site, whereas the other was in linkage with a SNP in MUC19, the gene proceeding LRRK2. The first, rs1472118, affected PPARGC1A expression in trans and was identified in pancreatic tissues. Rs7303059 was identified in adipose tissue and affected the expression of CISD1 in trans.

Co-inheritance of the LRRK2 cis-eQTLs was then investigated to explore the possibility of additive effects. The OPDC cohort was used as an exploratory dataset and the APDGC cohort was used as a replication dataset. The expected frequency of each allele combination was modelled using binomial distributions. This conditioned on minor allele frequencies in PD case individuals to account for possible bias owing to their proximity to GWAS regions.

Figure 2.6 shows that a number of allelic combinations were under- and over-represented among PD cases in the OPDC cohort. Significant associations were observed with both homozygous states of the rs1472118 variant for all rs7303059 genotypes. However there was no consistent relationship between

Figure 2.6: co-inheritance patterns between the two LRRK2 cis-eQTLs shown by the expected and observed number of individuals with each genotype in the OPDC cohort. All six genotypes associated with the homozygous states of rs1472118 demonstrated significantly different frequency to that expected under a null distribution. However the direction of enrichment was inconsistent.



SNP genotype and whether an increase or decrease over the expected number of individuals was observed.

This was also seen in the APDGC cohort. Figure 2.7 shows that both homozygous states of rs1472118 again demonstrated deviation from the expected population frequencies. The direction of enrichment was the same as that observed in the OPDC cohort for all associated genotypes. This was statistically significant for all except the homozygous minor allele state of rs7303059.

Using effect size values extracted from the eQTL analysis, the additive effect on LRRK2 expression of each allelic combination was predicted. This change was plotted against P value and fold enrichment in Figures 2.8 and 2.9 respectively. The distribution of control genotypes was mostly as expected under a null hypothesis. However one genotype was significantly enriched which represented the largest decrease in LRRK2 expression. This was also observed in PD cases indicating that this allelic combination was more prevalent in the general population than expected by chance.

Three out of five genotypes that caused an increase in LRRK2 expression showed significantly different frequency in PD cases to that expected under a null hypothesis. Those that caused a change in expression above 0.4 were most significant. However the direction of enrichment was inconsistent. Genotypes that caused an additive increase up to 0.411 were under-represented, whereas those that caused an increase of 0.7 were 1.28-fold over-represented compared to expectation. The relationship between genotype frequency and increased LRRK2 expression was therefore not straightforward.

Figure 2.7: co-inheritance patterns of the LRRK2 cis-eQTLs in the APDGC cohort shown by expected and observed genotype frequencies. This replicated the association of three genotypes significantly linked to PD onset in the OPDC cohort.

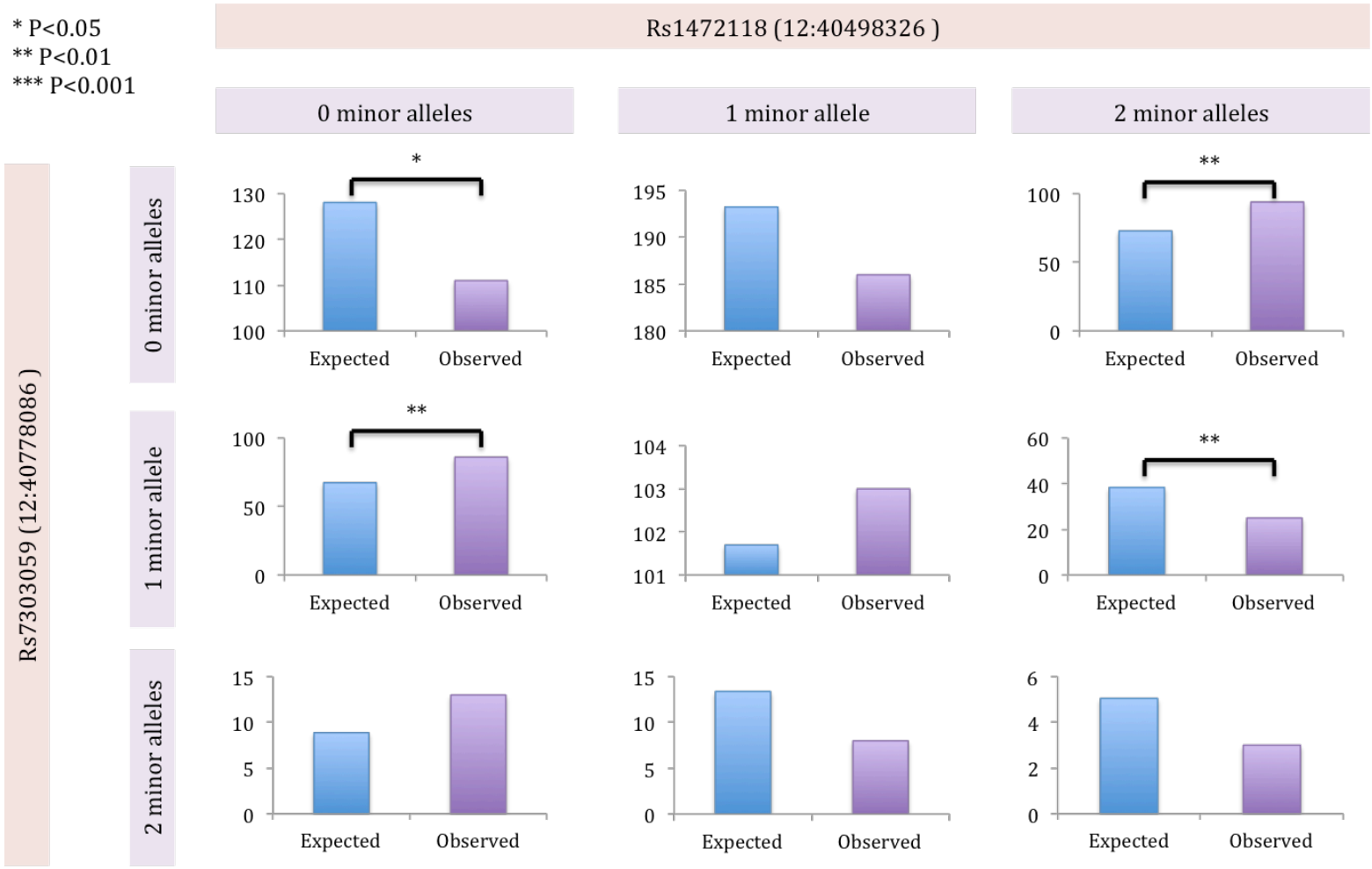


Figure 2.8: probability associated with the observed population frequency of each genotype compared to its additive effect on LRRK2 gene expression in the OPDC cohort. The black line indicates significance after Bonferroni correction. Genotypes that caused the greatest increases in LRRK2 expression were significantly associated with PD cases. The genotype that caused the greatest decrease was significantly associated with both PD case and control.

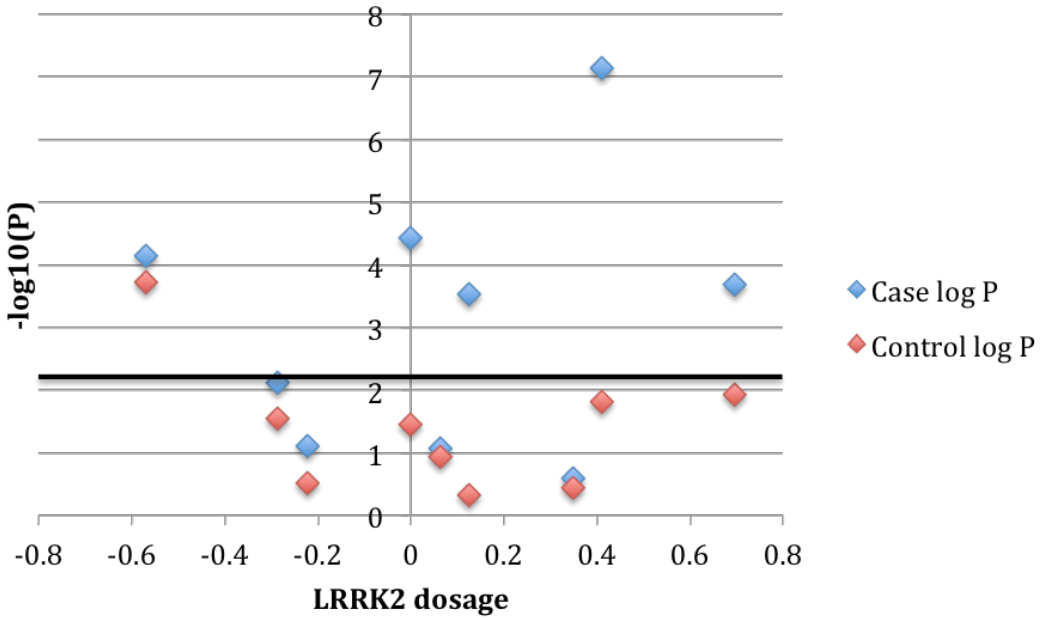
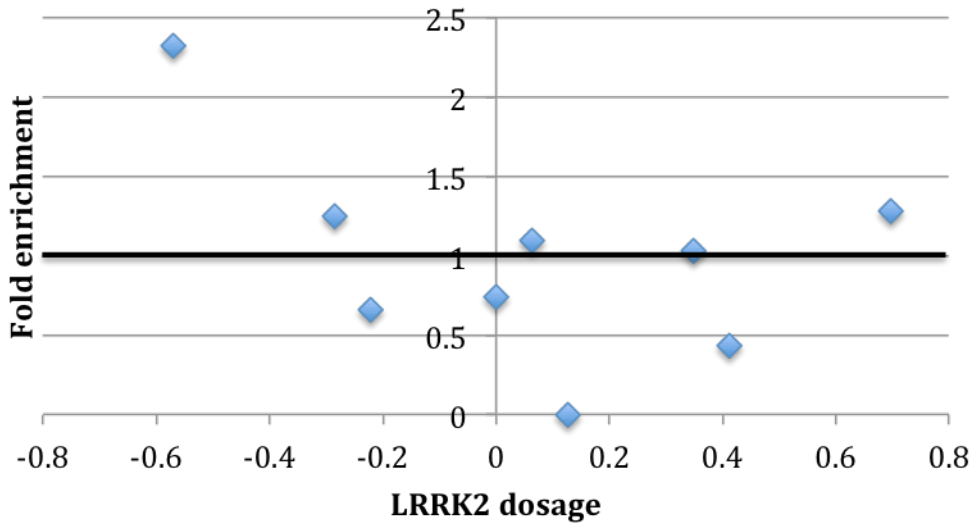


Figure 2.9: fold enrichment of each genotype in OPDC PD case individuals compared to additive effect on LRRK2 gene expression. Genotypes associated with the greatest increase and decrease in gene expression were significantly over-represented whereas those associated with intermediate expression levels were under-represented.



The double homozygous major genotype was significantly under-represented among PD cases by 26.3%. This genotype reflected the baseline level of LRRK2 expression attributable to these loci. Overall the OPDC PD case population appeared to contain an excess of individuals with genotypes causing the largest changes of LRRK2 expression in either direction, and a depletion of those with genotypes causing little or no change.

Analysis of the APDGC dataset is summarised in Figures 2.10 and 2.11 and replicated several results. Genotypes that conferred the greatest increases in LRRK2 expression again showed significantly different frequency to expectation. Direction of enrichment remained the same as in the OPDC cohort for all significant results. The genotype that conferred the greatest change was 1.3-fold over-represented among this case population, and that which conferred the 2nd greatest change was under-represented 0.652-fold. The double homozygous major genotype was also significantly under-represented in this sample.

The genotype that resulted in the greatest decrease of LRRK2 expression was over-represented in this cohort, but not significantly so. However significant over-representation was observed of the genotype causing the second greatest decrease of -0.285. There is therefore some evidence to suggest an excess of genotypes causing a reduction of LRRK2 expression in the general population, however these results were not conclusive.

Among PD cases in both cohorts there was a general trend towards under-representation of genotypes that caused small changes in LRRK2 expression, and over-representation of genotypes that caused large increases or decreases. In controls genotypes associated with highly decreased but not

Figure 2.10: probability associated with the frequency of each genotype in the APDGC cohort compared to additive effect on LRRK2 gene expression. The two genotypes that caused the greatest increase in expression showed significantly different frequency to that expected under a null distribution, replicating results from the OPDC cohort.

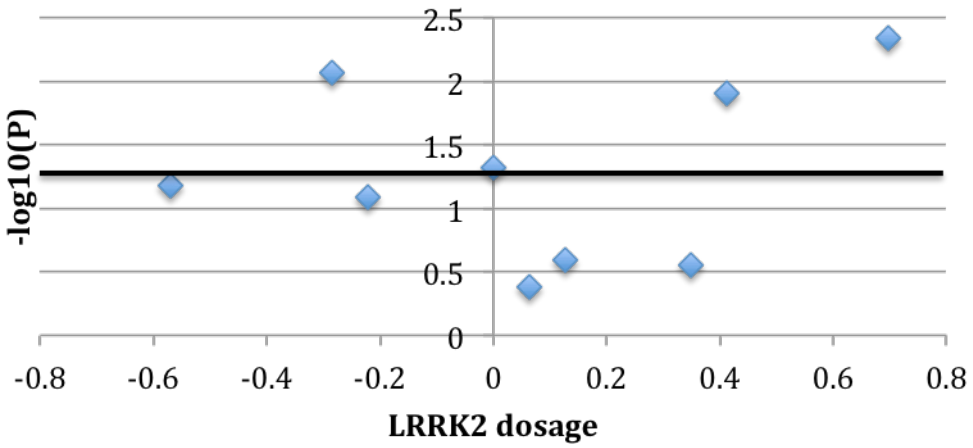
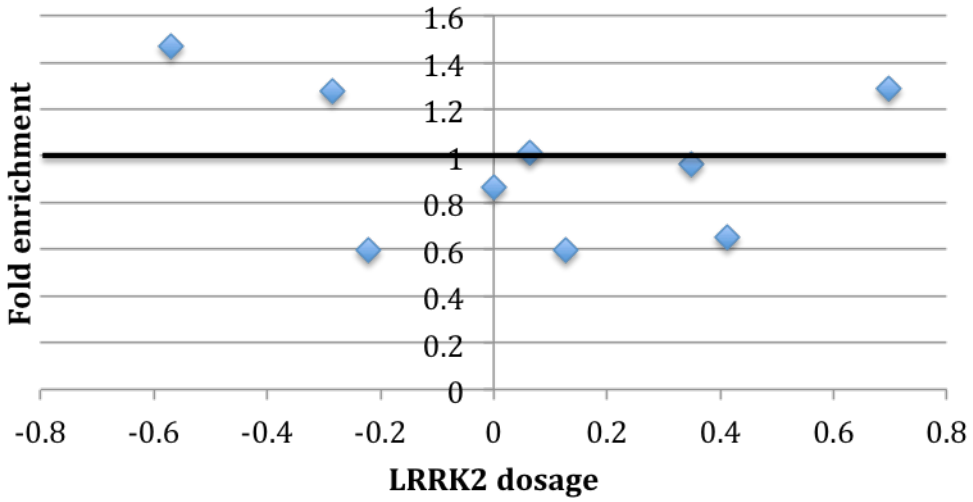


Figure 2.11: fold enrichment of each genotype in the APDGC cohort compared to its additive effect on LRRK2 expression. All significantly associated genotypes were enriched in the same direction as observed in the OPDC cohort.



increased expression were over-represented. Reduced LRRK2 expression therefore seems likely to be a favoured phenotype in the general population, whereas elevated LRRK2 expression was more prevalent among PD cases.

2.4) Discussion

2.4.1) Overview

In this analysis multiple data sources were integrated into a single gene network, which provided an effective means of comparing the similarity and linkage of multiple traits between genes. This was interrogated in tandem with eQTL data to elucidate SNPs that affected the expression of multiple PD-relevant and functionally linked genes. Four variants were identified within PD GWAS regions, two of which were cis-eQTLs of the known pathogenic gene LRRK2. Among these variants, allele combinations that caused the highest increase in LRRK2 expression were significantly enriched among PD cases.

2.4.2) Gene networks

The final gene network combined data from six different sources, including protein-protein interactions, co-expression and functional annotations among others. Integrating diverse data types such as these ensured that maximum information was considered, giving a more comprehensive understanding of each gene's function and interactions. By weighting information according to its relevance, noise from uninformative sources was removed, maximising specificity of the final network.

The network was made PD-specific by benchmarking against genes with PD-relevant phenotype annotations. Whole datasets and individual links not informative for PD were therefore removed. The resulting network was notably smaller, but offered a much greater degree of specificity.

This was demonstrated by the clustering of known pathogenic PD variants. Genes implicated by GWAS were significantly more linked than would be expected by chance in a PD-specific network, whereas no significant linkage was observed in a general network or a non-PD network. This confirmed that by using only phenotypes of interest the strength of disease-relevant links was up-weighted relative to uninformative ones, increasing the precision of the final gene network and aiding the identification of novel links.

Despite the use of both human and mouse phenotype annotations the network remained relatively sparse. Additionally, genes that may have been related to PD or connected to PD-linked genes, but did not yet possess PD-relevant annotations of their own, were absent. Although the PD-specific network offered increased capability for identifying novel links, it was therefore limited in its ability to identify novel genes compared to general networks.

2.4.3) EQTL analysis

In GWAS significant SNPs are often assumed to map to the closest gene, without consideration of any regulatory variants they might be tagging. Consequently links between GWAS hits and pathologically relevant changes in gene expression can remain undetected. Although cis-genes have a relatively high probability of being correctly assigned, the effects of trans-acting variants may be completely overlooked. In order to overcome this limitation, PD-

associated GWAS intervals were searched for eQTLs affecting the expression of a functionally linked cis- and trans- gene pair. It was hoped that this would identify variants responsible for the PD association of that region by affecting the expression of multiple genes in linked cellular processes.

Four PD-relevant eQTLs were identified, each located in a separate GWAS interval. Two affected expression in skeletal muscle, one in subcutaneous adipose tissue and one in pancreatic tissue. As PD pathology primarily affects the brain it would have been interesting to examine eQTLs affecting brain-specific gene expression, however there were not enough tissue samples available to confidently identify these variants.

Rs34179446 is a cis-eQTL of PDE10A and a trans-eQTL of NPY2R. NPY2R is a neuropeptide receptor involved in presynaptic inhibition of neurotransmitter release. Several SNPs within this gene have been associated with BMI and obesity [269]. PDE10A regulates cAMP- and cGMP-mediated intracellular signalling, particularly in the basal ganglia, and is key in the regulation of dopaminergic signalling [270, 271]. Loss of this protein is associated with longer PD duration and more severe motor symptoms [272]. Weight changes are predictive of motor symptom progression [273], so by affecting the expression of genes linked to both weight and motor phenotype severity this eQTL could explain otherwise inconspicuous links between phenotypes.

Rs12095503 influences the expression of FMO3 in cis and ARSA in trans. FMO3 catalyses the metabolism of certain xenobiotics. Among these are several modifiers of PD risk including nicotine and MPTP [274, 275]. Expression of this

gene can vary up to 20-fold between individuals, causing a highly variable metabolism rate of these compounds which may explain the incomplete penetrance associated with some environmental factors [276, 277]. ARSA mutations are associated with a neurological lysosomal storage disorder that shares some phenotypic similarities with PD. Lysosome function is linked to PD risk both physiologically and genetically through mutations within the GBA gene. This eQTL may therefore modulate PD risk through both innate cellular function and susceptibility to environmental factors.

Rs7303059 is a cis-eQTL of LRRK2 and a trans-eQTL of CISD1. CISD1 encodes a CDGSH iron sulphur domain protein known as mitoNEET. Both LRRK2 and mitoNEET are localized to the outer mitochondrial membrane and affect mitochondrial dynamics. However they act in opposing directions. LRRK2 promotes mitochondrial fission causing a depletion of functional mitochondria and production of reactive oxygen species (ROS) [125]. In contrast mitoNEET is involved in the regulation of oxidative phosphorylation and electron transport and therefore it's expression decreases ROS production [278]. Oxidative stress is central to PD pathology and this eQTL could alter the amount experienced by the cell through distinct pathways, affecting its susceptibility to cell death.

Rs1472118 is also a cis-eQTL of LRRK2 and regulates PPARGC1A expression in trans. PPARGC1A encodes PGC-1α, a protein involved in the regulation of energy metabolism and mitochondrial biogenesis, whose expression is reduced in PD patient brains [279]. Its activation results in increased expression of subunits of the mitochondrial respiratory chain and is required for the induction of many ROS-detoxifying enzymes, preventing dopaminergic neuron loss [279,

280]. However dysregulation of PGC-1α in either direction increases mitochondrial fission, exacerbating the mitochondrial fragmentation and oxidative stress phenotypes associated with increased LRRK2 expression [111, 281]. This eQTL is therefore linked to an increase in ROS production, whilst the simultaneous decrease of PPARGC1A expression reduces the cell's ability to cope with the additional burden. Mutations at this locus could therefore have a significant impact on ROS-mediated dopaminergic neuron loss.

2.4.4) Co-inheritance of LRRK2 variants

LRRK2 is a protein containing many interaction domains and is consequently involved in a variety of functions. A number of downstream effectors regulate mitochondrial fusion and fission, and it is also involved in the regulation of synaptic function. Therefore correct expression levels are a crucial component in maintaining healthy cellular function, whereas dysregulation is associated with neurodegeneration. Patterns of co-inheritance of the two LRRK2 eQTLs were investigated to identify any additive effects on gene expression that might affect risk of PD onset.

By modelling allelic co-inheritance under a null hypothesis, it was possible to identify genetic combinations that were over- and under-represented in the test populations. In total three genotypes were both significantly enriched among PD cases in the OPDC cohort and replicated in the APDGC cohort. One genotype was significantly over-represented in both PD cases and controls in the OPDC cohort, but this was not replicated. The total effect on LRRK2 expression for each genotype was calculated assuming an additive model.

The genotype that caused the greatest reduction in LRRK2 expression was significantly over-represented in both cases and controls in the OPDC cohort. This implies that there may be some phenotypic advantage conferred by this genotype that is unrelated to neurodegeneration. Both populations contained mostly elderly individuals so this could be connected to the aging process. Although this genotype was also over-represented in the APDGC cohort, it was not statistically significant. The role of reduced LRRK2 gene expression in the general population therefore remains unclear.

The genotype that conferred the greatest increase in LRRK2 expression was significantly over-represented in both PD case populations but not in controls. Genotypes that resulted in intermediate expression were consistently under-represented among PD cases. The overall trend therefore indicated that high LRRK2 expression was associated with PD onset, and correspondingly fewer PD patients produced normal transcript levels.

Over-expression of LRRK2 in mouse models has been shown to alter dopamine release, striatal signal transduction and dopamine tone, glutamatergic presynaptic plasticity and postsynaptic signalling [282]. Cell models demonstrate impairment of synaptic vesicle endocytosis and increased mitochondrial fission [111, 283]. Functional data therefore supports a role of increased LRRK2 expression as a major factor in susceptibility to PD.

The LRRK2 eQTLs were not in LD with any known pathogenic variants and are therefore likely to affect disease risk independently of them. However they may impart increased risk through similar cellular mechanisms. The G2019S mutation results in a 2-3 fold increase in LRRK2 kinase activity and the

I2020T mutation may cause an increase of up to 40% [284]. Increased LRRK2 gene expression is likely to increase total activity as more protein is present. Consequently eQTL variation could cause neurodegeneration via the same pathways as these known risk SNPs.

These results provide additional evidence for an important role of LRRK2 expression in PD pathology, whilst also implicating novel mechanisms by which changes may arise. However in a biological system both the quantity and activity of the protein mediate the total cellular effect. Variants influencing activity of the various protein domains should therefore be studied together with those affecting expression.

2.4.5) Conclusion

EQTLs are important sources of variation in gene expression, yet are often overlooked in the interpretation of GWAS results. In this analysis eQTLs were identified which were in high linkage with PD-associated SNPs and affected the expression of multiple PD-linked and functionally connected genes. Four variants were elucidated, which affect several PD associated pathways and that may explain correlations between diverse features of disease progression. Additive effects of two LRRK2 cis-eQTLs implicate increased LRRK2 expression in PD onset. Variants such as these may cumulatively explain some of the missing heritable component of PD risk, in addition to being a factor in sporadic disease.

Chapter 3: CNV analysis

3.1) Introduction

Copy number variants (CNVs) are regions of the genome greater than 1kb that have been duplicated or deleted. Despite their implication in a number of neurological disorders, including Alzheimer's Disease [285-287], they remain relatively unreported in PD. Targeted studies have confirmed effects of CNVs in PARK2 and SNCA [110, 288, 289]. However there have been few hypothesis-free studies aiming to identify novel variants.

Much of the remaining unexplained genetic component in PD is likely to be polygenic. Pathway- and network-based approaches were therefore employed to test functionally related groups of genes for association with the disease, in order to more accurately reflect the underlying biological system. Beyond the onset of disease, PD shows a high degree of heterogeneity in presentation and severity of phenotypes. Possible genetic causes of this were investigated by linking CNVs to phenotypic subgroups and the corresponding continuous measures of phenotypic variation.

3.2) Method

3.2.1) Data

3.2.1.1) dbGaP

Data was downloaded from dbGaP (Genome-Wide Association Study of Parkinson Disease: Genes and Environment; Study Accession phs000196.v2.p1) and consisted of genotype and demographic data for 2000 cases and 2000 controls [84]. The Illumina HumanOmniQuadV1 array was used, giving data on 500,000 roughly equally spaced SNPs along the genome. This dataset was used in the initial discovery phase.

3.2.1.2) OPDC

The replication dataset consisted of 991 cases and 270 controls from the OPDC Discovery cohort. Subjects were recruited from neurology clinics within the Thames Valley, with elderly healthy controls consisting of spouses and friends which had been approached by the participants themselves [290]. Genotype data was generated using the Illumina HumanCoreExome v1.1. This provided information on over 500,000 markers spanning the genome, including 250,000 concentrated in exonic regions. A number of phenotypic measurements were also collected, from a combination of self-evaluating questionnaires and in-clinic testing performed by trained neurologists and nurses [248].

3.2.2) CNV calling and Quality control

3.2.2.1) Initial quality control

Initial filtering for both cohorts was carried out in PLINK version 1.9 [259]. Individuals were excluded if genotypic and phenotypic sex was not concordant

or missing data was greater than 2%. SNPs were excluded if missing data was greater than 1%. In order to identify individuals of divergent ancestry, SNPs in linkage disequilibrium with r^2 above 0.2 were removed using the PLINK *indep-pairwise* function, based on a sliding window of 50 SNPs. This resulted in a subset of independently inherited SNPs on which to perform principle component analysis using EIGENSTRAT. Outliers represented samples of divergent ancestry and were subsequently removed. All SNPs were reinstated for the calling of CNVs.

3.2.2.2) Batch effects

Batch effects of microarrays, and consequently in calling CNVs, are widely acknowledged [291-295]. Small changes in environmental conditions such as operating procedures, temperatures and timings can affect the probe hybridisation and binding efficiency, which can cause systematic artefacts. When processed this can result in false associations that are a product of the experimental design rather than disease biology.

Guided Principle Component Analysis (gPCA) [296] is an extension of traditional Principle Component Analysis (PCA) that transforms a matrix of probe signal intensities (used to identify CNVs) by an indicator matrix encoding batch. This guides the analysis to look for batch effects, even if they are not the greatest source of variation. The proportion of total variance that can be attributed to each batch is quantified by comparing the first principle components from guided and unsupervised PCA over a series of randomisations of the batch vector.

Experimental procedures in the dbGaP dataset sought to minimise batch effects by randomising by disease state, recruitment site, healthy controls greater than 85 years of age, DNA extraction method and DNA storage time. Manual inspection of the files indicated that all samples were processed sequentially in a single run. In the absence of an obvious way to define discrete batches, the gPCA R package was run in unsupervised mode, equivalent to a standard PCA that identifies the greatest sources of variance. This was performed on SNP subsets consisting of ~125,000 autosomal and sex-linked probes equally distributed across the genome.

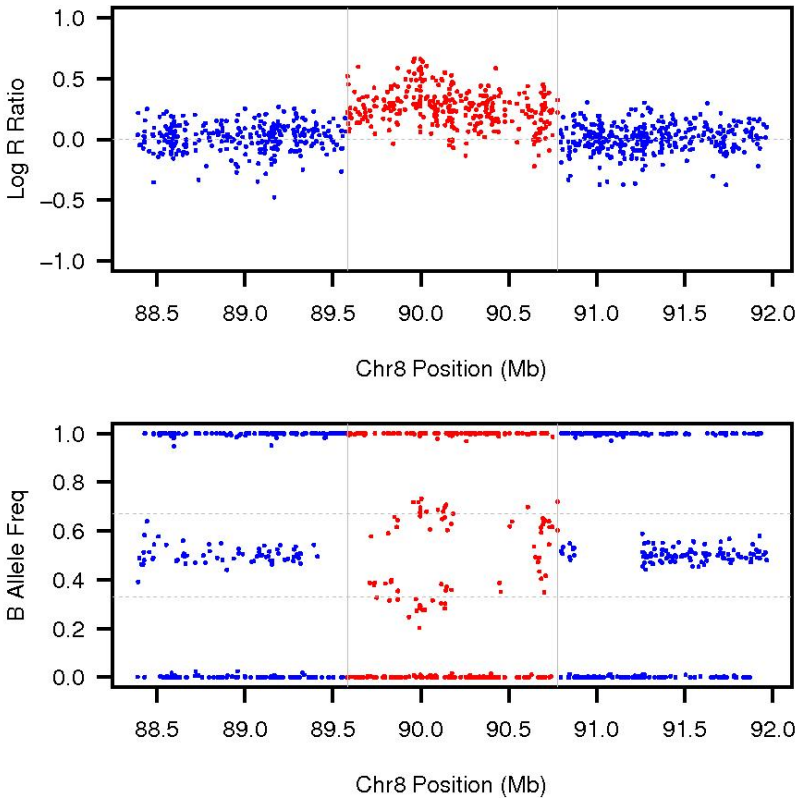
The OPDC data was generated at five different time points according to sample availability, and a batch vector was generated that corresponded to these groups. Guided PCA was then performed on SNP subsets of both autosomal and sex-linked variants as before. Default settings were used throughout.

3.2.2.3) CNV calling and quality control

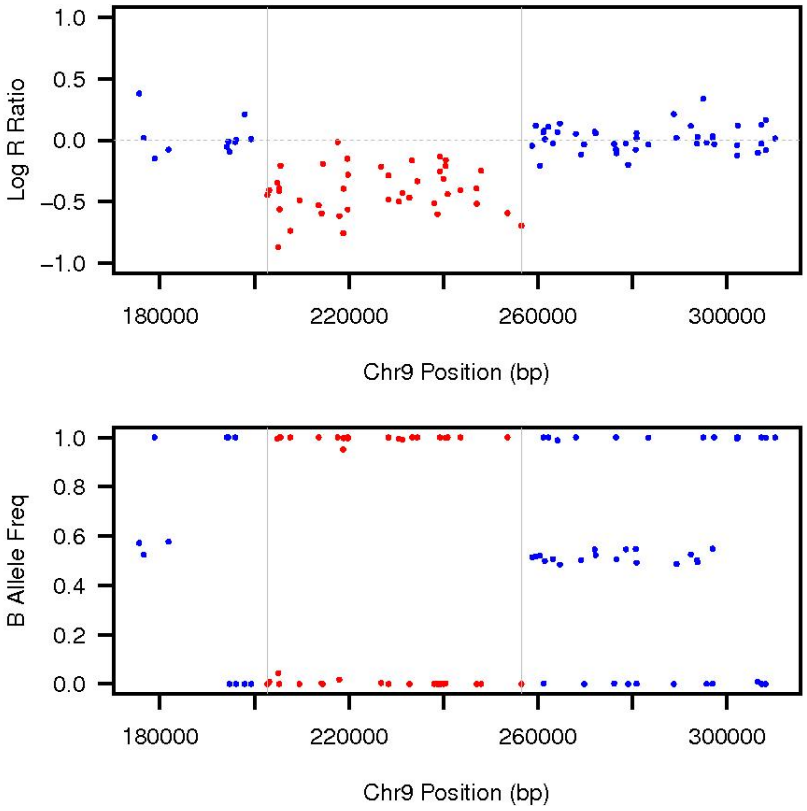
PennCNV was used to identify regions of copy number variation using a Hidden Markov Model (HMM), incorporating allele frequency and normalised signal intensity data for each probe[297]. Figure 3.1 shows the intensity (Log R Ratio) and allele frequency (B Allele Freq) of two high-confidence CNV calls for illustration. Part (a) shows a heterozygous deletion; probe intensity is lower compared to the surrounding region and the allele frequency segregates to 0 and 1 as only one allele is present at each locus. Conversely part (b) shows a heterozygous duplication, where probe intensity increases due to the extra copy

Figure 3.1: log R ratio and B allele frequency of high confidence duplication and deletion variants. At the duplication site (A) log R ratio is increased due to stronger probe binding and four allele states are observed. In contrast at the deletion site (B) log R ratio is reduced and only two allele states are present.

(A)



(B)



and allele frequencies segregate to 0, 1/3, 2/3 and 1, representing AAA, AAB, BBA and BBB alleles respectively.

Population-wide allele frequencies were used for comparison. These were provided for probes on the HumanOmni1QuadV1 array on which the dbGaP data were typed. However no such file was provided for the HumanCoreExome array. Consequently the "*compile_pfb"* script was used to discern estimates of allele frequencies for these probes based upon all individuals in the OPDC cohort.

The PennCNV algorithm was applied using the *"hhall.hmm"* model that is generalised for all arrays. In order to filter out poorly performing probes only those with less than 5% missing data were used. Occasionally large CNVs are called as several small, close variants. Such regions were merged using the *"clean_cnv"* Perl script provided, using default metrics.

Certain regions of the genome are prone to false positive calls, in particular telomeric, centromeric and immunoglobulin regions. Need *et al*. [298] have previously defined such regions and LiftOver was used to obtain these regions in UCSC Human Genome Build 19 co-ordinates (UCSC Batch Coordinate Conversion; https://genome.ucsc.edu/cgi-bin/hgLiftOver). CNVs overlapping any region by more than 50% of its length were excluded using the *"scan_region"* Perl script.

Individual sample quality control is summarised in Table 3.2. This included the removal of those with log R ratio standard deviation >0.28, B allele frequency drift >0.002, B allele frequency >0.55 or <0.45 or waviness factor >0.04, as deviation from these can indicate sample integrity issues. In order to

Table 3.2: Summary of quality control procedures and samples excluded at each step. The final dataset contained 719300 CNVs among 3806 individuals.

Quality control step	Number of samples excluded
Log R ratio standard deviation	0
B allele frequency drift	0
B allele frequency	0
Waviness factor	51
Total number	109
Total length	60

minimise false positive calls, only CNVs with at least 10 SNPs spanning the region and length between 1kb-500mb were included in further analysis. PennCNV also generates a confidence score equal to the difference between the log likelihood of the most and least likely copy number states. This was used to employ an additional quality control step whereby only variants with score greater than 10 were carried forward. Finally individuals with unusual numbers or aggregate length of CNVs were excluded, defined by greater than 3 standard deviations from the upper or lower quartiles. Thus, a list of high-confidence CNV calls was generated for both datasets independently.

3.2.3) Statistical analysis

PLINK v1.9 [259] was used to define rare CNVs in each population, for which no more than 50% of their length was subject to structural variation in more than 1% of individuals. The *cnv-indiv-perm* function was then used to examine overall enrichment of burden, length and proportion of events. P values are generated from 10,000 permutations of case-control status.

Rare CNVs were assigned to genes for which at least one base pair overlapped between the start of the first exon and the end of the last exon, according to the Human reference genome build hg19 (UCSC genome browser

August 2014). Pathway-based testing was then used. Analysing groups of genes that affect the same biological processes increases the power to detect disease association by incorporating additive effects and redundancy that may be undetected at the level of an individual gene.

In order to outline pathways mouse knockout phenotypes were used, as they have been studied and characterised more extensively than human phenotypes. Pathways were defined as groups of unique 1:1 human orthologues of mouse genes that when knocked down affect the same mouse phenotype (downloaded August 2014, Mouse Genome Database [255]). In this way 3330 pathways were defined.

Permutation-based testing was employed to examine the association of each pathway with disease. For each gene in a pathway a random replacement was chosen from the 100 genes most closely matched by length. This prevented false associations arising from the increased probability of CNV hits in longer genes. A null set of 'non-associated' pathways was thereby created. For each pathway the mean number of CNVs per person overlapping any gene within it was calculated for case and for control, and the test statistic defined as the difference between those means. This was repeated 1,000,000 times in order to generate an empirical P value for each pathway. FDR correction was performed.

In order to test the performance of this method, random pathways of varying lengths were selected in each category <50, 50-100, 100-250, 250-500, 500-1000 and >1000 genes. 1000 random case-control splits of the dbGaP dataset were performed to generate null populations. For each split this test was performed on all pathways and from this type I error rates were calculated.

3.2.4) Analysis of human PD phenotypic subgroups

Phenotypic subgroups defined by Lawton *et al.* were used to investigate the effects of CNVs beyond disease onset [248]. Psychological wellbeing, non-tremor motor and cognitive features define five discrete patient groups, as described in Table 3.3 [248]. For each variant, affected individuals were removed and the proportion of individuals expected in each subgroup calculated from the remaining cases. This was then compared to the proportions observed in affected individuals.

3.2.5) Network analysis

A disadvantage of pathway analysis is it's modular nature: although more relevant than the study of single variants genes are still classified into discrete groups, overlooking crossover between pathways. Phenotypic linkage networks developed by Honti *et al.* [253] were used to examine functional clustering of genes affected by CNVs in a more continuous framework. This approach integrates multiple data sources (such as functional annotations, protein-protein interactions and co-expression) in a weighted fashion that can prioritise PD-relevant links, more details of which can be found in chapter 1.

Four groups of genes were tested for functional clustering. Two groups

Table 3.3: The defining characteristics of each phenotype cluster identified by Lawton *et al.*

Phenotype cluster	Description
1	Mild motor and non-motor disease
2	Poor posture and cognition
3	Severe tremor
4	Poor psychological wellbeing, RBD and sleep
5	Severe motor and non-motor disease with poor psychological wellbeing

consisted of those genes hit either exclusively in cases or exclusively in controls, which indicates causation or protection under a full penetrance model. Due to the imbalance between case and control numbers in the OPDC cohort, it was required that a CNV be present in at least two cases to be included in this subset. Two further groups consisted of those genes that were significantly enriched ($p<0.05$) in either cases or controls as indicated by Fisher's exact test. Although not fully penetrant these variants may have an effect on disease under a polygenic model.

First, for each group the internal linkage was examined between genes within the set, to determine whether they were more functionally similar to each other than were random genes. This was carried out using the procedure developed by Frank Honti [253], with Python scripts provided by Cynthia Sandor. Functional similarity was quantified by summing direct links between all genes in the set. A null distribution of gene sets was then created, which consisted of permuting each gene from a pool of the 100 most closely matched genes in length and degree of connectivity in the network. This removed bias owing to hub genes (those which possess an especially high number and strength of links) and a general correlation between length and connectivity. From this an empirical P value was calculated.

Next, the linkage between each set and genes previously implicated in GWA studies was tested (as downloaded from the PDgene database; www.PDgene.org, March 2015). This was performed in a similar manner: permuting the gene set and examining the total linkage between genes in the test set and GWAS set. Finally, the same approach was used to determine whether

Table 3.4: summary of age and gender statistics. The proportion of males was significantly higher in PD cases than controls for both cohorts, likely due to the increased incidence of PD in men. The mean age of case and control individuals was also significantly different for both cohorts, however for all populations was above that of typical PD onset.

	dbGaP			OPDC		
	Case	**Control**	**P value**	**Case**	**Control**	**P value**
Age: mean (SD)	67.3 (10.7)	70.3 (14.1)	<0.0001	67.5 (9.61)	65.2 (10.29)	0.00269
Gender: % male	67.3	38.7	<0.0001	66.3	51.0	<0.0001

the gene sets associated with cases and with controls were linked to each other, in order to elucidate whether possible causative and protective variants functionally converge.

3.3) Results

The datasets differ only slightly in their demographic characteristics. Table 3.4 shows age and gender statistics for each cohort. Both showed a significantly higher proportion of males in cases than in controls, reflecting the increased incidence of PD in males. The average age of cases was similar in both cohorts. Although the control mean age differed, for both datasets this was at least 5 years older than the mean age of PD onset. The risk of individuals developing a neurodegenerative disorder in future was therefore low and consequently this difference should have had minimal effect. Both cohorts contained mostly individuals of European descent.

Figure 3.5 shows that the distribution of CNV length was similar in case and control for each cohort. In the dbGaP cohort this was similar to previous studies, with most variation made up of small CNVs less than 50kb in length. In

Figure 3.5: distribution of CNV length by population. CNV length is consistent between PD cases and controls within the same cohort, however is remarkably different between cohorts.

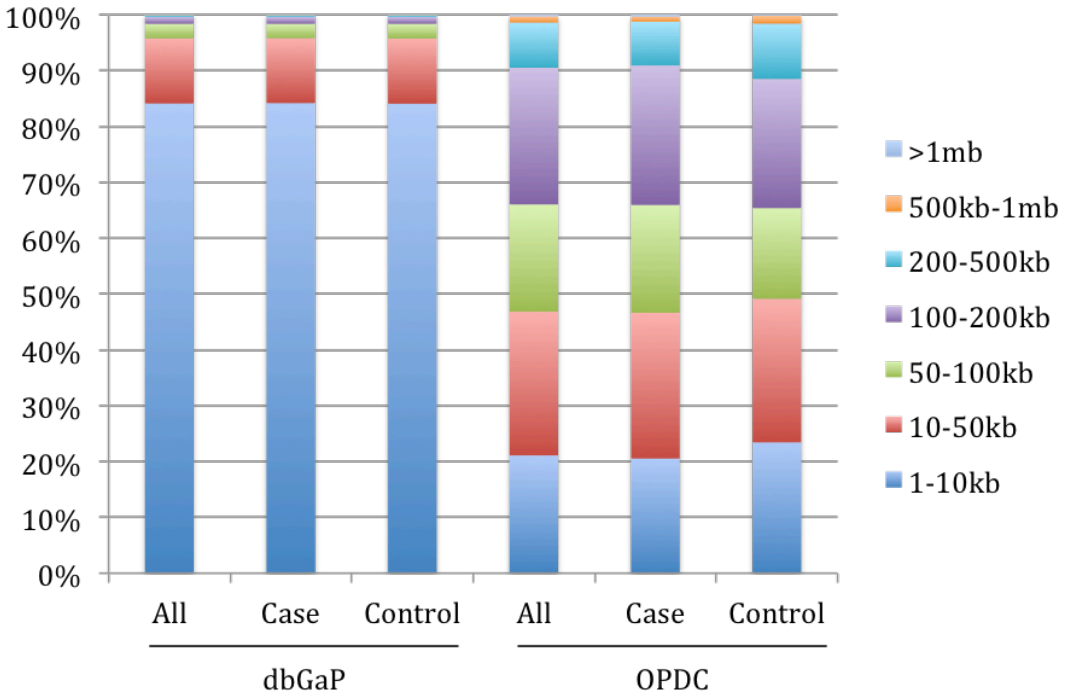
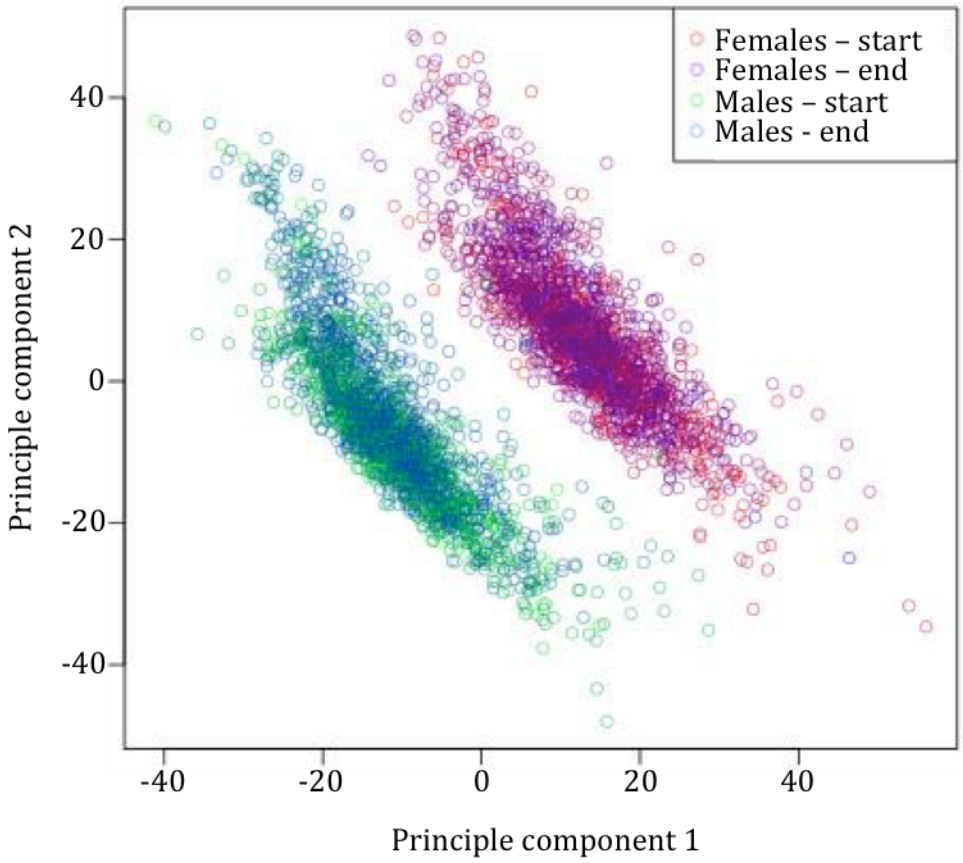


Figure 3.6: the first two principle components of PCA performed on the dbGaP cohort. Samples were genotyped sequentially and are shaded on a continuum to reflect this separately for males and females. This shows that genotyping order did not cause any confounding effects.



contrast these constituted less than 50% of the total variation in the OPDC cohort. In particular, variants less than 10kb comprised approximately 85% of all CNVs in the dbGaP cohort but fewer than 20% in the OPDC cohort. This is likely attributable to different resolution of the SNP array used for genotyping the OPDC cohort, a factor considered further within the discussion.

3.3.1) Batch effects

Figure 3.6 shows the first two principle components for the dbGaP cohort. The first two components primarily separate individuals of different gender, likely arising from the inclusion of sex-linked SNPs in the analysis as these vary systematically between males and females. Beyond that however no other sample groups are apparent. Furthermore each colour is graduated to reflect the position of the sample on the continuous genotyping run. Even scattering of colour demonstrates no systematic variation is present. This does not exclude the presence of batch effects but shows that there were greater causes of variance in the data than those caused by experimental design. Consequently they were not of primary concern.

During guided PCA applied to the OPDC cohort empirical P values were computed, equal to the probability that the proportion of total variance due to batch is greater than would be expected by chance. All SNPs were incorporated within 4 subsets, and values were consistently non-significant for the OPDC cohort (range 0.472 – 0.652). Figure 3.7 shows the first two principle components with individuals coloured according to their genotyping batch. Again, individuals are primarily clustered according to gender likely owing to the inclusion of sex-linked SNPs in each subset. Additionally the final genotyping run

Figure 3.7: the first two principle components for guided PCA carried out on the OPDC cohort. Genotyping was performed as seven independent procedures at five different time points. This graph shows that no batch effects were caused by this.

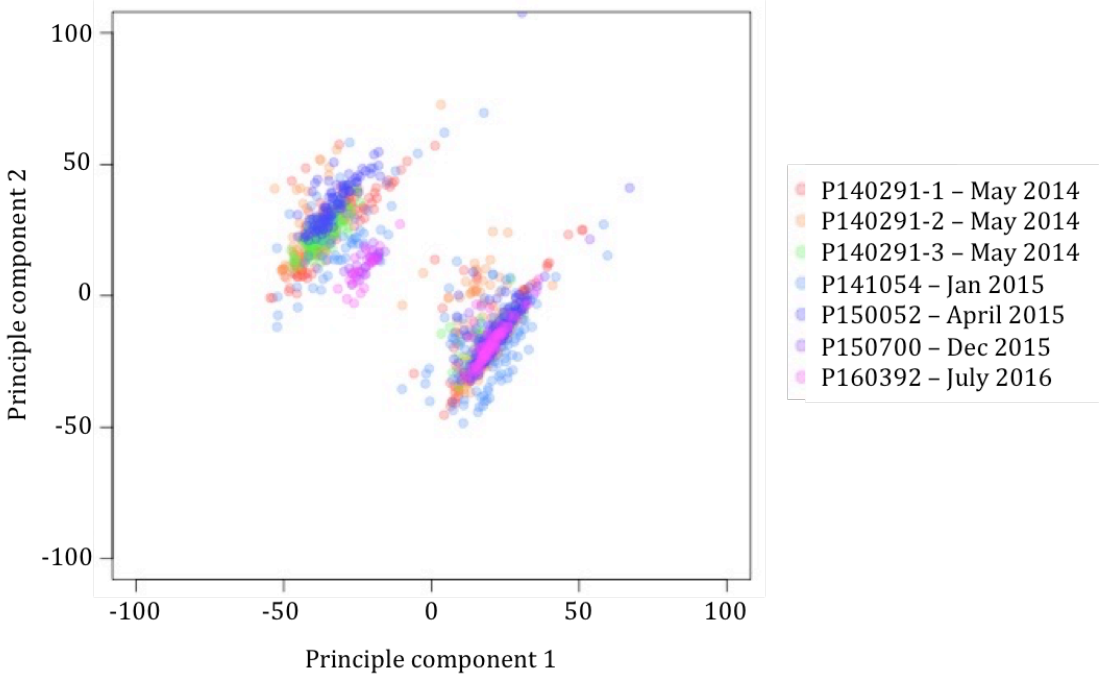
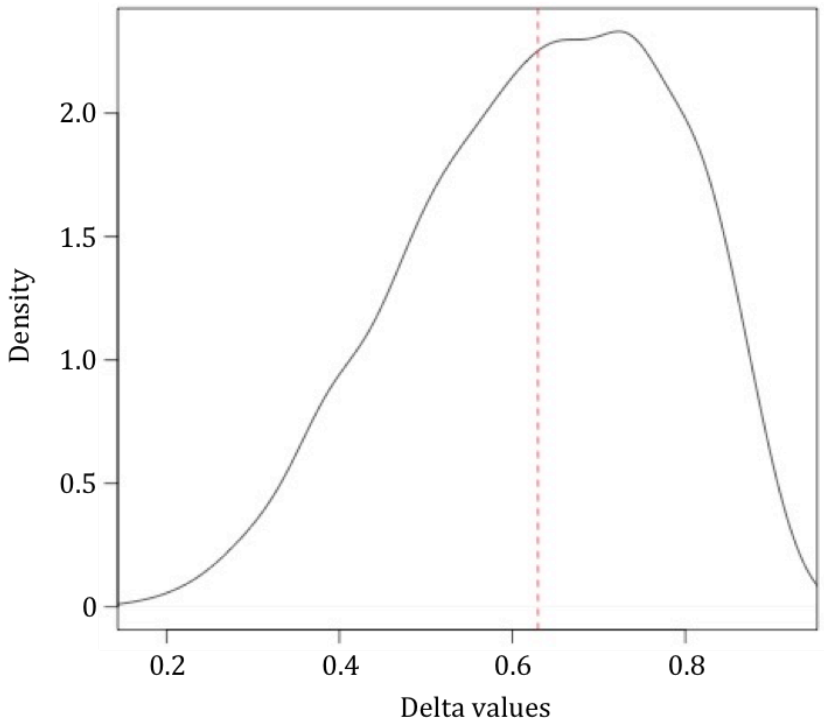


Figure 3.8: distribution of delta values from guided PCA permutations and the observed value in the OPDC dataset (red line). Delta quantifies the proportion of total variance that can be attributed to batch. This shows that in the OPDC cohort the effect of batch was no more than would be expected by chance.



performed in July 2016 shows some degree of separation from the remaining samples, which may be attributable to the use of an updated version of the SNP array. However figure 3.8 shows that the observed statistic was well within the range of those generated by permuted batches for a representative subset, indicating that the separation was not significant. Consequently it was concluded that batches should not be a confounding factor in the analysis of this cohort.

3.3.2) Enrichment analysis

The overall burden of CNVs in cases and controls was determined using the proportion of the population with at least one variant, average number of variants and average and total CNV length. This was performed both on the entire CNV set and rare variants present in <1% study population, then split further into deletions and duplications. Conservatively a Bonferroni correction was applied to account for multiple testing of six non-independent subsets (all and rare variants; all copy number states, deletions, duplications). Results are summarised in Tables 3.9 and 3.10**.**

Neither population showed significant difference between case and control in any measured characteristic when all CNVs were considered. A previous study in an Ashkenazi Jewish population found the total size of rare deletions to be significantly greater in PD cases than in controls, which was replicated in the dbGaP population [299]. In contrast the total length of duplications was less in cases but this difference was not significant. No other CNV characteristics were linked to PD onset in this cohort.

The OPDC cohort did not replicate the association with total rare deletion length. However around 6% more cases than controls carried at least one rare

Table 3.9: results from enrichment analysis carried out in the dbGaP cohort.

		Overall			Deletions			Duplications		
		Case	Control	P	Case	Control	P	Case	Control	P
All CNVs	Rate	188.7	189.3	0.793	153.4	153.1	0.326	35.28	36.15	0.991
	Proportion	1	1	1	1	1	1	1	1	1
	Total length	1915	1935	0.884	1210	1194	0.0667	704.9	740.9	0.997
	Mean length	10.17	10.23	0.777	7.894	7.791	0.053	20.99	21.46	0.874
Rare CNVs	Rate	7.935	7.867	0.36	4.847	4.672	0.128	3.088	3.195	0.855
	Proportion	0.998	0.999	0.887	0.976	0.976	0.509	0.903	0.9068	0.683
	Total length	347	343.9	0.402	156.4	143.6	**0.0407**	214.4	224.3	0.826
	Mean length	46.85	47.07	0.539	33.7	32.87	0.325	69.44	69.35	0.485

Table 3.10: results from enrichment analysis carried out in the OPDC

		Overall			Deletions			Duplications		
		Case	Control	P	Case	Control	P	Case	Control	P
All CNVs	Rate	3.162	3.038	0.24	1.406	1.409	0.522	1.756	1.63	0.187
	Proportion	1	1	1	0.726	0.726	0.537	0.785	0.7837	0.514
	Total length	307.1	307	0.498	161.5	146.8	0.159	242.1	255.8	0.745
	Mean length	108.2	116.3	0.833	89.48	93.28	0.692	123.3	136.3	0.857
Rare CNVs	Rate	1.882	1.764	0.194	0.868	0.860	0.491	1.015	0.9038	0.152
	Proportion	0.826	0.7644	**0.0271**	0.526	0.510	0.366	0.569	0.5144	0.0867
	Total length	240.2	251.7	0.695	145	142.2	0.451	214.6	233.2	0.757
	Mean length	117.2	130.9	0.865	95.31	107.9	0.852	139.4	151.3	0.741

variant: a nominally significant result that did not pass multiple testing corrections. Neither deletions nor duplications were in any way associated with disease phenotype, implying this was an overall enrichment. As mentioned previously however this could be due to the array design and consequent lack of detection of small variants, especially given almost 100% of the dbGaP population have at least one rare CNV.

3.3.3) Pathway analysis

Pathway analysis serves two main benefits over that of individual variants. Firstly, functionally similar but distinct genes may cause disease in different individuals. At a low frequency or with small population size statistical analysis of single genes is often underpowered to detect association, whereas the analysis of functionally similar groups has increased statistical power. Secondly, the majority of genes exert an effect as members of biological pathways containing compensatory mechanisms and redundancy. Disease caused by an accumulation of several mutations can therefore only be detected by pathway-type analyses.

Pathways were defined using human orthologues of mouse genes that affect the same phenotype in a knockout model. In a hypothesis-free approach 3330 pathways were tested for association with disease onset by permutation. An FDR correction was applied to compensate for multiple testing.

The method of pathway testing was verified using a null distribution that consisted of the random shuffling of case-control phenotype status. Under this model no association was expected and therefore the empirical false positive rate could be estimated. A variety of pathway lengths were tested. Figure 3.11 shows P values were uniformly distributed for all pathways, showing that this

Figure 3.11: quantile-quantile plots of P values associated with 6 pathways for 1000 random case-control splits of the dbGaP cohort. These demonstrate that there was no systematic bias in the significance values produced by this procedure.

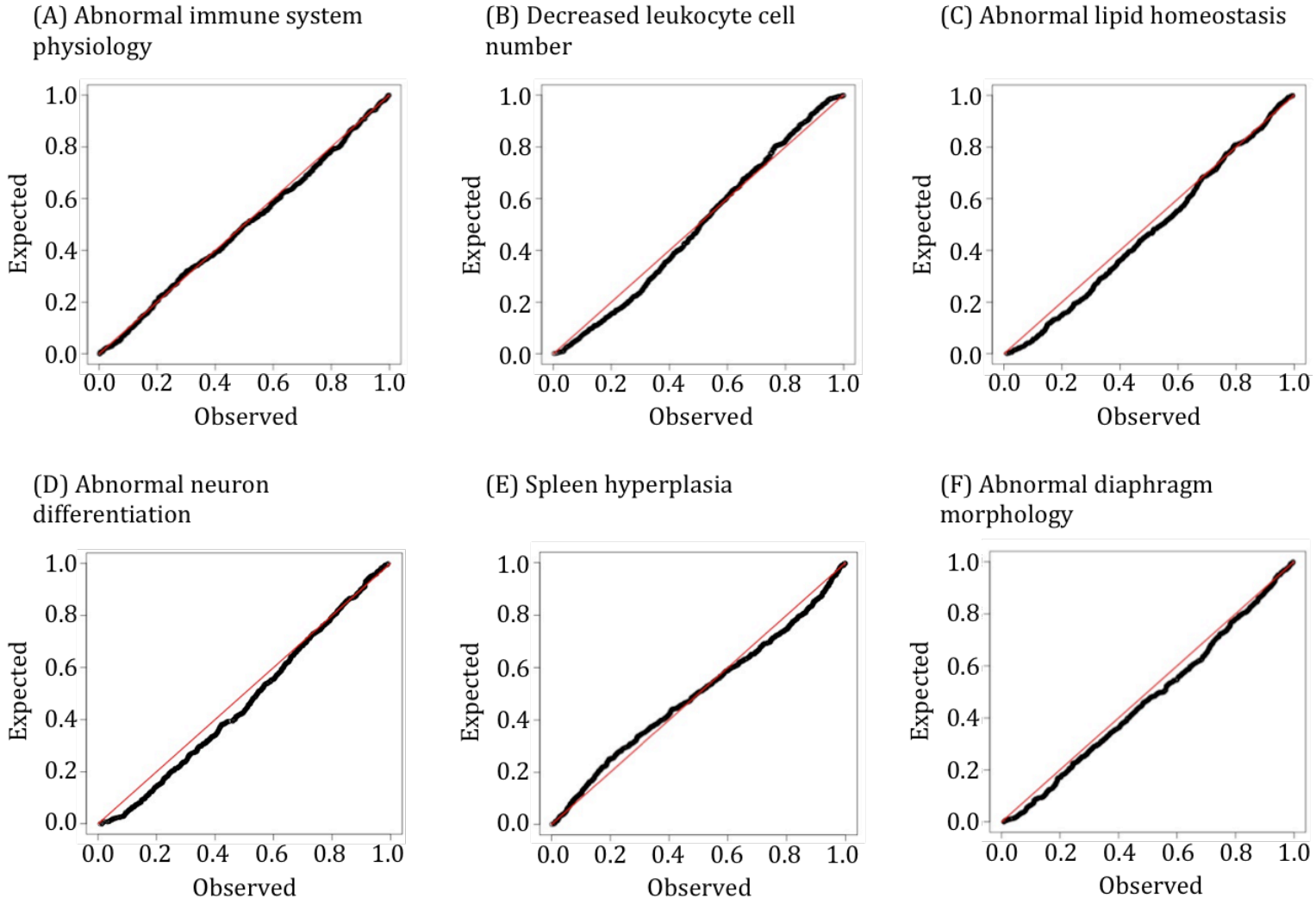
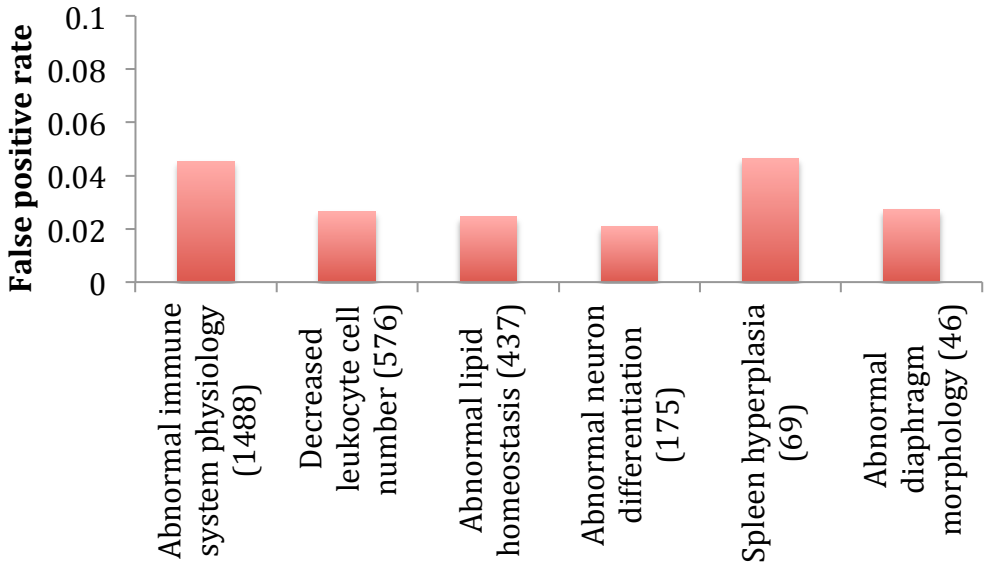


Figure 3.12: False positive rates calculated from 1000 random case-control splits of the dbGaP cohort. For all pathway lengths tested (the number of genes within each is stated below) the false positive rate was an acceptable level and this method was therefore suitable for use in statistical analysis.



test was unbiased. For all lengths of pathway tested the false positive rate was below or very close to 0.05 (Figure 3.12).

3.3.3.1) dbGaP dataset – discovery cohort

In the set of rare CNVs, seven groups showed significant enrichment in cases compared to controls: *abnormal mitochondrial morphology, abnormal pons morphology, abnormal substantia nigra morphology, abnormal tail movements, brain vacuoles, hippocampal neuron degeneration* and *decreased exploration in new environment.* Results are summarised in Table 3.13. On manual inspection of each pathway it emerged that this association was mainly driven by events within the PARK2 gene, whose almost 2 fold enrichment was nominally significant in itself (Fisher's test: p=0.0217).

Table 3.13: results with and without PARK2 for seven pathways that were significantly associated with PD onset in the dbGaP cohort.

Pathway	P value (uncorrected)	P value (FDR)[1]	P value without PARK2 (uncorrected)
Abnormal tail movements	0.000007	0.00999	0.278
Brain vacuoles	0.000009	0.00999	0.622
Abnormal substantia nigra morphology	0.000014	0.0117	0.625
Abnormal pons morphology	0.00002	0.0133	0.0662
Hippocampal neuron degeneration	0.000036	0.02	0.569
Abnormal mitochondrial morphology	<0.000001	<0.000001	**0.0291**
Decreased exploration in new environment	0.000061	0.029	**0.0346**

On exclusion of PARK2 from the analysis, only *abnormal mitochondrial morphology* and *decreased exploration in new environment* retained association with disease. The association of the former was again driven by mutations within just a single additional gene: PDSS2. This gene was 2.3 fold enriched in cases, however it was not significantly associated with disease onset (Fishers test:

Table 3.14: results of independent Fisher's exact tests for association of each gene within the *decreased exploration in new environment* pathway in the dbGaP cohort.

Gene	Number case CNVs	Number control CNVs	P value
PARK2	52	27	0.023
ALS2	8	3	0.226
ATXN1	8	3	0.226
RGS7	12	7	0.358
GRID2	34	29	0.612
GRM8	2	1	1
CPLX2	1	0	1
NAV2	1	0	1
TRMT1L	1	0	1

[1] Also known as Q-value

p=0.115).

On the other hand the *decreased exploration in new environment* pathway contained a number of variants contributing to the overall association, a full list of which can be found in Table 3.14. ATXN1 and ALS2 were both 2.7 fold enriched (p=0.226 individually) and RGS7 was 1.7 fold enriched (p=0.358) in cases. A number of additional genes contained 1 or 2 variants only in cases. Distribution of duplications and deletions varied, implying different roles in disease onset. ALS2 duplications appeared only in female cases with most variation comprising of deletions. In contrast, ATXN1 and RGS7 CNVs were dominated by duplications, ATXN1 entirely so. Heterozygous deletions of RGS7 were found in both case and control.

3.3.3.2) OPDC dataset – replication cohort

Due to the small size of this cohort and relatively few controls, it was likely that any test would be underpowered to detect association. For this reason only controls with no indication of REM sleep behaviour disorder (RBD) were used in the analysis, on account of the increased risk of developing PD in individuals with this disorder [300-303]. This increased discovery power by enriching for "true" PD controls that were unlikely to develop the disease in future.

Table 3.15 shows that of the seven pathways initially discovered in the dbGaP cohort the association of four with case status was replicated: *abnormal tail movements, brain vacuoles, hippocampal neuron degeneration* and *abnormal mitochondrial morphology.* Similarly to before however, once CNVs overlapping

Table 3.15: results of independent Fisher's exact tests in the OPDC cohort for association of genes within the *decreased exploration in new environment* pathway

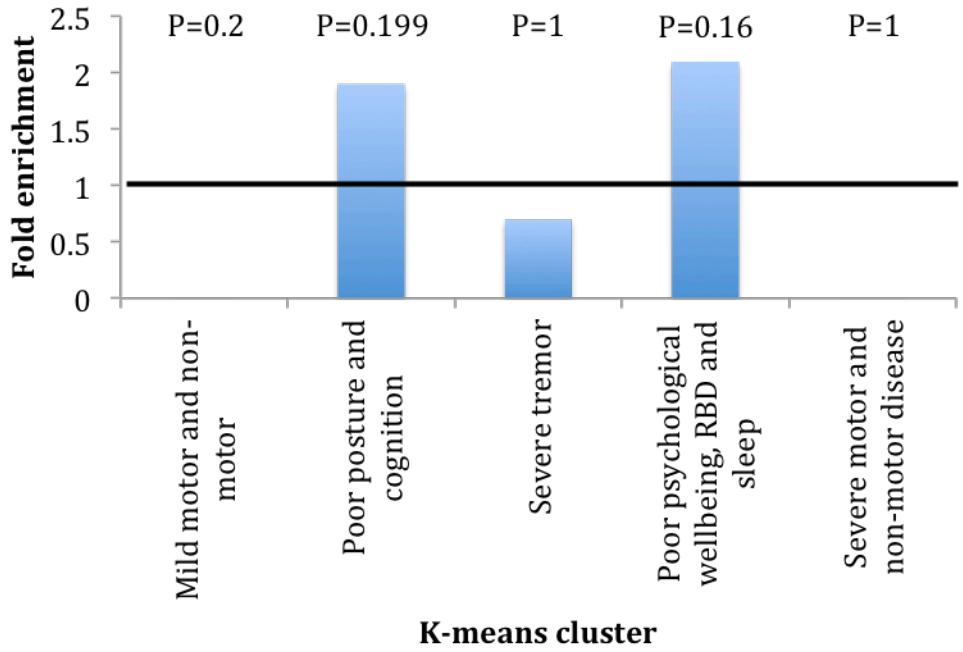
Pathway	P value	P value without PARK2
Abnormal tail movements	0.00176	0.489
Brain vacuoles	0.0001	0.159
Abnormal substantia nigra morphology	0.279	0.991
Abnormal pons morphology	0.136	0.681
Hippocampal neuron degeneration	0.00452	0.948
Abnormal mitochondrial morphology	0.03526	0.792
Decreased exploration in new environment	0.419	0.653

PARK2 were removed no association remained significant. PARK2 is not significantly associated with disease in this cohort (Fisher's test: P=0.544), despite being approximately 2-fold enriched once case:control ratio is accounted for. Furthermore, few other previously highlighted genes were affected by CNV in this cohort. Only a single event in RGS7 and four variants in TRMT1L were observed, all of which were in cases.

3.3.4) Phenotypic analysis in the OPDC dataset

Phenotypic subgroups were used to test whether genetic variants in PARK2 predispose patients towards a particular set of disease phenotypes, as this was the only gene with CNVs in enough patients to perform reasonable analysis. The subgroup distribution of PD patients with PARK2 CNVs is significantly different to that of the remaining case population (multinomial test, p= 0.00273). However no single cluster is significantly enriched in this subset (Fisher's test, see Figure 3.16). Instead there appears to be a general trend

Figure 3.16: fold enrichment of each phenotype cluster among PD patients with PARK2 CNVs in the OPDC cohort. P values are provided by Fisher's exact test.



towards clusters 2 and 4, characterised by poor posture and cognition and severe RBD respectively. Both demonstrate approximately 2-fold enrichment above that expected from a random sample of the PD population (Figure 3.16).

Two-sample T tests were used to compare a number of continuous phenotypic measures between cases with and without PARK2 CNVs. Only individuals within three years of diagnosis were used, in order to remove effects owing to disease progression rather than the disease subtype. Between the chosen phenotypes all defining features of the groupings were examined. Table 3.17 shows that no phenotype is significantly more severe in PD patients with PARK2 CNVs than in those without.

Table 3.17: results of Student's t-tests comparing continuous traits associated with phenotypic groups 2 and 4 between PD patients with and without PARK2 CNVs.

Phenotype	Feature	Population mean (SD)		P value
		No PARK2 CNV	PARK2 CNV	
UPDRS postural	Posture	2.75 (2.61)	3.71 (3.00)	0.2548
Semantic fluency (adjusted)	Cognition	9.96 (3.41)	8.93 (2.56)	0.161
Phonemic fluency (adjusted)	Cognition	10.9 (3.88)	10.9 (3.86)	0.977
MMSE total	Cognition	27.3 (2.27)	27.6 (2.50)	0.638
MOCA total (adjusted)	Cognition	24.8 (3.48)	24.5 (4.45)	0.797
RBD total	RBD	4.75 (3.04)	5.64 (3.61)	0.376

3.3.5) Network analysis

By using networks incorporating a variety of gene information it was possible to examine the degree of functional similarity within gene sets implicated by rare CNVs. Four groups of genes were defined. Firstly genes affected by CNVs either exclusively in cases or exclusively in controls, as such variants imply full penetrance in onset or protection. Secondly, genes significantly associated (Fisher's test $P<0.05$) with case or control status, which may be involved in PD within a polygenic framework.

The degree of functional linkage between genes in the same set was tested to examine how functionally related they were to each other. Linkage with known PD GWAS genes was also examined in order to test whether genes affected by SNP variation and by CNV functionally converge. A Bonferroni correction was applied to account for testing of four non-independent gene subsets.

Linkage testing was performed using the strongest 500,000 links in the non-specific network and 50,000 links in the PD-specific network, allowing for a similar proportion of noisy links in both analyses. 10,000 permutations were carried out.

3.3.5.1) dbGaP CNV clustering

All results are summarised in Table 3.18. Both case-associated gene sets demonstrated clustering among themselves to a nominally significant level, indicating some degree of functional similarity in those variants implicated in disease onset. Furthermore those statistically associated with case status were highly functionally linked to genes previously implicated in PD onset by GWAS studies, showing convergence of molecular mechanisms of SNP and CNV disease-associated mutation.

Both control-associated gene sets also showed internal clustering, and this was most significant in those variants exclusively found in controls. Neither of these sets was significantly associated with known PD onset genes. Despite this there is strong linkage between the disease-associated CNV set and the control-associated CNV set (P<0.0001), showing that mechanisms of onset and

Table 3.18: significance of functional clustering of genes associated with PD case or control status in the dbGaP cohort. Both internal links and links to genes implicated in PD onset by GWAS were examined in PD-specific and general networks

		General network		**PD-specific network**	
		Internal	**PD GWAS**	**Internal**	**PD GWAS**
Statistically associated with	**Case**	*0.025*	**<0.0001**	-	-
	Control	*0.0256*	*0.0424*	-	-
Only present in	**Case**	*0.0146*	0.435	0.157	0.054
	Control	**0.0001**	0.0949	0.0679	0.0906

protection are not dissimilar.

Finally, applying the analysis to a PD-specific gene network demonstrates clustering that is almost significant between GWAS genes and those with CNVs only present in cases. This lends some additional support to the functional convergence of SNP- and CNV- driven risk variation. Although more focussed this network is also more sparse, consequently there were not enough links to confidently test statistically associated genes.

3.3.5.2) OPDC CNV clustering

Due to the excess proportion of cases over controls in this cohort over 1700 genes were present exclusively in cases. Consequently it was required that at least two CNVs were observed to be included in this set. Internal clustering of genes only present in case or in control was almost significant in the general network. This lent some support to results observed in the dbGaP cohort, which showed functional similarity of genes linked to disease onset and protection. However no significant linkage is observed in a PD-specific network or with PD GWAS genes for either set (Table 3.19). The cohort was underpowered to detect association of single variants, therefore gene sets statistically associated with disease phenotype were not defined.

Table 3.19: significance of clustering of genes linked to PD case and control status in the OPDC cohort. Internal links and links to genes implicated in PD GWAS were examined in the general and PD-specific networks

		General network		PD-specific network	
		Internal	**PD GWAS**	**Internal**	**PD GWAS**
Statistically associated with	**Case**	-	**-**	-	-
	Control	-	-	-	-
Only present in	**Case**	0.0612	0.387	0.5207	0.1244
	Control	0.0678	0.9258	0.5458	0.8237

3.4) Discussion

Overall it was demonstrated that some of the remaining unexplained genetic component in PD may be explained by copy number variation. The effect of global CNV burden was examined and previous results were replicated associating an increase in deletion length with disease onset. Next a hypothesis-free pathway analysis was performed, highlighting novel genes involved in PD-relevant pathways. Functional similarity of implicated genes was investigated via network approaches and showed that those associated with disease onset and neuroprotective effects converge to similar mechanisms. Finally, the effect of CNV on phenotype presentation was examined.

3.4.1) Overall enrichment analysis

A significantly greater total size of rare deletions in cases was observed in the dbGaP cohort, replicating the only other study examining the effect of global CNV burden in PD [299]. In these individuals a greater number of genes are therefore likely to be affected by mutation, which points toward an additive effect of structural variation on disease risk. No enrichment is observed in total deletion number or average length, indicating a slight increase in both features not of statistical significance individually.

No link between case status and deletions is observed within the OPDC cohort. Instead, there is nominally significant difference between the proportion of case and control individuals with at least one rare variant. 82.6% of cases and 76.4% of controls have at least one rare CNV. However almost all individuals in the dbGaP population have at least one such variant. Furthermore the average number of CNVs identified per individual was comparatively fewer in the OPDC

cohort, and the contribution of small variants to this total was reduced by over 60% compared to other studies. The SNP array used in this cohort contains half the number of equally spaced SNPs, and although the remainder are more focussed in coding regions this could have resulted in lower resolution across some areas of the genome. Due to this, or perhaps other features of the array, it may therefore be possible that several smaller variants escaped discovery. Consequently this enrichment could be an artifact of array design and the lack of identified small CNVs, rather than due to genuine underlying biology.

Overall these results do not give a coherent conclusion. However given the previous study and the characteristics of each cohort the findings from the dbGaP study are likely to be more reliable. Consequently an increased total size of rare deletions may be linked with PD onset.

3.4.2) Pathway approaches

A hypothesis-free genome-wide study of rare copy number variation would have been greatly underpowered to detect association of single variants or genes in these cohorts. Consequently a pathway-based approach was adopted, which by aggregating the analysis of several functionally linked variants increases the power to detect association. Such analyses are also more biologically relevant to polygenic diseases, where regulatory and compensatory networks can mask the effect of individual variants. This is particularly applicable to PD, where variable penetrance of the majority of known risk variants indicates the presence of modifying factors.

Seven pathways showed association with disease onset in the dbGaP cohort: *abnormal mitochondrial morphology, abnormal pons morphology,*

abnormal substantia nigra morphology, abnormal tail movements, brain vacuoles, hippocampal neuron degeneration and *decreased exploration in new environment.* Several of these brain regions and biological processes are known to be involved in human PD pathology. On further inspection it was established that the association of five were driven solely by mutations within the PARK2 gene. This gene is located within a common fragile site and is therefore highly susceptible to copy number events [304]. CNVs at this locus have previously been implicated in PD [289]. However not all pathways containing PARK2 were associated with disease. So although single genes rather than pathways were driving the associations, pathways were identified in which PARK2 imparts an effect that is subsidised by additional variation.

The *decreased exploration in new environment* pathway remained disease-associated on exclusion of PARK2. This is a behavioural and neurological phenotype descending from either *abnormal emotion/affect behaviour* or *abnormal learning/memory/conditioning*. Such defects likely reflect the cognitive phenotypes, commonly depression and memory impairment, observed in human PD.

Among the remaining genes most case-enriched variation was observed within ATXN1, ALS2 and RGS7: genes linked to ataxia, amyotrophic lateral sclerosis (ALS) and G-protein mediated signalling respectively. RGS7 encodes a protein that accelerates GTPase activity in striatal neurons [305], controlling dopamine signalling together with RGS9-2 and R7BP [306]. Ataxia and ALS2 have also been linked to GTPase activity [307-309], as have known PD risk mutations PINK1, Parkin and LRRK2 [128, 310-312].

Both ataxia and ALS manifest clinical motor phenotypes similar to those observed in PD. Furthermore genes causing hereditary ataxia have been linked with Parkinsonism [313]. Misdiagnosis is unlikely to be the cause of this enrichment given patients whose PD diagnosis changed within the 12 year follow-up period were excluded. The molecular cause of ataxia is not well understood. However ALS is associated with neuronal death, albeit affecting different brain regions and neuron types to those implicated in PD pathology.

Abnormal mitochondrial morphology was the only other pathway to remain associated with PD when PARK2 was removed from the analysis. Mitochondrial dysfunction is a known factor in PD onset and is associated with a number of known PD risk variants. Neurons have high energetic demand and mitochondria are actively moved to areas of high activity such as synaptic regions [314-317]. Mutations affecting a cell's ability to generate ATP where it is needed will therefore manifest phenotypes in neurons first.

Disease association was largely driven by a single other variant, PDSS2. Although not previously implicated in PD onset, PDSS2 missense mutant and conditional knockout mice show neuromuscular defects and TH+ neuron pathology similar to that observed in sporadic PD [318]. Neither PARK2 nor PDSS2 were significantly associated with disease onset in GWAS performed on this cohort, implying this association is not attributable to underlying SNP variation.

Analysis performed in the OPDC cohort did not elucidate any additional disease-linked pathways. However the association of four pathways identified in the dbGaP cohort was replicated. *Abnormal tail movements, brain vacuoles,*

hippocampal neuron degeneration and *abnormal mitochondrial morphology* all showed an enrichment of CNVs in PD individuals.

Abnormal tail movements likely reflect a rodent manifestation of motor dysfunction. The remaining cellular phenotypes have all been previously linked to PD. An increased number of autophagic vacuoles is present in the brains of PD patients [83], which may represent abnormal activation of macroautophagy [319]. Reduced hippocampal activity and increased hippocampal alpha-synuclein deposits are associated with Parkinson's disease dementia, although not with PD overall [320].

As before, these associations were driven by variants in PARK2 and none remained associated with disease when this gene was excluded from analysis. However not all previous associations were replicated, and the strongest associations were not always observed in the smallest pathways as would be expected if this was the only contributing gene. It is therefore likely that other variants in the pathway also play an important role in disease aetiology. In this study the remaining genes contain few variants however, so a larger cohort would be required to separate disease-associated variants from benign ones.

PARK2 again demonstrated a high mutational rate in this population. It's approximately 2-fold enrichment in cases is similar to that observed in the dbGaP cohort. All variants were heterozygous, and both deletion and duplication events were present in both case and control. CNVs in this gene therefore demonstrate incomplete penetrance, as has been previously observed.

3.4.3) Network analysis

Network analysis has similar benefits to pathway analysis: increasing power of discovery and biological relevance by considering functionally similar genes. However continuous linking of genes, rather than their classification into discrete subgroups, allows a greater degree of flexibility in the analysis as indirect and variable strength links can be considered. Additionally more information is incorporated than the mouse phenotypes used to define pathways, providing more extensive information on gene relatedness.

In the dbGaP cohort gene sets statistically associated with either case or control status both showed a high degree of functional similarity in a non-specific network. Variants implicated in PD onset or prevention therefore converge to similar molecular mechanisms, supporting a polygenic disease model. Furthermore genes affected by CNVs exclusively in case or in control were also functionally similar. Although these sets are likely to contain some benign variants, in particular those present only once or twice in the population, genuine associations confer high penetrance. Functional convergence indicates that a number are likely to be true associations and that individual rare CNV events may be an important factor in the onset and prevention of PD.

Variants associated with case status also showed high linkage with known PD GWAS genes. This demonstrates that different structural variation within similar pathways can cause the same disease phenotype. Previously this has been observed in hypothesis-led studies of genes already identified via SNP variation, such as SNCA and PARK2, however has not been investigated in a genome-wide manner. Consequently rare CNV variants affecting known disease

pathways may explain some of the gap between expected heritability and known genetic causes in PD.

Case associated genes also show a high degree of linkage to control associated genes, indicating that mutations increasing or decreasing disease risk affect similar molecular pathways. Variants acting within the same pathways to cause opposing effects imply that direction of effect is crucial. Analysis of deletions and duplications separately would therefore be beneficial in a larger cohort.

Controls in this study are screened against a number of neurological disorders not restricted to PD. They are also age-matched and consequently unlikely to develop any neurological disorders in the future. The internal clustering observed is therefore likely to reflect a module protective against neurodegeneration in general.

Clustering in the PD-specific network could only be tested for genes exclusively present in case or control, due to the size of the gene sets. Although the PD-specific network is more focussed than a general network it is also sparser, as only a small fraction of genes are annotated with PD-linked phenotypes. Consequently smaller gene sets did not contain enough genes within the network to perform informative testing. No significant results were observed for either cohort, which indicates that genes affected by CNVs in cases and in controls affect a wide range of functions amongst those already linked to PD.

Owing to a lack of power there were not enough statistically associated genes, with either case or control, to test functional clustering in the OPDC

cohort. Those genes affected by CNVs exclusively in case or control were defined, however given the small size of this cohort it is expected that many of these are not truly disease-associated. Instead their exclusivity can be attributed to a shortage of samples in which to identify recurrent variants. As expected, functional clustering is not observed in either a non-specific or a PD-specific network for these genes.

3.4.4) Phenotype analysis

Previous studies have shown that phenotypic patterns at onset are predictive of disease progression [321]. SNP variation is also associated with disease course [322]. Here it is shown that CNV variation in PARK2 can also influence subphenotype, resulting in an approximately 2-fold increase in risk of developing a phenotypic classification associated with severe RBD or poor posture and cognition at disease onset. However this is not reflected in differences of observed continuous variables measuring these characteristics. Additionally this contradicts previous literature linking PARK2 to a decreased risk of dementia [313].

Due to small sample size this could be attributable to either low statistical power or a false positive result. Although PARK2 is particularly susceptible to copy number events only seven were observed in cases within this cohort. Consequently tests of continuous measures were underpowered to detect differences between these individuals and the remaining PD patients. The enrichment of phenotypic groups observed here may therefore reflect true effects of these variants, but more study is required to confirm enrichment of specific phenotypes. Alternatively these findings could represent a false positive

result, as with such a small sample just a few patients with abnormal phenotypes could significantly alter the observed phenotype distribution. The effect of CNVs in PARK2 on PD phenotypes remains unclear and additional study is required.

Phenotype classifications defined by Lawton *et al.* [248] were used in this analysis because of their increased specificity over the more commonly used Postural Instability Gait Disorder (PIGD) / Tremor Dominant (TD) subtypes. However many attempts have been made to define discrete phenotypic subgroups that accurately reflect the variation in phenotypic presentation across the population. It is not possible to say which, if any, is biologically relevant. Consequently discrete groups may provide a good starting point, but in the long term they may hide underlying continuous variation not reflected in the categorical nature of groups.

3.4.5) Study limitations

Many studies of PD show varying results concerning CNVs. Environmental effects are well known to affect PD risk, yet even considering this individual studies demonstrate a lack of coherency. For example, previous studies have shown CNVs in PARK2 to be not associated with PD [323], associated with PD [289] , associated with ADHD [324], and both associated and not associated with PD using different methods applied to the same data [325].

Largely this is attributable to variation between CNV calling algorithms. Applied to the same data, CNV calls from different software demonstrate <50% concordance [326]. Furthermore the same algorithm applied to replicate data show <70% reproducibility [326]. Even consensus calls (those identified by 2 or more algorithms) can sometimes not be validated [325], indicating significant

artifacts somewhere within the experimental or analytical pipeline. Given such uncertainty it would have been desirable to validate interesting variants experimentally, for example by qPCR, however this was not feasible.

Although identical methodology was used for analysis of both cohorts these factors may still confer differential effects. Different array types were used for each cohort and could cause systematic artefacts. The HumanOmni1QuadV1 used for the dbGaP cohort contains 500,000 roughly equally spaced variants, whereas the HumanCoreExome used for the OPDC cohort contains only 250,000 spaced variants, the remaining 250,000 concentrated in exonic regions. This results in varying coverage among different regions of the genome between the array types, which may affect the number and reliability of CNV calls within different genes.

As a result certain regions have much poorer resolution in this array, which reduces the ability to detect CNVs and accurately map their breakpoints. This is reflected in the lower number of small variants detected in the OPDC cohort, contributing towards a lower total number of variants called overall. In contrast, CNVs in the dbGaP cohort are of a similar number and size distribution to previous studies.

There is currently no consensus for annotating CNVs to the genes that they affect. Annotation in this work used a minimum of one base pair overlap between the beginning of the first exon and the end of the final exon. A more stringent threshold commonly used requires that a CNV overlap a minimum number of exons. However CNV boundaries called by algorithms are inaccurate, as they are calculated from probabilistic inference between SNP markers that

can be 1000s of base pairs apart. This method could therefore underestimate the true number of CNVs affecting the gene. Furthermore a variant need only overlap one exon to affect the translation of the whole gene, for example by deleting a start codon. Consequently by relaxing somewhat the threshold for gene assignment of a CNV it was hoped that more variants affecting the gene could have been included. This carries a risk of increased false associations, however such events should affect cases and controls equally and not affect the analysis.

Sample size and population structure is also a limiting factor in this analysis. In particular, the OPDC cohort used for replication contained only 250 controls and 850 PD patients. Excess numbers of cases compared to controls limits the power to discover control-associated variants and can inflate the significance of case-associated variants. Additionally such a small total number, especially when examining rare variants, means that only the strongest associations can possibly be detected. The probability of false positive associations is also increased, particularly when sample size is further reduced as in the phenotypic analysis. Unfortunately CNV studies cannot be combined or imputed due to batch variation. They therefore remain limited in power until methods are developed that allow the reliable comparison of data from different studies.

3.4.6) Conclusion

Overall it has been shown that copy number variation affects PD onset, appearing to converge on genes involved in mitochondria- and GTPase-related processes. No association appears to be fully penetrant, indicating that other factors – genetic or environmental – are involved. It has also been demonstrated

that variants not associated with disease individually are linked with onset as members of a genetic pathway. Consequently pathway analysis is shown to be a valuable tool to elucidate functionally similar and interacting genes that otherwise escape disease association. Network analysis confirms that CNVs associated with disease onset affect functionally similar genes and that these are also related to genes affected by SNP variation. Additionally genes affected by CNVs that may reduce the risk of PD demonstrate functional similarity both between themselves and with those implicated in PD onset.

Chapter 4: Binary phenotypes in Parkinson's Disease

4.1) Introduction

A genome-wide association study (GWAS) was carried out to compare Parkinson's Disease (PD) case and aged control. Individual SNP variants are of limited clinical use, so enrichment analysis was carried out to elucidate the underlying pathways associated with disease onset. Genes annotated with *nucleotide binding* and *nuclear heterochromatin* were enriched among SNPs linked to PD onset. This implies that distinct genetic variants implicated in PD onset may converge towards similar biological functions.

The analysis was then extended to compare different phenotypic subgroups of patients. Each subgroup was systematically compared to the remaining four using the GWAS approach. Enrichment analysis was carried out using those results. Cluster 5, associated with severe motor and non-motor phenotypes with cognitive decline, was significantly linked to *guanyl nucleotide exchange factor activity*. Genes annotated with this term affect guanyl nucleotide binding and are a subset of *nucleotide binding.* This demonstrates that although broadly similar mechanisms may cause disease onset overall, perturbations affecting specific pathways within that may cause the onset of particular phenotypes.

4.2) Methods

4.2.1) Datasets

Two datasets were used in this analysis. The Oxford Parkinson's Disease Centre (OPDC) Discovery dataset provided 991 cases and 270 controls recruited from the Thames Valley area. Cases were diagnosed using UK-PD Brain Bank criteria and only individuals with greater than 90% chance of PD at 18-month follow-up were included in the analysis. Control individuals were mainly spouses and carers of PD patients who had no past or present diagnosis of PD or other neurological condition. Data was also collected for 107 at-risk individuals, comprised mainly of first-degree relatives of PD patients. An additional 108 individuals diagnosed with REM sleep behaviour disorder (RBD) were recruited to the study due to the increased incidence of PD associated with this condition. Blood samples were drawn by clinician and processed by Sam Evetts. Genotype data was generated at the Wellcome Trust Centre for Human Genetics using the Illumina HumanCoreExome-12 v1.1 and Illumina InfiniumCoreExome-24 v1.1 SNP arrays. Each consisted of 500,000 SNPs, half of which were exome variants.

An additional 3007 aged controls were used from a study of Late-Onset Alzheimer's Disease (LOAD) (National Institute on Aging - Late Onset Alzheimer's Disease Family Study: Genome-Wide Association Study for Susceptibility Loci; dbGaP accession phs000168.v2.p2). All participants underwent neurological evaluation and were screened for any past or present diagnosis of PD in addition to a number of other neurological disorders. The Illumina Infinum II assay protocol was followed and genotyping was carried out on the Illumina Human610 Quadv1 array.

4.2.2) Quality control

Quality control was carried out independently for the two datasets using PLINK v1.9 [259], as summarised in Table 4.1. Variants were excluded if minor allele frequency (MAF) was less than 0.01, Hardy-Weinberg Equilibrium (HWE) P value was less than 0.00001 or missing data rate was above 5%. Individuals were excluded if genotypic and phenotypic sex was discordant, missing data was greater than 2% or heterozygosity rate was greater than two standard deviations from the mean. Principle component analysis (PCA) was carried out using EIGENSTRAT, with additional individuals of Central European descent from the International HapMap Project (release 23) [260, 327]. Samples of non-European ancestry were then identified. Individuals were excluded whose score for any of the first 10 principle components was greater than 6 standard deviations from the mean.

Pedigree information was provided for samples from the LOAD cohort. Identity By Descent (IBD) analysis was carried out in the OPDC cohort to quantify relatedness between individuals from their genotypes. SNPs in linkage disequilibrium (LD) were removed using the PLINK *indep-pairwise* function, using a sliding window of 50 SNPs which shifted by 5 at each step. This created a

Table 4.1: summary of quality control procedures and samples excluded at each step for the OPDC and LOAD cohorts

	Number samples excluded	
Quality control step	**OPDC**	**LOAD**
Discordant sex	7	37
Heterozygosity	23	140
Missing data rate	2	0
Divergent ancestry	11	25

set of SNPs inherited approximately independently. For every pair of individuals the proportion of alleles inherited from a common ancestor was quantified. Relatives up to second degree were then inferred.

After all quality control metrics had been implemented the two datasets were merged and SNPs in LD removed as before. EIGENSTRAT was used to perform PCA to ensure the two populations were not genetically divergent. This also identified any underlying genetic artefacts, which were then controlled for during further analysis.

4.2.3) Genotype imputation

Imputation of unobserved and missing variants was carried out separately for each dataset. A reference panel containing comprehensive SNP data was used to identify extended patterns of LD and co-inherited alleles. The samples first underwent phasing to identify which alleles originated from the paternal and maternal chromosomes. Next, patterns of LD were applied to the phased samples to infer the likelihood of each possible genotype at unknown SNPs.

Reformatting and sorting of data was carried out using VCFtools [261]. Only individuals and SNPs passing all quality control stages were used. The Michigan Imputation Server was used to phase and impute data in both cohorts separately using Eagle and Minimac3 respectively [262, 264]. The 1000 Genomes project (phase 3, release 5) contains data for over 80million variants in 503 individuals of European descent and provided the reference panel [263].

For each SNP an r^2 value was produced reflecting the accuracy of imputation at that locus. Variants were filtered for r^2 greater than 0.3 to ensure

that only well-imputed SNPs were used in further analysis. Only variants with a minor allele frequency above 0.01 in both cohorts were used in association testing, as rare alleles were difficult to impute reliably and could have been prone to false positive associations in small samples.

4.2.4) Case-control GWAS

SNPtest v2.5.1 [328] was used to perform case-control logistic regression for each variant under an additive model. These tests conditioned on age, sex and the first two principal components to account for population substructure. Individuals classified as "at-risk" or "RBD" in the OPDC cohort were not used because although they were asymptomatic at the most recent follow-up their risk of developing PD in future was relatively high. Genotype probabilities were used to take into consideration the uncertainty of imputed variants.

The OPDC cohort was drawn from a relatively confined geographical area compared with the LOAD cohort. Furthermore there was a strong case-control ascertainment bias. Both of these factors could have inflated association P values and increased the false positive discovery rate. Consequently an extra binary covariate was added that represented study of origin to better control population stratification.

4.2.5) Discrete phenotype GWAS

Phenotypic presentation and progression in PD is diverse. Some pathogenic genetic variants segregate with characteristic sets of phenotypes, yet the cause of this heterogeneity remains largely unknown. Disease-linked genetic variation may associate with different phenotypes but this would be masked during analysis of all PD cases simultaneously. Consequently analysis of phenotypic

subgroups was performed in order to identify variants linked to the onset of particular phenotypes.

Patients in the OPDC cohort were assigned to one of five phenotypic sub-groups based predominantly on psychological well-being, cognitive impairment and non-tremor motor features such as posture and rigidity, detailed in Table 4.2 [248]. This system was developed by Michael Lawton and Yoav Ben-Shlomo, who also provided the classifications for each individual patient. In total five GWAS were carried out. For each phenotypic group patients within that group were classified as 'case' and patients in the remaining four groups were classified as 'control'. Individuals who were not assigned to any group were excluded from the analysis altogether.

Logistic regression was carried out for each subgroup under an additive model, conditioning on age, gender and the first two principle components. As all patients in this analysis were from the OPDC cohort the additional covariate encoding study of origin was not required. Only variants with minor allele frequency above 0.05 were used in this analysis. The small size of case and control groups was not sufficiently powered to detect association of rare variants and would have been prone to false positive associations.

Table 4.2: The defining characteristics of each phenotype cluster identified by Lawton *et al.*

Phenotype cluster	Description
1	Mild motor and non-motor disease
2	Poor posture and cognition
3	Severe tremor
4	Poor psychological wellbeing, RBD and sleep
5	Severe motor and non-motor disease with poor psychological wellbeing

4.2.6) Enrichment analysis

Multiple testing corrections employed within GWAS ensure few false positive associations but at the cost of a high false negative rate. Consequently variants that only impart a small effect or one mediated by other variants are likely to remain undetected, especially in small samples. Meta-Analysis Gene-set Enrichment of variaNT Associations (MAGENTA) provided a gene set enrichment analysis function which tested for over-represented groups of functionally similar genes among sub-threshold SNPs [329].

SNPs were assigned to genes based on their physical proximity. Variants were mapped to genes for which they were within 110kb upstream of the gene's most extreme transcript start site or 40kb downstream of the gene's most extreme transcript end site. This ensured that 99% of cis-eQTLs were mapped to the relevant gene, capturing signals from regulatory and promoter regions in addition to coding regions [329, 330].

For each gene an association score was calculated from the most significantly associated 5^{th} and 25^{th} percentile of SNPs, whilst controlling for confounding effects such as gene length. Sets of functionally linked genes were then tested for over-representation among those genes with the most significant association scores. Only the most strongly associated gene was used from all of those assigned to the same best SNP to prevent false associations arising from physical clustering of functionally similar genes.

This was carried out for the case-control analysis and independently for the analysis of each phenotype cluster. Gene sets were defined by Gene Ontology

annotations [265]. A false discovery rate (FDR) correction was applied to account for multiple testing.

4.3) Results

Quality control was first carried out for all individuals in the OPDC cohort. IBD analysis identified a total of 130 related individuals within 57 families. This included 8 parent-offspring pairs, 45 sibling groups and a number of more extensive pedigrees. Most contained at least one PD case individual and one or more individuals classed as at-risk.

Figure 4.3 shows the PCA plot produced to identify individuals of divergent ancestry. HapMap individuals clustered together some distance away from the main group of OPDC samples, which was likely to be due to the small geographical area from which the OPDC samples were collected. Distance from the HapMap reference sample was therefore not a good indicator of divergent ancestry in this case. Instead outliers were identified using PCA on the OPDC cohort only.

Figure 4.4 shows the first two principle components for just this cohort. This identified several distinct clusters separated from the main group. On further examination a number of these were formed of large families. All related individuals clustered closely to each other in this analysis, although smaller families were largely indistinguishable from the main group of unrelated individuals. It was likely that the genetic similarity between individuals in large families, rather than divergent ancestry, was driving their separation from the

Figure 4.3: the first two principle components generated from PCA of the OPDC cohort with CEU Hapmap. The HapMap population clusters together away from the main body of OPDC samples, indicating it is not a good measure of divergent ancestry in this case.

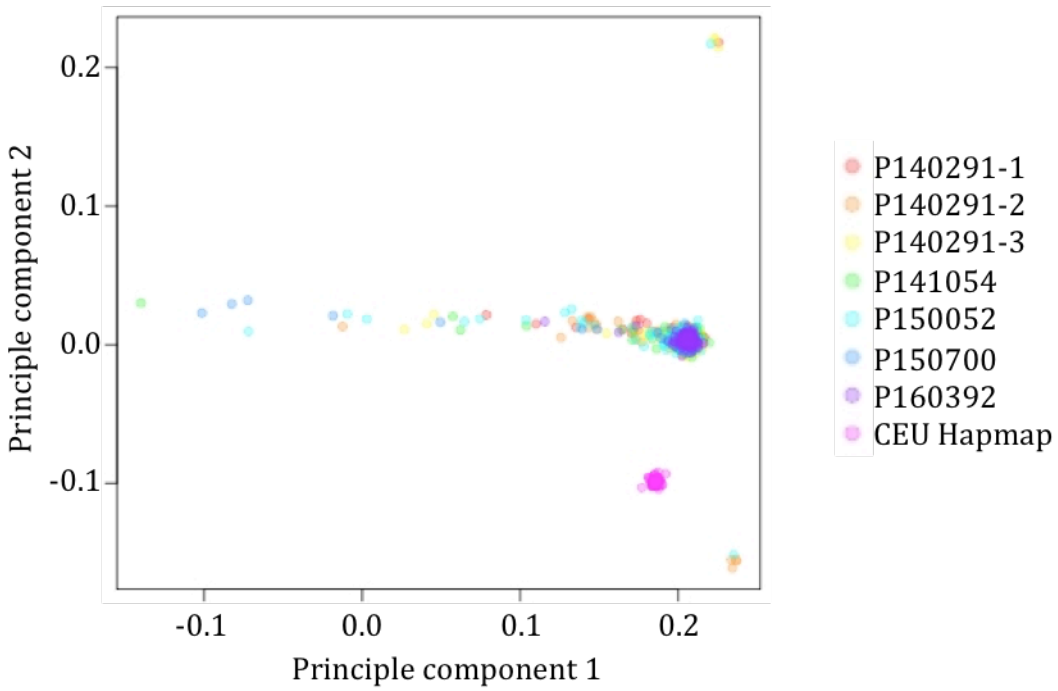
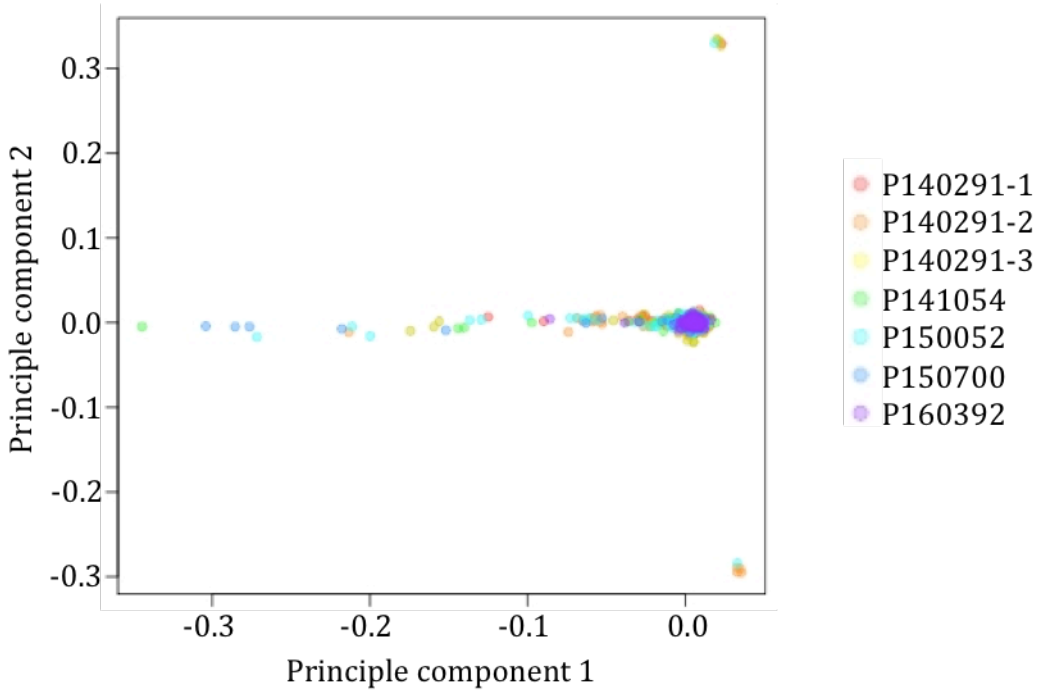


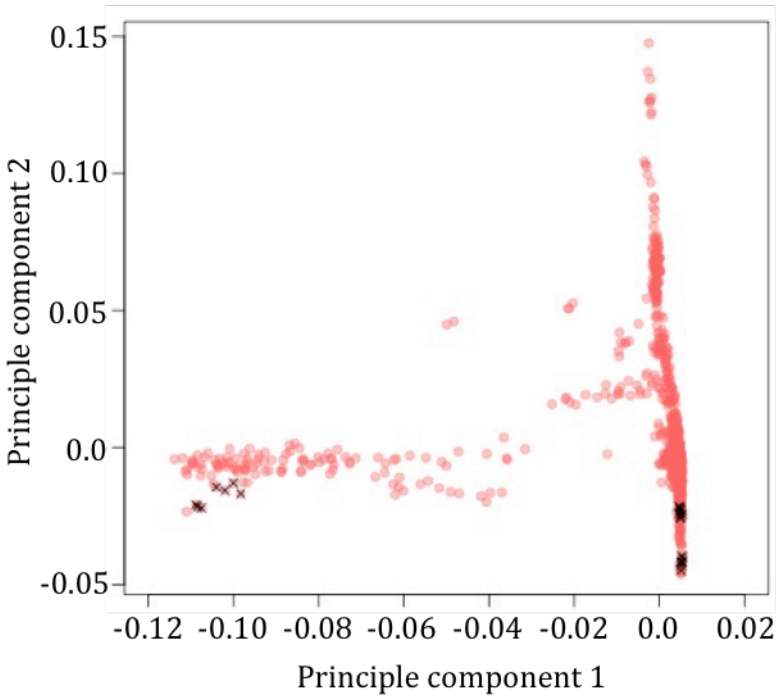
Figure 4.4: the first two principle components from PCA of the OPDC cohort. Individuals are coloured by genotyping run, showing that no batch effects were present.



main group. Consequently these individuals were exempt from PCA-based exclusion criteria and ethnicity was confirmed from clinical data. Figure 4.4 also demonstrates that no batch effects were present despite different versions of the SNP array having been used at different time points.

Quality control was then carried out for the LOAD cohort independently. The procedure was identical to that employed for the OPDC cohort. Divergent ancestry was again identified using deviance from the study population rather than by comparing with HapMap individuals. 25 samples were excluded on this basis and these individuals are shown in Figure 4.5. This figure also highlights the presence of genetic gradients within this population and underlines the importance of controlling for these factors during analyses.

Figure 4.5: the first two principle components from PCA carried out on the LOAD cohort shows genetic gradients were present in the population. Black crosses represent individuals excluded for extreme values of any of the first ten principle components.



4.3.1) Case-control analysis

Once all quality control steps were carried out 882 cases and 265 controls remained from the OPDC cohort and 2813 controls remained from the LOAD cohort. Both datasets were merged and PCA was carried out. This showed that the two datasets overlapped in the principle component space and were therefore genetically comparable (Figure 4.6). Systematic genetic gradients remained and this procedure provided covariates with which to control for this in further analysis.

Gender and age demographics are summarised in Table 4.7. The proportion of males was notably higher in the OPDC cohort. This was likely due to the increased incidence of PD in men given all cases originated from this dataset. Mean age was above the typical age of onset in both cohorts. Consequently the probability of a control individual developing PD at a later date was low.

After filtering on r^2 value and minor allele frequency, over 9 million SNPs were used for analysis. Logistic regression was carried out under an additive model comparing PD case with control. The QQ plot in Figure 4.8 shows that the included covariates adequately accounted for population substructure. It also demonstrates that there was inadequate statistical power to detect the most strongly associated variants.

The Manhattan plot in Figure 4.9 shows that no variants reached genome-wide significance. Known PD risk SNPs were then examined individually. This analysis replicated the association of five SNPs linked to MCCC1, SNCA, GPNMB, CCDC62 and MAPT (Table 4.10).

Figure 4.6: the first two principle components of PCA performed on the OPDC and LOAD cohorts combined. The two populations overlap and were therefore ancestrally similar and genetically comparable.

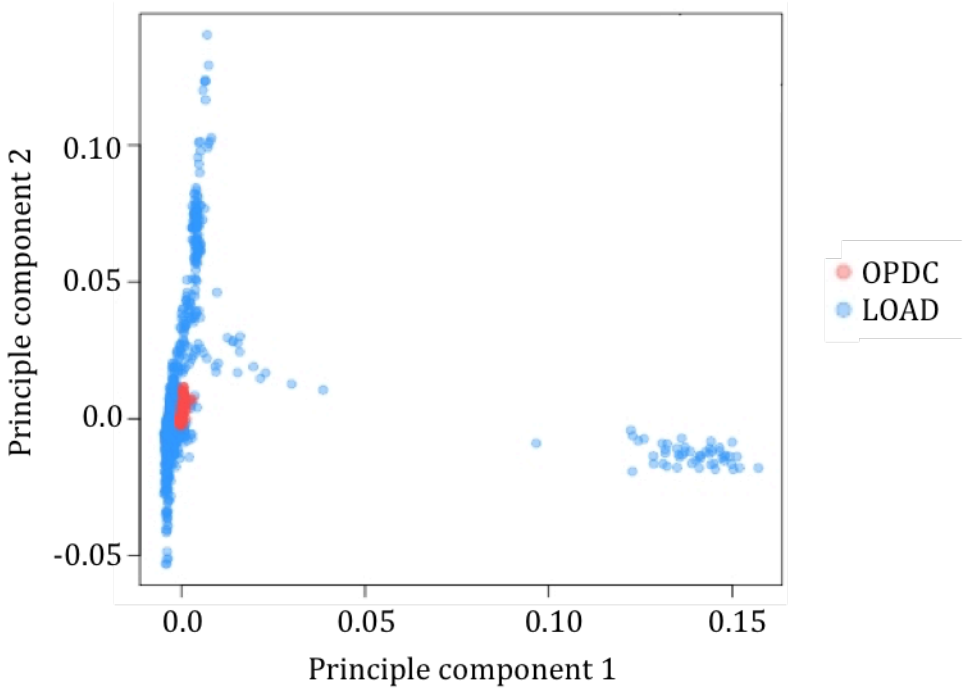


Table 4.7: summary of age and gender variables for the OPDC and LOAD cohorts.

	OPDC	**LOAD**
Mean age (standard deviation)	66.7 (9.75)	82.2 (9.98)
Percentage male	61.7%	36.8%

Figure 4.8: Quantile-Quantile plot of P values from the PD case-control GWAS. This shows that population stratification was adequately controlled, however there was insufficient statistical power to detect genome-wide significant variants.

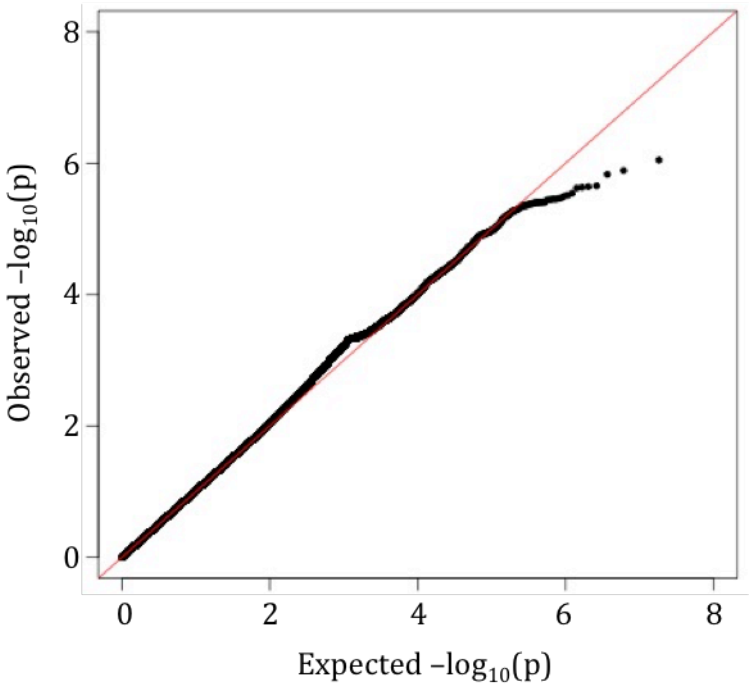


Figure 4.9: Manhattan plot from the GWAS comparing PD case with control shows that no variants were genome-wide significant.

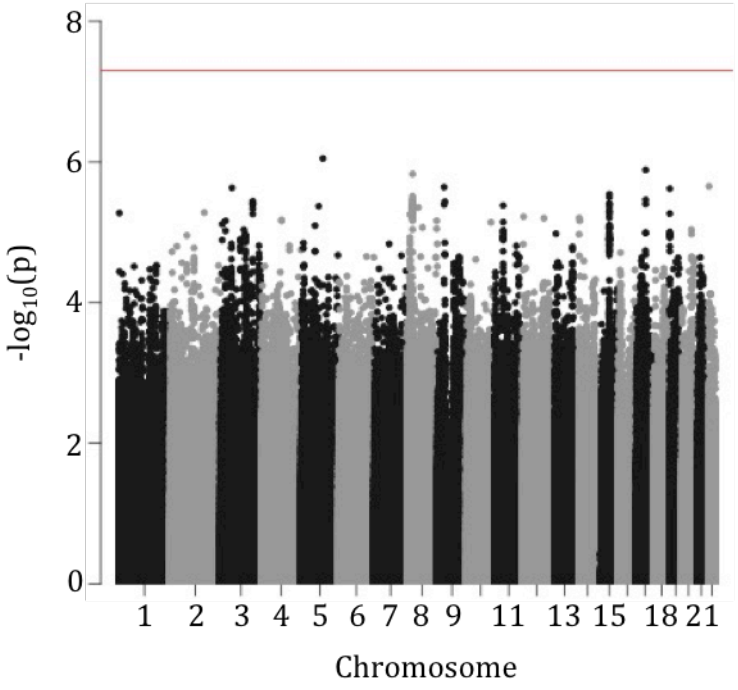


Table 4.10: Association P values of known PD risk SNPs show that five previous associations were replicated in this analysis, and that their direction of effect is consistent with previous study (represented by asterisk).

SNP	Associated gene	P value	Beta
rs6430538	ACMSD-TMEM163	0.805	0.0262
rs14235	BCKDK-STX1B	0.952	0.00662
rs11724635	BST1	0.0639	0.189
rs11060180	CCDC62	**0.00806**	**-0.409***
rs8118008	DDRGK1	0.106	-0.17
rs3793947	DLG2	0.889	0.0156
rs6812193	FAM47E-SCARB2	0.115	-0.164
rs591323	FGF20	0.0667	-0.215
rs35749011	GBA-SYT11	0.2	0.486
rs11158026	GCH1	0.845	-0.0223
rs199347	GPNMB	**0.0222**	**-0.235***
rs9275326	HLA-DQB1	0.598	-0.0922
rs117896735	INPP5F	0.0681	1.07
rs7077361	ITGA8	0.252	-0.166
rs115185635	KRT8P25-APOOP2	0.611	0.182
rs76904798	LRRK2	0.352	0.129
rs17649553	MAPT	**0.000457**	**-0.429***
rs12637471	MCCC1	**0.0213**	**-0.300185***
rs329648	MIR4697	0.222	-0.129
rs60298754	MMP16	0.34	0.342
rs34016896	NMD3	0.0663	0.206
rs823118	RAB7L1-NUCKS1	0.391	0.0904
rs12456492	RIT2	0.116	0.171
rs10797576	SIPA1L2	0.967	-0.00683
rs356182	SNCA	**0.0349**	**-0.236***
rs62120679	SPPL2B	0.123	0.19
rs11868035	SREBF1-RAI1	0.631	-0.0534
rs1474055	STK39	0.0863	0.267
rs34311866	TMEM175-GAK-DGKQ	0.596	0.0709
rs1555399	TMEM229B	0.189	0.132
rs2823357	USP25	0.747	0.0337
rs2414739	VPS13C	0.352	0.107

Gene set enrichment analysis was carried out using MAGENTA to identify common functional categories among the most strongly associated SNPs. Two Gene Ontology annotations were significantly associated with disease onset after FDR multiple testing corrections were applied (Table 4.11). Both were over-represented among the upper 25th percentile of associated SNPs, the threshold recommended for highly polygenic diseases.

Nuclear heterochromatin is linked to relatively few genes, but was the most strongly associated Gene Ontology term. Among genes linked to the highest 25th percentile of disease-associated SNPs three were predicted to possess this annotation under a null distribution. This was exceeded three-fold giving an FDR corrected P value of 0.026. The most significant 5th percentile of SNPs was expected to associate with only one gene with this annotation. Two were observed, but this was not a statistically significant enrichment (uncorrected P=0.107).

Nucleotide binding is a larger Gene Ontology term that describes over 1500 genes. Of these 406 were expected among genes linked to the top 25th percentile of disease-associated SNPs under a null distribution. In total 468 were observed, exceeding expectation 1.15-fold. This term was also 1.31-fold over-represented among the upper 5th percentile of associated SNPs. This was initially statistically significant (p= 0.000293) but did not surpass multiple testing corrections.

Table 4.11: significant results from enrichment analysis of Gene Ontology annotations among SNPs associated with PD onset. This was carried out for the most highly associated 5th (A) and 25th (B) percentile of SNPs using MAGENTA.

(A)

GO term	Expected genes	Observed genes	P value (uncorrected)	P value (FDR)[2]
Nuclear heterochromatin	1	2	0.107	0.828
Nucleotide binding	81	106	0.000293	0.547

(B)

GO term	Expected genes	Observed genes	P value (uncorrected)	P value (FDR)[2]
Nuclear heterochromatin	3	9	0.0002	0.0296
Nucleotide binding	406	468	0.000004	0.0422

[2] Also known as Q-value

4.3.2) Phenotypic analysis

Logistic regression was then carried out separately for each of the five phenotypic subgroups within the OPDC cohort. Table 4.12 shows the number of individuals that were in each subgroup and describes briefly the characteristic phenotypes of each. The QQ plots in Figure 4.13 demonstrate that population substructure was generally well controlled.

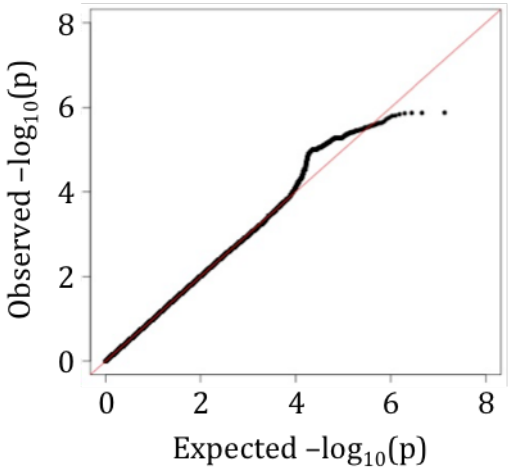
No variants achieved genome-wide significance, however a number demonstrated suggestive association ($p<5x10^{-7}$) (Figure 4.14). Most interestingly two peaks, each consisting of several variants, were associated with cluster 5. Individuals within this group displayed the most severe phenotypes overall but particularly in psychological and motor domains. The largest peak on chromosome 15 consisted of 25 suggestively associated variants upstream of ST8SIA2 (Figure 4.15). The other peak was located on chromosome 12 between RAB3IP and MYRFL genes (Figure 4.16). Additionally two isolated variants were significant at this level: rs539097805 was associated with cluster 5 and rs11276363 was associated with cluster 3.

Table 4.12: summary of characteristic phenotypes and the number of case and control individuals analysed for each phenotypic group

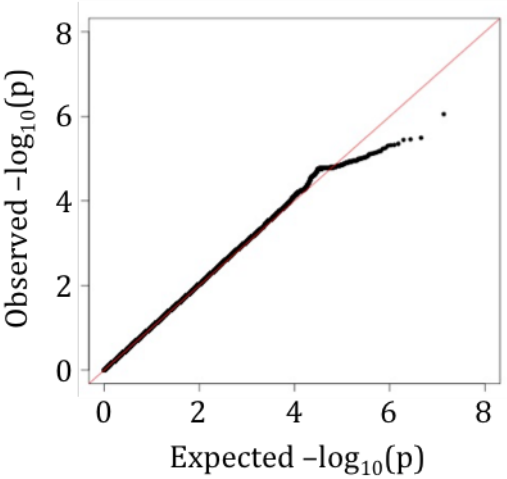
Phenotypic group	Number cases	Number controls	Description
1	176	512	Mild motor and non-motor
2	153	535	Poor posture and cognition
3	149	539	Severe tremor
4	131	557	Poor psychological well-being, RBD and sleep
5	79	609	Severe motor and non-motor disease with poor psychological well-being

Figure 4.13: Quantile-Quantile plots for GWAS performed on each phenotypic subgroup show that population substructure is adequately controlled.

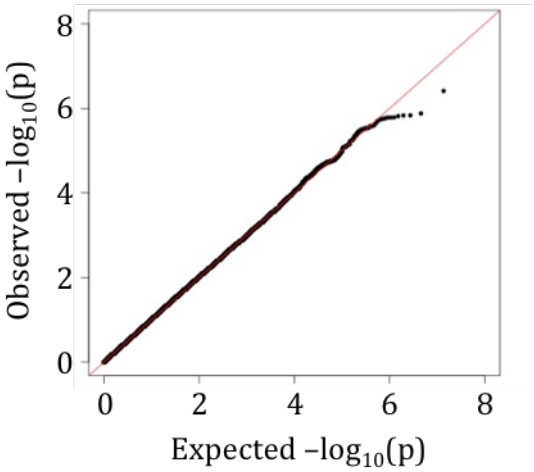
(A) K-means cluster 1



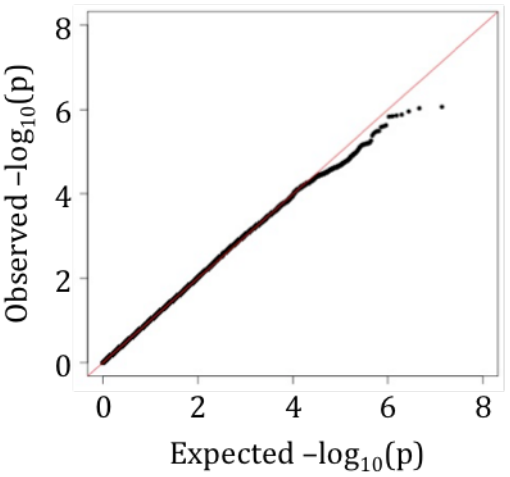
(B) K-means cluster 2



(C) K-means cluster 3



(D) K-means cluster 4



(E) K-means cluster 5

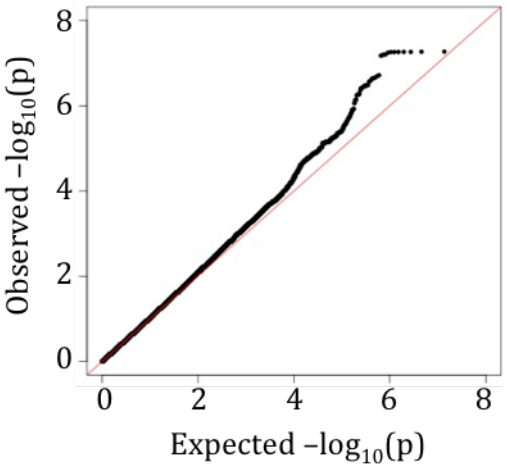
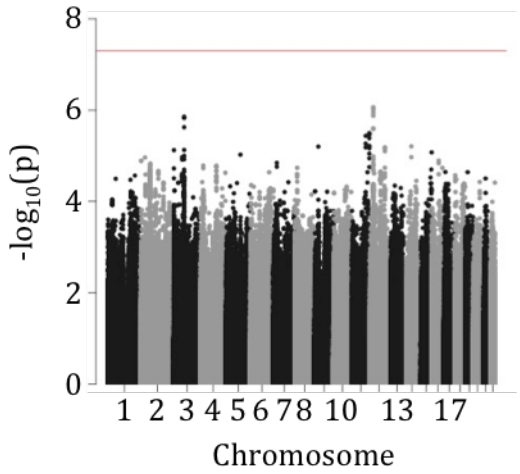
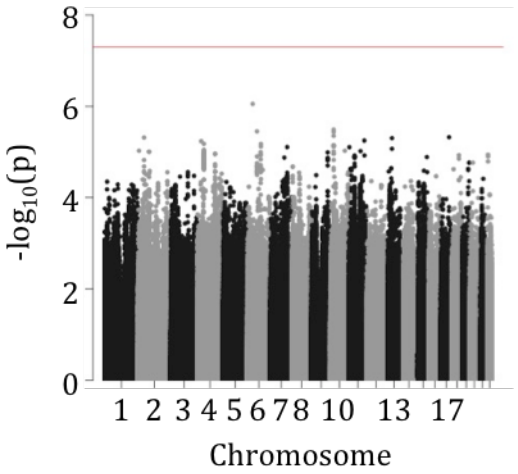


Figure 4.14: Manhattan plots for GWAS performed on each of the five phenotypic clusters (A-E).

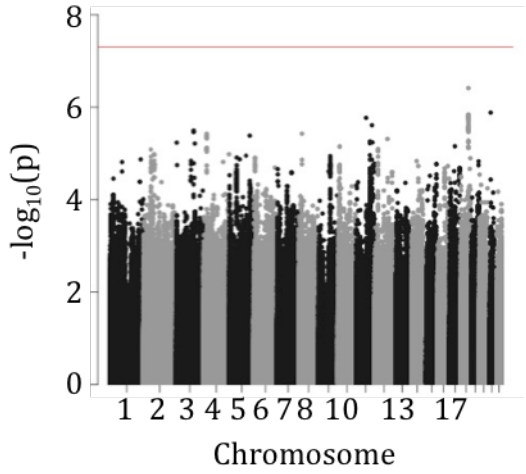
(A) K-means cluster 1



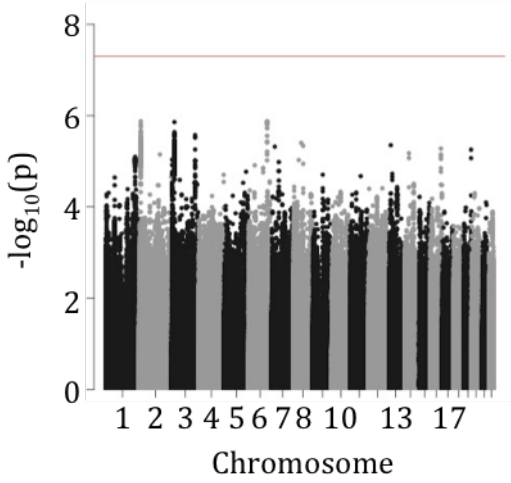
(B) K-means cluster 2



(C) K-means cluster 3



(D) K-means cluster 4



(E) K-means cluster 5

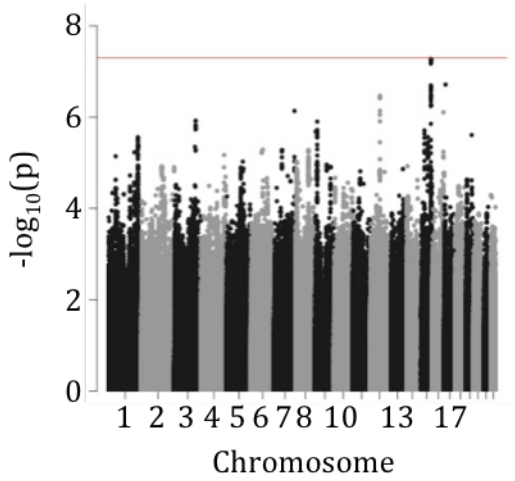


Figure 4.15: Visualized association peak of SNPs on chromosome 15 associated with phenotype cluster five

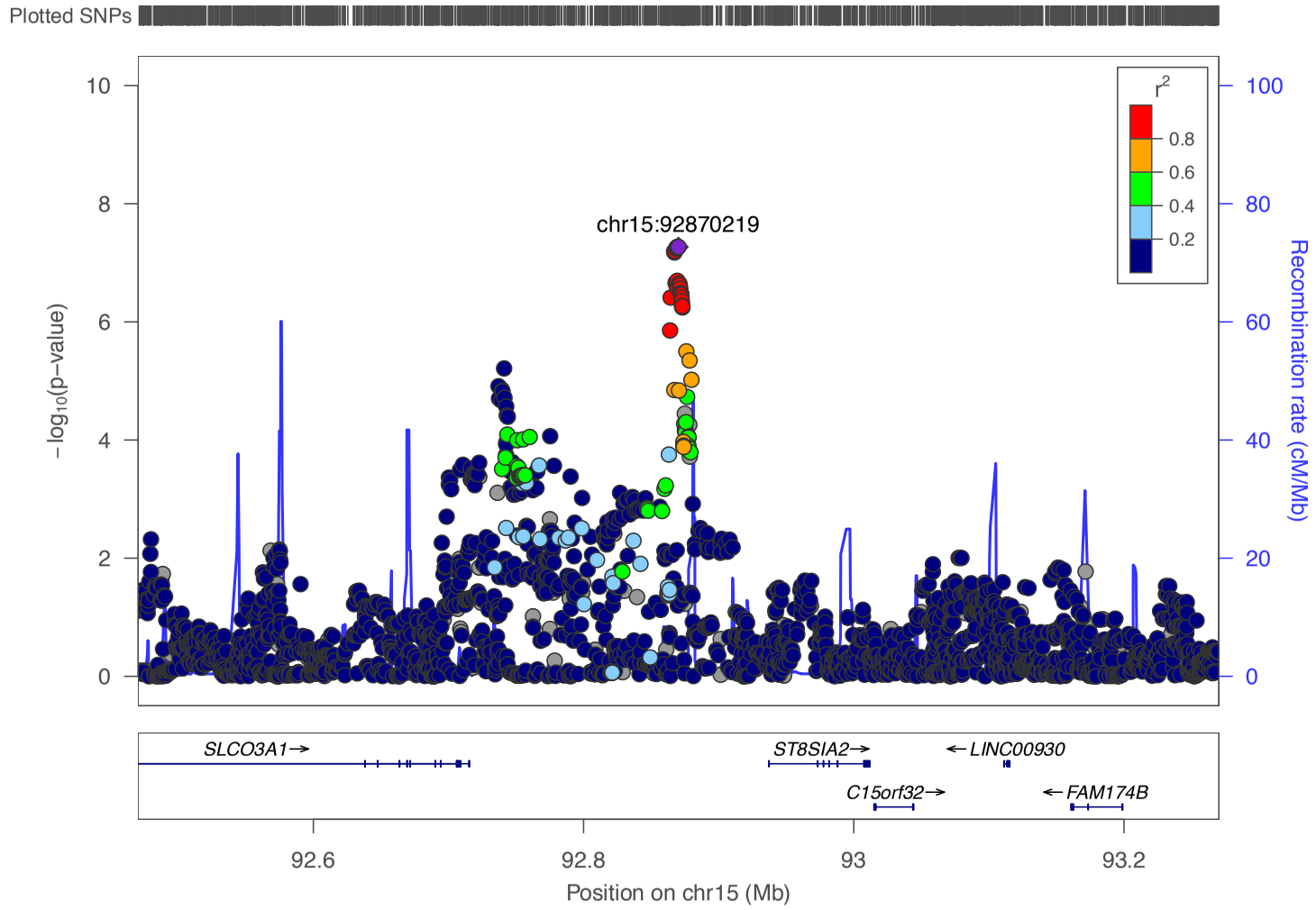
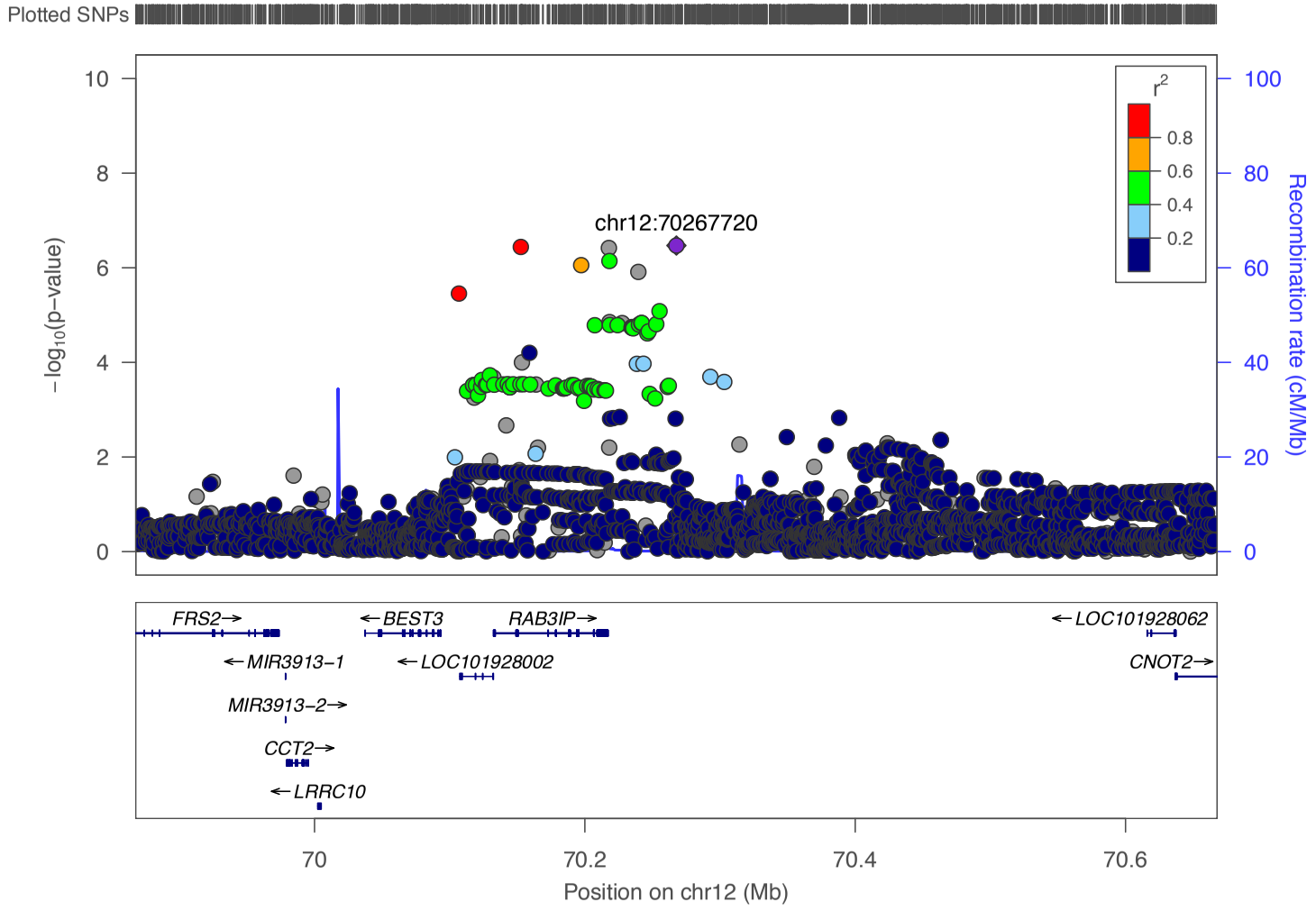


Figure 4.16: Visualized association peak of SNPs on chromosome 12 associated with phenotype cluster five



Enrichment analysis was performed independently for each phenotypic subgroup to identify over-represented functional gene classes linked to particular phenotypes. Only cluster 5 was significantly associated with any Gene Ontology annotation after FDR corrections were applied (Table 4.17). Genes annotated with *guanyl-nucleotide exchange factor activity* were almost two-fold enriched among those linked to the upper 25th percentile of SNPs. This represented an increase of 23 genes over expectation. Similar fold-enrichment was also observed among genes linked to the upper 5th percentile, although this was only nominally significant.

Table 4.17: significant results from pathway analysis among SNPs in the most highly associated 5th (A) and 25th (B) percentiles with phenotype cluster 5.

(A)

GO term	Expected genes	Observed genes	P value (uncorrected)	P value (FDR)
Guanyl-nucleotide exchange factor activity	6	10	0.0471	0.693

(B)

GO term	Expected genes	Observed genes	P value (uncorrected)	P value (FDR)
Guanyl-nucleotide exchange factor activity	28	51	0.00000099	0.0017

4.4) Discussion

In this analysis the GWAS framework was used to carry out genetic comparison between case and control and between different phenotypic subgroups within the case population. Enrichment analysis was performed on these results to identify functionally linked sets of genes over-represented among disease- or phenotype-associated SNPs. Two Gene Ontology terms were significantly associated with disease onset and one with a phenotypic subgroup. This shows that some genetic variants converge on common molecular pathways involved in PD onset. Furthermore results suggest that phenotype presentation may be influenced by genetic factors.

4.4.1) Imputation

The majority of exome variants typed on the array were mono-allelic, consequently only 300,000 of the 500,000 directly observed loci were informative in the OPDC cohort. Imputation of unknown variants provided information for an additional 9 million SNPs after filtering. Coverage of the genome was therefore increased 30-fold, allowing more variants to be tested for association.

Use of the Michigan Imputation Server was far superior to performing the analysis in-house. It was significantly faster, saving both personal and computational time. Ease of use was good and improved over that of the phasing and imputing software independently.

4.4.2) Case-control analysis

GWAS have identified over 30 PD risk SNPs, however few examine any functional links between them that might highlight disease-associated pathways.

In this study cases were compared with aged controls to identify SNPs associated with disease onset. Functional enrichment was then carried out among the most strongly associated variants. In this way two functional gene annotations were significantly associated with PD onset.

In the case-control analysis no variants were genome-wide significant. This was an expected result in a study of this size. For complex disease several thousand each of case and control individuals are generally required to achieve the necessary statistical power. This study was less than $1/10^{th}$ of the size of the most recent PD GWAS.

Five previously identified associations were replicated. In previous studies MAPT and SNCA have been the most strongly linked to PD and both genes were associated with disease onset in this study. Association of MCCC1, GPNMB and CCDC62 was also nominally significant. Replication of known risk variants demonstrated that this cohort was genetically representative of the general PD population.

GWAS have considerably aided the understanding of PD pathology, however are limited by testing each variant independently of all others. Consequently there is no way to identify additive or complex effects of interacting SNPs, nor the common biological pathways in which they act. Cellular pathways provide a more comprehensive understanding of the disease process and are easier to target pharmacologically than are individual variants. To address this MAGENTA was used to identify functional gene classes enriched for PD-linked genetic mutation.

The most strongly associated annotation was *Nuclear heterochromatin.* 3-fold enrichment was observed within the upper 25th percentile of disease-associated SNPs. Loss of heterochromatin markers is associated with the aging process and its acceleration is linked to other neurodegenerative disorders [331-333]. This can alter gene transcription, resulting in expression of genes that should remain silenced [333]. Genetic susceptibility to heterochromatin changes could therefore exacerbate the cellular phenotype associated with normal aging, causing neurodegeneration. Heterochromatin relaxation can also occur as a result of oxidative stress, so may also interact with other aspects of PD pathology.

Nucleotide binding was also significantly associated with PD onset. This annotation was enriched among the most highly associated 25th and 5th percentiles of disease-related SNPs. This term includes genes whose proteins interact selectively and non-covalently with any nucleotide, such as adenyl and guanyl nucleotides and nicotinamide adenine dinucleotides (NAD). Consequently a relatively large number of genes carry this annotation

Binding of specific nucleotides has previously been implicated in PD onset. Several pathogenic mutations in the LRRK2 gene increase GTP binding but not GTPase activity [334]. ATP-binding genes are also associated with disease onset and to changes arising from treatment [335]. Although nucleotides are diverse in structure and function this analysis demonstrated a convergent effect on PD onset. This diversity could however reflect differences in pathological mechanisms or phenotype presentation. The stratification of individuals by

disease sub-type was required to elucidate interactions of this kind and was explored in the next stage of analysis.

4.4.3) K-means cluster analysis

Phenotypic differences are widely observed among PD patients. However the molecular perturbations underlying these differences, and whether they can be attributed to genetic or environmental causes, remains largely unknown. Certain risk genes have been linked to characteristic patterns of disease progression, implying at least some genetic component. Consequently different phenotypic subgroups of patients were compared in the hope of identifying genetic factors influencing phenotypic presentation.

The most commonly used subtypes of PD split individuals into three groups: postural-instability gait disorder (PIGD), tremor-dominant (TD) and indeterminate. Previous analysis of the OPDC cohort elucidated five groups, two of which were similar to the PIGD and TD subtypes [248]. This novel classification system therefore seemed to support previous results whilst granting better resolution over the remaining "indeterminate" individuals. Consequently these five phenotypic clusters were used for genetic analysis.

Each subgroup was examined in turn, comparing individuals within the cluster ("cases") to those in other clusters ("controls"). The number of case and control samples was significantly decreased using this method. This reduced statistical power and was the main limitation of this analysis. However specificity of the case population was increased and it was hoped that this would overcome the reduction in total numbers.

No genome-wide associations were observed with any phenotypic cluster, however a number of suggestively associated variants were identified ($p<5x10^{-7}$). Cluster 3, defined by severe tremor, was associated with a single insertion variant. Rs11276363 lies in an intron of CCBE1, a gene encoding a collagen and calcium binding protein with an EGF domain. This protein regulates VEGFC, a neurotrophic and neuroprotective factor for dopamine neurons [336, 337]. The onset and severity of tremor phenotypes is correlated with dopamine loss in specific brain regions [338]. Alterations of CCBE1 could therefore cause dopamine depletion through dysregulation of neuroprotective factors and contribute to the onset of severe tremor in these patients.

The remaining significant variants were associated with cluster 5. Individuals in this group were characterized by severe motor and non-motor symptoms with poor psychological wellbeing. Two independent association peaks were identified in addition to a single isolated SNP.

The first peak was located downstream of the RAB3IP gene on chromosome 12. RAB3IP interacts with RAB3A, a protein highly expressed in neurons where it reduces alpha-synuclein binding to membranes [339, 340]. PD-linked mutation in the SNCA gene is associated with similar characteristic phenotypes to this cluster [104, 105]. This could indicate a common role of alpha-synuclein function in the development of severe motor and cognitive phenotypes.

The largest peak was situated on chromosome 15 at the upstream end of ST8SIA2, a gene previously linked to schizophrenia and bipolar disorder [341]. It encodes a type II membrane protein involved in the production of polysialic acid

[342]. This compound is thought to have roles in emotion, learning, memory, behavior and circadian rhythm [343]. Alterations in ST8SIA2 may therefore influence the severe cognitive and psychiatric phenotypes displayed by patients in this cluster through modulation of polysialic acid production.

The single associated SNP also linked with phenotype cluster 5 is an intron variant of NTN1. This gene is expressed in striatal and nigral neurons throughout life and may be involved in the development and maintenance of connectivity to the substantia nigra and corpus striatum [344]. Deficiency of NTN1 or its receptors is associated with reduced dopamine levels in the medial prefrontal cortex [345]. This brain region is linked to memory, therefore mutations affecting NTN1 function could influence the onset of the severe cognitive phenotypes associated with this phenotypic cluster. Its receptors have also been implicated in SNP models predicting PD outcome, indicating that other genes acting within the same pathway might also have an effect on disease progression [346, 347].

In addition to having the most SNP associations, cluster 5 was the only phenotypic subtype to be significantly enriched for a functional gene annotation. *Guanyl-nucleotide exchange factor activity* annotated genes were almost two-fold enriched among the SNPs most strongly associated with this group. This term includes any gene whose product stimulates the binding of guanyl nucleotides associated with a GTPase, but not those that affect the catalytic activity directly.

GTPase proteins form five different classes, several of which have been linked to PD, and are involved in a number of diverse roles within the cell. Guanine nucleotide Exchange Factors (GEFs) promote the activation of small

GTPases by catalysing the release of GDP and binding of GTP. GTPases may respond to several GEFs, each of which responds to different upstream stimuli [348].

Rab GTPases regulate intracellular traffic and are involved in both synaptic vesicle fusion and endocytosis in neurons. They are also linked to PINK1, whose activity results in phosphorylation of certain Rabs. This impairs their activation by GEFs causing a corresponding reduction in overall GTPase activity [349]. In patient-derived PINK1 Q456X mutant fibroblasts phosphorylation is absent [349]. This is likely to result in altered GEF activation, causing aberrant activity of GTPases and other downstream proteins.

LRRK2 contains a Roc GTPase domain that is also regulated by GEFs, one of which is ARHGEF7. This protein influences neurite outgrowth and promotes the binding of GTP to wild type LRRK2 [350, 351]. However the R1441C mutant shows impaired interactions with this GEF, resulting in reduced GTP binding affinity [351]. The effect of GTP binding on LRRK2 toxicity directly remains unknown. However kinase activity is dependent on this function and is directly implicated in the toxicity of familial mutations [128, 352]. Consequently genetic variants that modify GEF interactions and therefore GTP binding could alter LRRK2 toxicity.

GEFs may alter known aspects of PD pathology through several diverse effectors. However the mechanisms that link GEF activity with the development of severe motor and non-motor phenotypes are unknown. Further study of this association would be beneficial, as this could provide the basis for symptomatic treatment targeted specifically to a disease subtype.

Gene Ontology annotations are hierarchical and the *Nucleotide binding* term implicated in the case-control analysis is a parent of the *Guanyl-nucleotide exchange factor activity* term associated with phenotype cluster 5. This demonstrates that similar molecular processes could underlie PD onset for all patients, but that perturbations of specific pathways within that could cause the development of particular phenotypes. This concept aligns with the phenotypic similarity observed between patients carrying known genetic mutations such as SNCA.

The association of molecular mechanisms with specific phenotypic subgroups indicates that therapeutic agents might be more efficacious among some patients than others. Given the limitations of population size on statistical power in this study it is likely that larger cohorts would elucidate many more mechanisms associated with each phenotypic cluster. Consequently stratification by disease subtype would be beneficial during clinical trial to account for the possibility of diverse biological mechanisms underlying different phenotype presentations.

4.4.4) Conclusion

Overall it has been shown that PD-associated SNPs converge on genes involved in common molecular pathways. Nuclear heterochromatin and nucleotide binding were linked to disease onset, both of which are factors known to influence neurodegeneration but are relatively unexplored in PD pathology. Furthermore guanine nucleotide exchange factor activity, a regulator of a specific type of nucleotide binding, was significantly associated with the phenotypic

subgroup of PD characterised by severe motor, non-motor and psychological phenotypes.

This demonstrates that common molecular mechanisms might underlie PD onset as a whole, but that genetic mutations affecting particular pathways within that could cause the onset of diverse phenotypes. Therefore different therapeutic strategies may be optimal for different patient subgroups. By targeting more selective molecular pathways it should be possible to maximise efficacy and minimise off-target effects of treatment. Phenotypic groups are inferred from clinical observations and questionnaire data and do not require genotype information. Consequently this provides a convenient way of stratifying patients for targeted treatment.

Chapter 5: Phenotype-Genotype Analysis of the OPDC Discovery Cohort

5.1) Introduction

Clinical heterogeneity is a known hallmark of Parkinson's Disease (PD): patients present with a diverse selection of neurological, motor and autonomic phenotypes, each of which also demonstrates variable severity. However the pathological mechanisms underlying these differences remain largely unknown. Previous research has suggested that certain PD genetic risk variants may influence phenotype presentation, but they explain only a small proportion of the observed variance.

In this work the first hypothesis-free genome-wide investigation into the effect of genotype on deep clinical phenotype was carried out. Axes of latent phenotypic variation were defined, each of which represented the severity of a number of covarying observed phenotypes from complete absence to most severe, capturing the entire spectrum of severity observed in the clinic. Logistic regression along these axes identified a number of strongly associated variants, each of which was associated with at least one measured phenotype. In total 10 genomic regions were significantly linked to the severity of clinical phenotypes in this analysis.

5.2) Methods

5.2.1) Data

Data from 843 PD cases from the OPDC Discovery Cohort was used in this analysis. Individuals were required to have at least 90% chance of PD according to UK-PD brain bank criteria, no alternative diagnosis and disease duration less than 3.5 years. Details of quality control and genotype imputation are discussed in chapter 3, section 2. Phenotype data was collected for over 50 attributes listed in Table 5.1, encompassing autonomic, neurological and motor phenotypes.

5.2.2) Phenotype imputation and generation of latent components

Michael Lawton compiled the phenotype data collected by the clinicians and performed quality control. Missing data rates were generally less than 5%. However if samples were excluded on the basis of a single missing data entry then the proportion of retained samples decreased rapidly. Consequently missing values were imputed using PHENIX [353], developed by Andy Dahl and Jonathan Marchini. Unlike most imputation methods taken from mainstream statistics, this programme is specifically designed to impute phenotypes and therefore outperforms several more general methods for this purpose.

PHENIX models phenotypes as a product of random effects and residuals within a multiple phenotype mixed model. It then exploits sample relatedness to further decompose random effects into kinship effects between individuals, genetic covariances between phenotypes and residual covariances between phenotypes. In this way correlations both between phenotypes and between individuals are identified to maximise the accuracy of the predictive model.

Table 5.1: Summary of phenotype data collected for individuals in the OPDC cohort, provided by Michael Lawton

Phenotype	Variable	Details
Subject ID	String	Unique patient identifier
Discovery	Categorical: {PD, Controls, At Risk}	The variable that denotes which group the individual belongs to. Note that the PD group has >= 90% probability of Parkinson's Disease
Gender	Binary: {Male, Female}	
Age	Continuous	Age at first assessment in years
Age onset	Continuous	Age at onset of PD in years
Disease duration	Continuous	Disease duration since onset in years
Disease duration diag	Continuous	Disease duration since diagnosis in years
Ethnicity	Binary: {white, non-white}	Due to small numbers this was dichotomised into white vs. non-white
BMI	Continuous	Body Mass index
BFI extra total	Ordinal: 8 to 40	Big Five inventory Extraversion. 8 questions rated from 1 to 5. Higher score corresponds to more extravert personality
BFI agree total	Ordinal: 9 to 45	Big Five inventory Agreeableness. 9 questions rated from 1 to 5. Higher score corresponds to more agreeable personality
BFI consci total	Ordinal: 9 to 45	Big Five inventory conscientiousness. 9 questions rated from 1 to 5. Higher score corresponds to more conscientious personality
BFI neuro total	Ordinal: 8 to 40	Big Five inventory neuroticism. 8 questions rated from 1 to 5. Higher score corresponds to more neurotic personality
BFI open total	Ordinal: 10 to 50	Big Five inventory Openness. 10 questions rated from 1 to 5. Higher score corresponds to more open personality
ESS total	Ordinal: 0 to 24	Epworth sleepiness scale. A measure of daytime sleepiness, higher score corresponds to more daytime sleepiness. 8 questions rated from 0 to 3

ESS bin	Binary: {0,1}	ESS total dichotomised at 11 or more. 0 = normal; 1=daytime sleepiness problem
RBD total	Ordinal: 0 to 13	REM Sleep behaviour disorder screening questionnaire. Higher score corresponds to greater REM sleep behaviour problems. 13 questions rated as 0 or 1.
RBD bin	Binary: {0,1}	RBD total dichotomised at 5 or more. 0 = normal; 1=REM sleep behaviour disorder
EQ5D index	Discrete	EuroQol health states. Quality of life index. 5 questions rated 1-3 and then transformed into index using a equation
EQ5D vas score	Continuous: bounded 0 to 100	Health quality visual analogue scale, 0 is worst health, 100 is perfect health
Constip cat	Categorical: {<1 or laxative use, 1, 2, >2}	Average daily bowel movements, merged with use of laxatives. Quantitative measure of constipation.
Leeds anxiety total	Ordinal: 0 to 18	Leeds anxiety scale. 6 questions rated 0-3 higher score corresponds with more anxiety
Leeds anxiety bin	Binary: {0,1}	Leeds anxiety total dichotomised at 7 or more. 0= normal; 1=clinically anxious
Leeds depression total	Ordinal: 0 to 18	Leeds depression scale. 6 questions rated 0-3 higher score corresponds with more depression
Leeds depression bin	Binary: {0,1}	Leeds depression total dichotomised at 7 or more. 0= normal; 1=clinically depressed
BDI total	Ordinal: 0 to 63	Becks Depression Inventory. 21 questions rated 0-3. Higher score corresponds with more depression
BDI cat	Categorical: {minimal depression, mild depression, moderate depression, severe depression}	BDI total categorised into. 0-13 minimal depression; 14-19 mild depression; 20 to 28 moderate depression; 29 to 63 severe depression
QUIP all	Binary: {0,1}	Questionnaire for Impulsive-

		Compulsive Disorders. 13 questions rated 0-1 split into 8 domains. This variable denotes the presence of at least one QUIP domain.
QUIP gamb	Binary: {0,1}	QUIP. 2 questions for gambling domain. This variable denotes the presence of at least one question in gambling domain
QUIP sex	Binary: {0,1}	QUIP. 2 questions for sex domain. This variable denotes the presence of at least one question in sex domain.
QUIP buy	Binary: {0,1}	QUIP. 2 questions for buying domain. This variable denotes the presence of at least one question in buying domain.
QUIP eat	Binary: {0,1}	QUIP. 2 questions for eating domain. This variable denotes the presence of at least one question in eating domain.
QUIP med	Binary: {0,1}	QUIP. 2 questions for medication domain. This variable denotes the presence of at least one question in medication domain. This is missing for any person not on PD medication.
QUIP hobby	Binary: {0,1}	QUIP. 1 question for hobby domain.
QUIP pund	Binary: {0,1}	QUIP. 1 question for punding domain.
QUIP walk	Binary: {0,1}	QUIP. 1 question for walkabout domain.
MOCA total	Ordinal: 0 to 30	Montreal Cognitive Assessment. A battery of different questions measuring cognitive ability. A higher score corresponds with better cognition.
MOCA total adj	Ordinal: 0 to 30	MOCA total adjusted for education years, +1 to score if education years <= 12 to a maximum of 30
MOCA cat screen	Categorical: {dementia, MCI, normal}	MOCA total using screening cut-points categorised into: 0-20 dementia; 21-25 mild cognitive impairment (MCI); 26-30 normal
MOCA cat diag	Categorical: {dementia, MCI, normal}	MOCA total using diagnostic cut-points categorised into: 0-21 dementia; 22-23 Mild cognitive

		impairment (MCI); 24-30 normal
MOCA cat screen adj	Categorical: {dementia, MCI, normal}	MOCA total adj using screening cut-points categorised as above
MOCA cat diag adj	Categorical: {dementia, MCI, normal}	MOCA total adj using diagnostic cut-points categorised as above
Education years	Ordinal	Number of years of education
MMSE total	Ordinal: 0 to 30	Mini-mental state examination. A battery of different questions measuring cognitive ability. A higher score corresponds with better cognition.
MMSE bin	Binary: {0,1}	MMSE total dichotomised at 23 or less. 0 = normal, 1 = dementia
Phen fluency score	Ordinal: 0 , no upper bound	Number of words beginning with a particular letter. 60 seconds for each of the letters: F, A and S. A higher score corresponds with better cognition.
Phen fluency adj score	Ordinal: 0 , no upper bound	Phen fluency score adjusted for age
Seman fluency score	Ordinal: 0, no upper bound	Number of animals and the number of boy's names. 60 seconds for each test. A higher score corresponds with better cognition.
Seman fluency adj score	Ordinal: 0, no upper bound	Seman fluency score adjusted for age
Purdue total	Ordinal: 0, no upper bound	Total of the first three Purdue pegboard test subtests. Test to measure manual dexterity. A higher score corresponds with better dexterity
Purdue assembly	Ordinal: 0, no upper bound	Total for the assembly part of the Purdue pegboard Test to measure manual dexterity. A higher score corresponds with better dexterity
Getgo average	Continuous	Get up and Go Test. Time (minutes) taken for an individual to get up from a chair, walk three meters, turn around, walk back to the chair and sit down. Average of three attempts. A higher score corresponds with worse motor function.
Getgo aban	Marker: {1}	1 = get go test abandoned or not done

Flamingo time	Continuous: 0 to 30 secs	Flamingo test. Time (seconds) that a person can stand on one leg. If the patient can do this for 30 seconds the test is stopped. A lower score corresponds with worse motor function.
Flamingo aban	Marker: {1}	1 = flamingo test abandoned or not done
SBP postural drop	Continuous	Systolic blood pressure postural drop. 2 measurements made whilst lying down and then another made after standing up. Mean of systolic BP lying - systolic BP standing
DBP postural drop	Continuous	Diastolic blood pressure postural drop. Mean of diastolic BP lying - diastolic BP standing
Pulse postural drop	Continuous	Pulse postural drop. Mean of lying pulse - standing pulse
Vascular cat	Categorical: {None, 1, 2 or more}	Number of vascular diseases
LEDD total	Continuous	Levodopa equivalent daily dose, a quantitative measure of the amount of PD medication a person is taking.
Sniffin total	Ordinal: 0 to 16	Sniffin sticks identification test. 16 smells to identify, from 16 multiple-choice questions with four possible answers. Higher score corresponds with better smell
Sniffin bin	Binary: {0,1}	Sniffin total dichotomised using published age and gender corrected normative values (10th centile or lower). 0 = normal smell, 1= hyposmia (diminished sense of smell)
Hoehn Yahr stage	Ordinal: 0 to 5	Hoehn and Yahr stage. Parkinson's disease severity. 0 is no disease; 5 is bedridden
Schwab England scale	Ordinal: 0 to 100	Modified Schwab & England Activities of Daily Living. Variable is measured from 0% to 100% in 10% increments. 0% is bedridden, 100% completely independent
UPDRS I	Ordinal: 0 to 52	Movement Disorder Society version of the Unified

		Parkinson's Disease Rating Scale (MDS-UPDRS) part I "Non Motor Aspects of Experiences of Daily Living. 13 questions rated from 0: normal to 4: severe. A higher score corresponds with worse non-motor problems
UPDRS II	Ordinal: 0 to 52	MDS UPDRS part II "Motor Aspects of Experience of Daily Living". 13 questions rated 0-4. A higher score corresponds with worse motor problems. Self-reported.
UPDRS III	Ordinal: 0 to 132	MDS UPDRS part III "Motor Examination". 33 questions rated 0-4. A higher score corresponds with worse motor problems. Clinician assessed.
UPDRS IV	Ordinal: 0 to 24	MDS UPDRS part IV "Motor Complications". 6 questions rated 0-4. A higher score corresponds with worse motor complications.
UPDRS apathy	Ordinal: 0 to 4	MDS UPDRS part I apathy question rated 0-4. A higher score corresponds with greater apathy.
UPDRS hallucinations	Ordinal: 0 to 4	MDS UPDRS part I hallucination question rated 0-4. A higher score corresponds with more hallucinations.
UPDRS fatigue	Ordinal: 0 to 4	MDS UPDRS part I fatigue question rated 0-4. A higher score corresponds with more fatigue.
UPDRS pain	Ordinal: 0 to 4	MDS UPDRS part I pain question rated 0-4. A higher score corresponds with more pain.
UPDRS constipation	Ordinal: 0 to 4	MDS UPDRS part I constipation question rated 0-4. A higher score corresponds with more constipation.
UPDRS rigidity	Ordinal: 0 to 20	MDS UPDRS part III, all 5 rigidity questions rated 0-4. A higher score corresponds with more rigidity
UPDRS bradykinesia	Ordinal: 0 to 48	MDS UPDRS part III, all 12 bradykinesia questions rated 0-4. A higher score corresponds with more bradykinesia.

UPDRS postural	Ordinal: 0 to 20	MDS UPDRS part III, all 5 postural questions rated 0-4. A higher score corresponds with worse posture.
UPDRS tremor	Ordinal: 0 to 40	MDS UPDRS part III, all 10 tremor questions rated 0-4. A higher score corresponds with more tremor.
UPDRS speech	Ordinal: 0 to 4	MDS UPDRS part III, 1 question rated 0-4. A higher score corresponds with more speech problems.
UPDRS faceneck	Ordinal: 0 to 16	MDS UPDRS part III, 4 faceneck questions rated 0-4. A higher score corresponds with more face/neck problems
UPDRS arms	Ordinal: 0 to 56	MDS UPDRS part III, 14 arms questions rated 0-4. A higher score corresponds with more arm problems
UPDRS legs	Ordinal: 0 to 32	MDS UPDRS part III, 8 legs questions rated 0-4. A higher score corresponds with more leg problems
UPDRS laterality	Ordinal: 0 to 44	MDS UPDRS part III 11 questions on each side. Absolute difference between left and right side question. Higher score corresponds with more unilateral disease.
UPDRS laterality bin	Binary: {0,1}	UPDRS laterality dichotomised at four or more. 0 = two-sided disease; 1 = one-sided disease.
UPDRS phenotype cat	Categorical: {TD, indeterminate, PIGD}	MDS UPDRS. Common phenotype. Mixture of part II and part III questions transformed into an equation to classify an individual as: TD = tremor dominant; PIGD = Postural Instability Gait Disorder; or indeterminate.

Inputs to this programme consisted of a matrix of phenotype values and a kinship matrix reflecting sample relatedness. Within the phenotype matrix all values were quantile normalised. For phenotypes where categories were defined according to a test score, the continuous score only was imputed and the category inferred from this. The kinship matrix was generated using GEMMA, from SNPs pruned to be in approximate linkage equilibrium [354]. The PLINK *indep-pairwise* function was used to produce this SNP set by removing SNPs with r^2 greater than 0.2 [259]. PHENIX was then run using default settings but without removing extreme values.

Performance of the predictive model for each phenotype was evaluated by the random removal and imputation of 5% observed data. From this the correlation between observed and estimated values could be quantified as an r^2 value. This was repeated 1000 times and the mean r^2 taken to give a measure of imputation accuracy for each phenotype.

During the imputation process latent components were generated equal to the number of input phenotypes. These described underlying patterns within the data and each one represented a unique direction of variation drawn from several phenotypes. Consequently these components were termed phenotypic axes. Every individual was assigned a score for each axis that reflected the severity of his or her associated phenotypes.

For samples in the replication dataset not included in the generation of the latent components, scores along each axis were calculated independently using their observed phenotype data. Phenotype measurements for each sample were first individually quantile normalised with respect to the original dataset.

Normalised phenotype values were then weighted according to variable loading values and residual environmental covariances to produce a severity score for each component using Equation A.

Equation A

$$\mu = [I + \beta\Sigma^{-1}\beta^T]^{-1}\beta\Sigma^{-1}y_*^T$$

Where μ = *matrix containing every individual's score for every phenotype axis*

I = *identity matrix*

β = *matrix of variable loadings*

Σ = *posterior expectation of environmental covariance across phenotypes*

y_* = *matrix of normalised phenotype values*

5.2.3) Quantitative trait GWAS

A quantitative trait GWAS was carried out using SNPtest [328]. Logistic regression was performed on scores for each phenotype axis to identify SNPs affecting its severity. Effects were analysed under an additive model that conditioned on age, sex and the first two principle components to account for any underlying population substructure. Only SNPs with minor allele frequency greater than 0.01 were investigated, as this cohort was relatively small and consequently underpowered to detect true associations of rare variants.

Any SNP significantly associated with a phenotype axis was investigated further. All observed phenotype measures correlated with that axis, defined by an absolute correlation coefficient greater than 0.3, were extracted and quantile

normalised. Logistic regression was then carried out comparing SNP genotype with the severity of all correlated clinical phenotypes. This enabled the identification of which specific phenotypes each variant was affecting and the quantification of effect size and direction.

5.2.4) Phenotype regression model

The next stage of analysis aimed to create genetic risk models of phenotype severity. SNPs were defined as suggestively associated if their logistic regression P value was less than 0.001. For each phenotypic axis, variants fulfilling this criterion were carried forward to the development of an elastic net regression model. This aimed to model phenotypic axis score from a combination of weighted genotype dosages according to Equation B.

Equation B

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Where $\hat{y}$ *= predicted score for genotype axis*

β_i *= coefficient for* SNP_i

x_i *= genotype dosage at* SNP_i

To solve this equation, elastic net regression seeks to minimise the absolute difference between observed and predicted phenotype scores (Equation C), subject to the sum of genotype dosage coefficients remaining less than α (Equation D). This constraint shrinks the coefficients, which should

prevent over-fitting. When α=0 or α=1 this is equivalent to ridge regression and lasso regression respectively, whilst elastic net fills the sliding scale between these two extremes.

Equation C

$$\sum (y - \hat{y})^2$$

Equation D

$$\sum |\beta_i| \leq \alpha$$

Before regression was carried out 10% of the total sample was removed to provide an independent set of observations with which to test the performance of the final model. The population was split into ten quantiles and 10% randomly removed from each. This ensured a distribution of phenotype scores in both the model generation and testing stages.

Genetic models were then created independently for each phenotype axis using suggestively associated SNPs from the corresponding logistic regression. Optimal values for α and λ, an additional parameter that controls shrinkage, were calculated simultaneously using cross-validation. This was performed using the R package 'Caret' [355]. With these defined parameters a bootstrapping approach was employed to identify SNPs robustly influencing phenotype scores. For each of 1000 iterations the R package 'glmnet' was used to

fit an elastic net model to a percentage of the population [356]. SNPs whose beta coefficients were non-zero were recorded for each and bootstrapping values calculated as the proportion of iterations for which a given SNP contributed to the model. Final genetic models were then generated using only those variants satisfying a minimum bootstrap threshold.

5.2.5) Polygenic risk scoring

Polygenic risk scores provide an additive measure of genetic risk for a trait calculated directly from the results of logistic regression. Theoretically this methodology is inferior as scores are derived from the analysis of each SNP independently of all others, unlike elastic net where the simultaneous analysis of all SNPs allows for the modelling of interactions and redundancies. However in some cases their performance is equal. Polygenic scoring was attempted in this analysis to try and avoid possible over-fitting by other methods.

Scores were calculated separately for each phenotypic component. Associated variants were weighted by their odds ratio, inferred from logistic regression, to reflect their estimated effect size. They were then summed to provide an additive measure of total genetic load. This is summarised in Equation E. Only SNPs for which their association P value was lower than a pre-determined threshold were included in the model. Several values were explored for this threshold ranging $1x10^{-7}$ to 0.5, as previous literature has shown that the optimal threshold varies widely for different disorders.

Equation E

$$\hat{y} = \sum \beta_i x_i$$

Where $\hat{y}$ *= polygenic risk*

β_i *= log of odds ratio for* SNP_i

x_i *= genotype dosage at* SNP_i

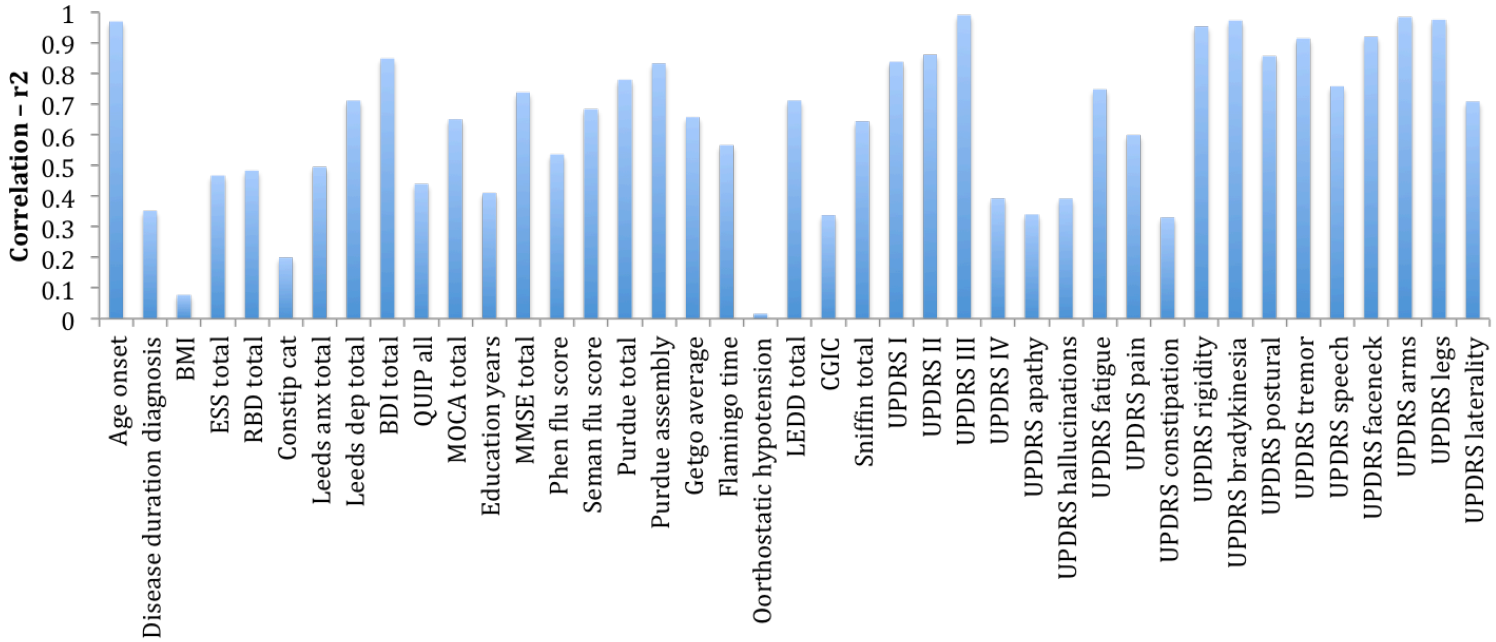
5.3) Results

5.3.1) Phenotypic imputation

PHENIX was used to impute missing phenotype data for all samples. 57 families were present in this cohort, providing a level of relatedness that should have significantly improved the identification of genetic effects and therefore the overall performance of this method. The accuracy of imputed values was measured by randomly removing and re-imputing 5% of observed data. Figure 5.2 shows the r^2 value for each phenotype, equal to the mean squared correlation between observed and imputed values. This demonstrates that the majority of phenotypes were imputed well. In general r^2 greater than 0.3 is accepted as a threshold for good quality imputation and the majority of phenotypes exceeded this.

This procedure was also carried out separately with only PD cases to estimate heritability. This ensured that only PD-specific effects were captured and avoided bias arising from some attributes not being measured in controls. Due to the shrinkage imposed by a low-rank model within PHENIX heritability

Figure 5.2: r^2 values between observed data and imputed data using PHENIX show that the majority of phenotypes were imputed well

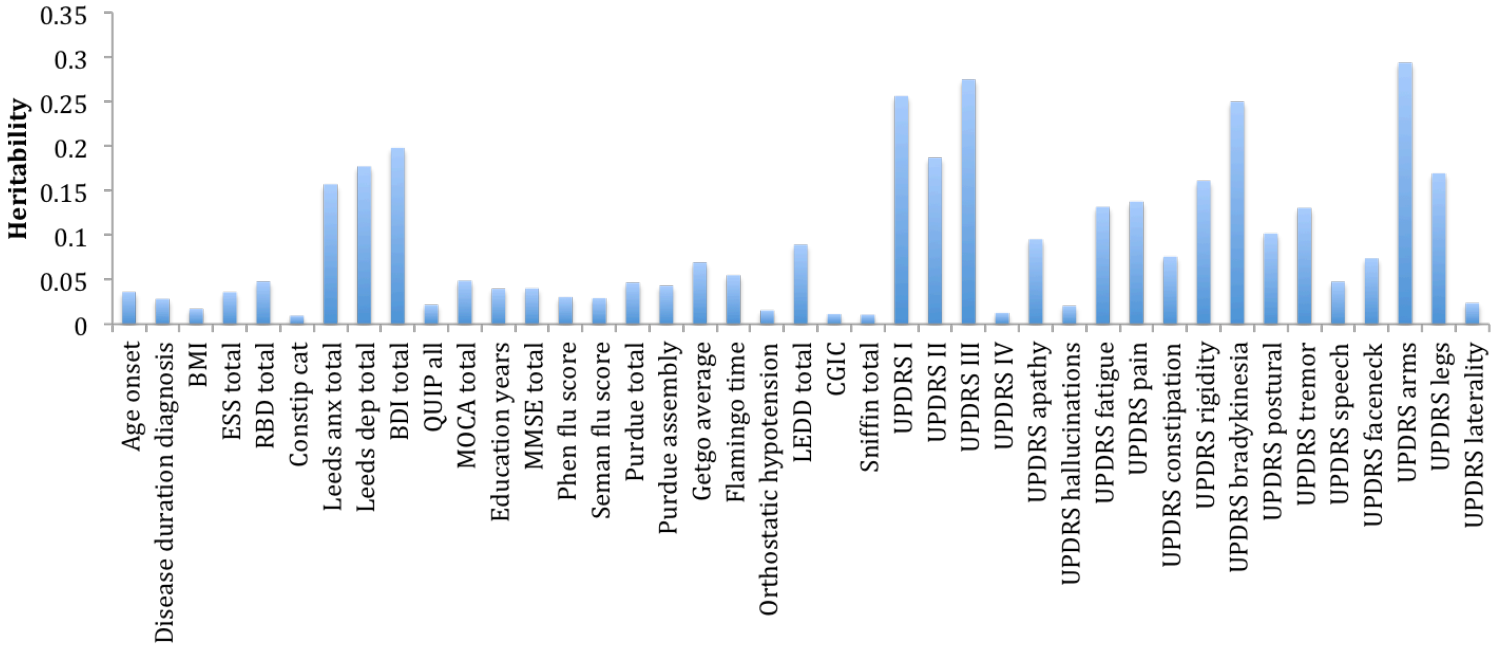


values are underestimated by this programme, and there is no straightforward way of normalising to return more accurate approximations. It can be inferred however that these represent a minimum degree of heritability.

Figure 5.3 shows the heritability estimates for individual phenotypes generated by PHENIX. Parts I, II and III of the UPDRS questionnaire were the most heritable indicating that broad features of disease progression may have a significant genetic component. Of these the UPDRS III, representing overall motor symptom severity, had highest heritability estimated at 0.27. Heritability of specific aspects of motor function was variable. It was highest for the particular limbs affected and type of movement dysfunction. However whole-body function, such as that measured by flamingo and get-up-and-go tests, had almost no genetic component. Consequently the heritability of motor symptom severity seems relatively high overall, but the particular subtype in which it manifests may be more determined by environmental factors.

Non-motor phenotypes also showed variable degrees of heritability. The UPDRS I measures mental activity, behaviour and mood and its heritability was estimated at 0.26. Within this category depression and anxiety independently showed the strongest genetic component. Pain and fatigue were also moderately influenced by genetic factors. In contrast cognitive impairment had almost no genetic component and this was consistent throughout questionnaire and verbal fluency tests. Daytime sleepiness and features of RBD also showed minimal heritability. The severity of both motor and non-motor symptoms overall therefore seemed to be moderately heritable, but the influence of genetic factors

Figure 5.3: heritability values of each phenotype as estimated in PD cases using PHENIX



on specific phenotypes was more variable and generally less than for general symptom severity.

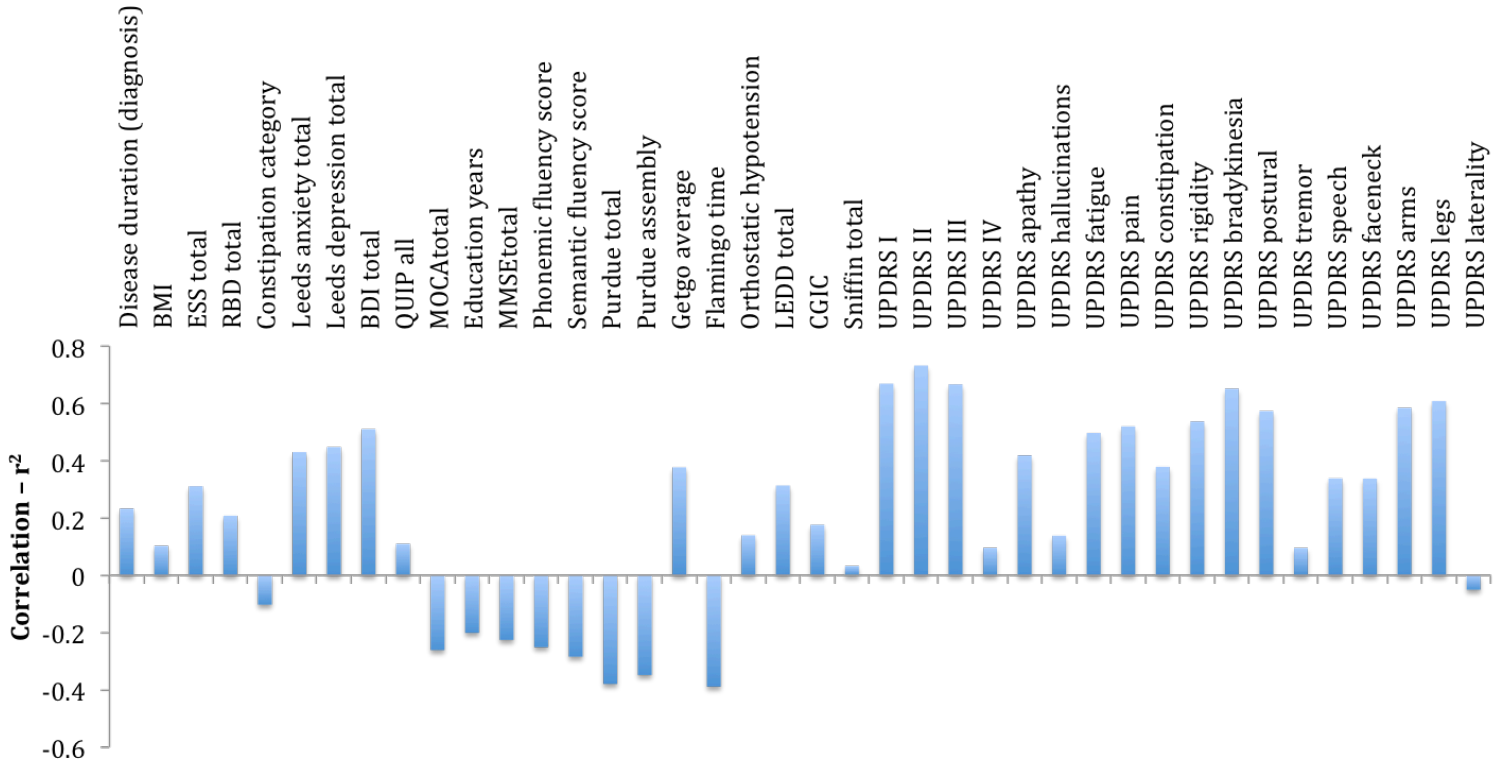
5.3.2) Generation of phenotype axes

Continuous phenotypic axes were generated using PHENIX, each of which represented the severity of a number of co-varying observed phenotypes. Although 45 axes were produced not all represented pertinent variation. To determine which were clinically relevant, the correlation between each axis and each observed phenotype was examined. This was performed beginning at the axis that explained the most variance and moving stepwise down. Only axes correlated with at least one clinically observed phenotype were used in analysis, defined by an absolute correlation coefficient greater than 0.3. Beyond the 6th component no observed phenotypes were strongly correlated with the axes, indicating they were not useful representations of the clinical spectrum. Therefore components 1-6 were used in analysis.

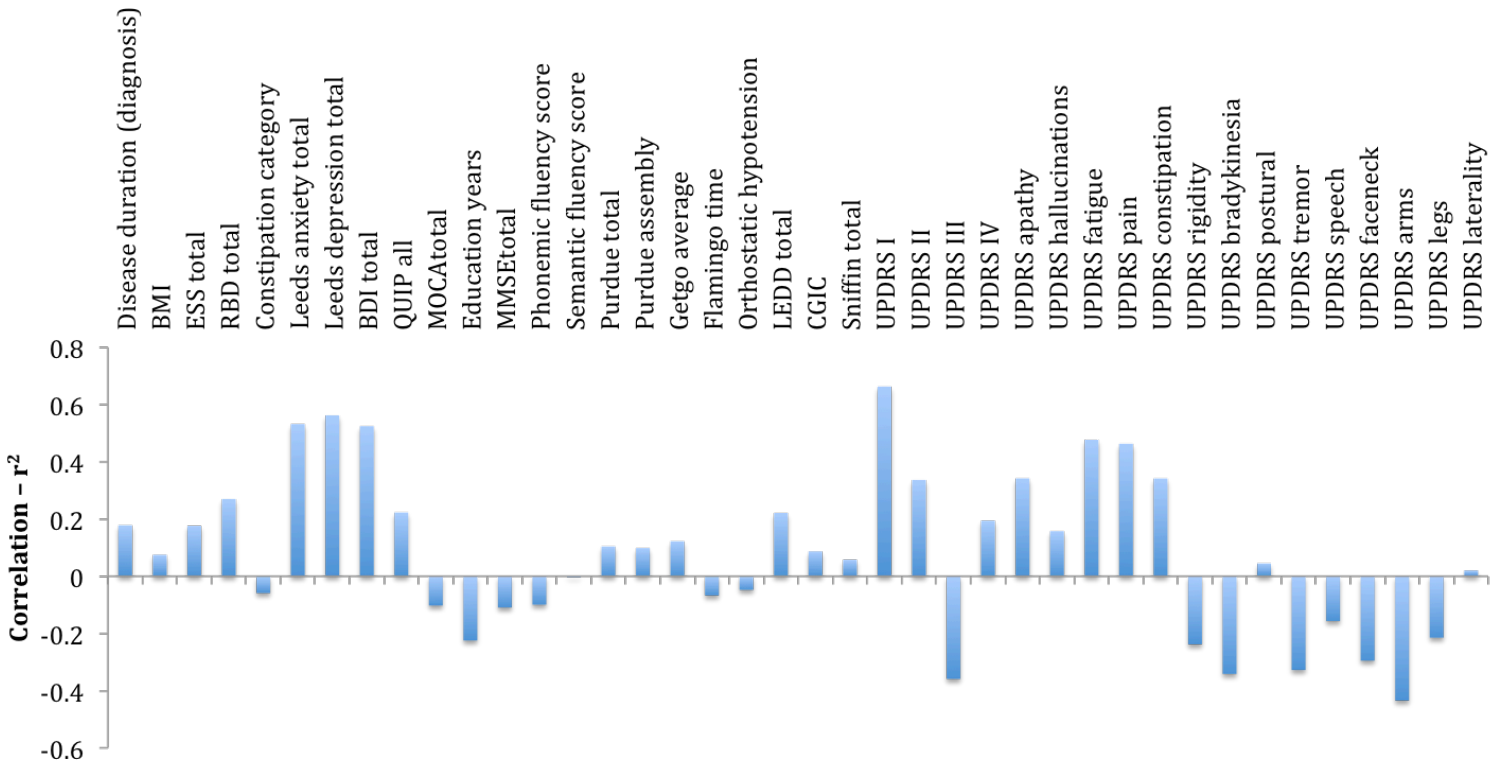
The correlation of individual observed phenotypes with each component is shown in Figure 5.4. This shows that every component represented a unique collection of phenotypes to differing extents and in different directions of severity. Axis 1 represented worsening non-tremor motor phenotypes, anxiety and depression accompanied by an opposing improvement in cognitive function. Anxiety and depression were also features of axis 2, in addition to an increase in the severity of autonomic symptoms and decrease in motor dysfunction. Axis 3 represented general motor symptom severity including rigidity, bradykinesia and tremor of the whole body independently of non-motor features. Tremor was also represented independently of almost all other phenotypes by the 4th axis.

Figure 5.4: the correlations between each phenotypic axis (A-F) and each observed phenotype

(A) Phenotype axis 1



(B) Phenotype axis 2



(C) Phenotype axis 3

(D) Phenotype axis 4
Correlation – r²
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
Disease duration (diagnosis)
BMI
ESS total
RBD total
Constipation category
Leeds anxiety total
Leeds depression total
BDI total
QUIP all
MOCAtotal
Education years
MMSEtotal
Phonemic fluency score
Semantic fluency score
Purdue total
Purdue assembly
Getgo average
Flamingo time
Orthostatic hypotension
LEDD total
CGIC
Sniffin total
UPDRS I
UPDRS II
UPDRS III
UPDRS IV
UPDRS apathy
UPDRS hallucinations
UPDRS fatigue
UPDRS pain
UPDRS constipation
UPDRS rigidity
UPDRS bradykinesia
UPDRS postural
UPDRS tremor
UPDRS speech
UPDRS facenеck
UPDRS arms
UPDRS legs
UPDRS laterality

(E) Phenotype axis 5
Correlation – r²
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
Disease duration (diagnosis)
BMI
ESS total
RBD total
Constipation category
Leeds anxiety total
Leeds depression total
BDI total
QUIP all
MOCAtotal
Education years
MMSEtotal
Phonemic fluency score
Semantic fluency score
Purdue total
Purdue assembly
Getgo average
Flamingo time
Orthostatic hypotension
LEDD total
CGIC
Sniffin total
UPDRS I
UPDRS II
UPDRS III
UPDRS IV
UPDRS apathy
UPDRS hallucinations
UPDRS fatigue
UPDRS pain
UPDRS constipation
UPDRS rigidity
UPDRS bradykinesia
UPDRS postural
UPDRS tremor
UPDRS speech
UPDRS faceneck
UPDRS arms
UPDRS legs
UPDRS laterality

(F) Phenotype axis 6
Correlation – r²
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
Disease duration (diagnosis)
BMI
ESS total
RBD total
Constipation category
Leeds anxiety total
Leeds depression total
BDI total
QUIP all
MOCAtotal
Education years
MMSEtotal
Phonemic fluency score
Semantic fluency score
Purdue total
Purdue assembly
Getgo average
Flamingo time
Orthostatic hypotension
LEDD total
CGIC
Sniffin total
UPDRS I
UPDRS II
UPDRS III
UPDRS IV
UPDRS apathy
UPDRS hallucinations
UPDRS fatigue
UPDRS pain
UPDRS constipation
UPDRS rigidity
UPDRS bradykinesia
UPDRS postural
UPDRS tremor
UPDRS speech
UPDRS faceneck
UPDRS arms
UPDRS legs
UPDRS laterality

Axis 5 was associated with increasing severity of posture and speech phenotypes and a reduction in tremor severity. Axis 6 was linked to reduced mobility and control over whole-body movement, also with an improvement in tremor phenotypes. No component was highly correlated with disease duration, indicating that none merely represented a general worsening of disease with time.

The contribution of different phenotypes to these axes was therefore highly variable. Specific aspects of motor dysfunction were important factors in defining the majority of axes. Anxiety and depression were also relatively important features but only for axes explaining the largest amounts of variation. Conversely cognitive impairment was comparatively insignificant, as severity of this phenotype was represented by only one phenotype axis.

5.3.3) Quantitative trait analysis of phenotypic axis scores

A quantitative trait GWAS was carried out independently for each phenotype axis to identify genetic variants affecting the severity of patient scores. QQ plots (Figure 5.5) show that population stratification was adequately controlled by the included covariates. A departure from the expected quantiles was observed for axes 1 and 4, which corresponded to variants surpassing genome-wide significance in both cases (Figure 5.6).

Seven variants demonstrated genome-wide significance in total, 6 of which belonged to a peak associated with axis 4 (Figure 5.7). These were examined further in addition to 53 additional variants with P values less than $5x10^{-7}$. For each SNP logistic regression was performed against all directly

Figure 5.5: QQ plots for the quantitative trait GWAS carried out for each phenotype axis (A-F)

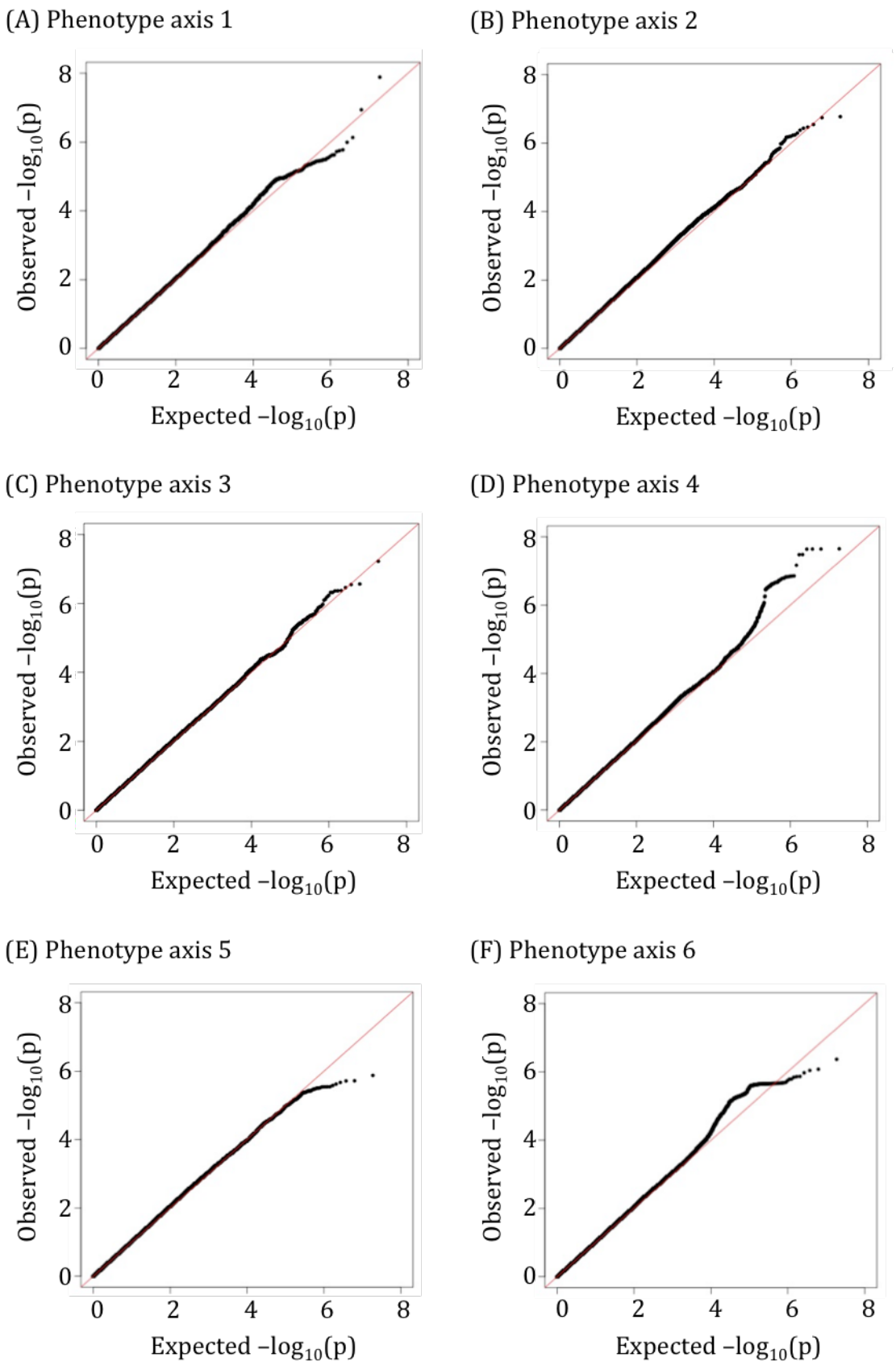
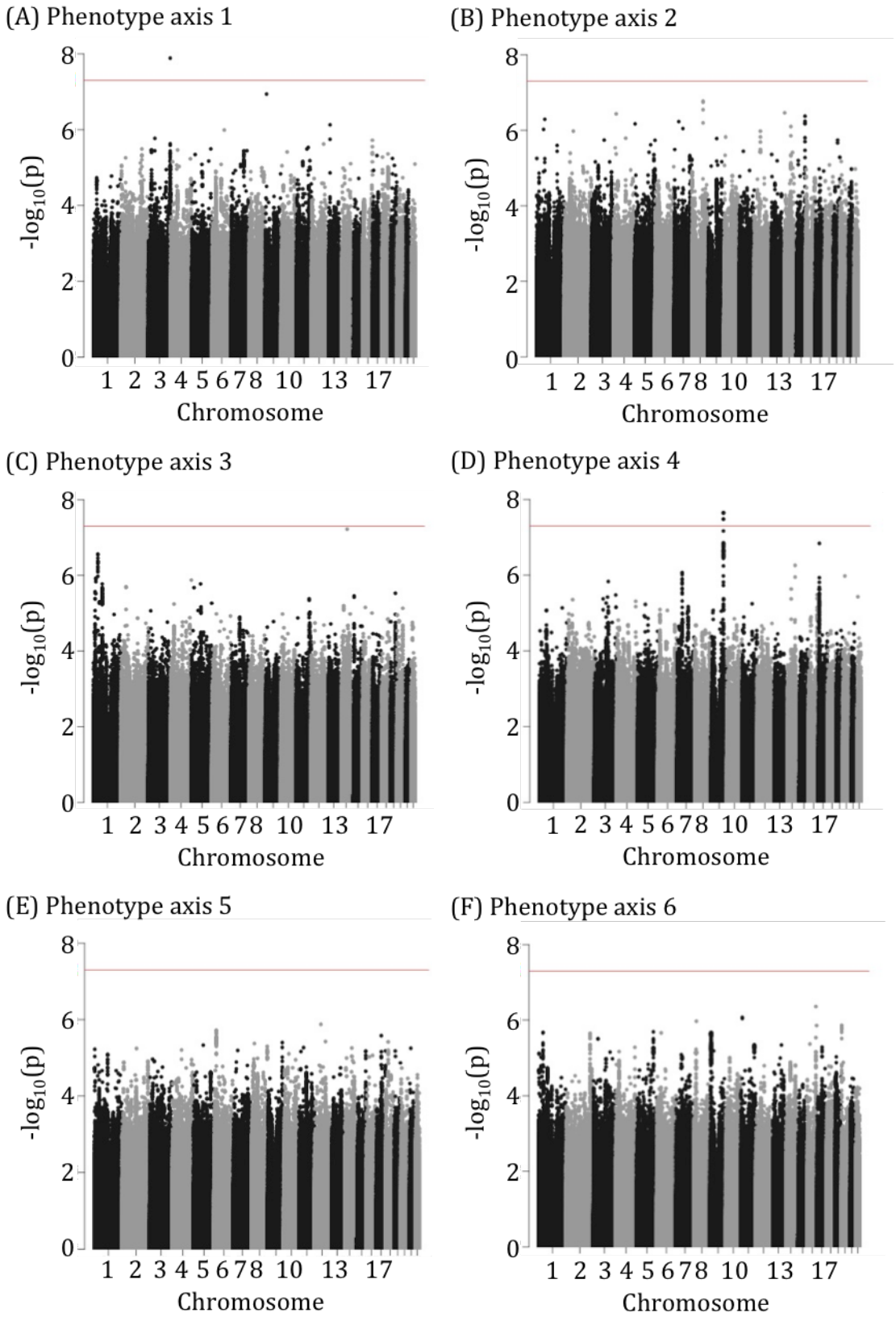


Figure 5.6: Manhattan plots for the quantitative trait GWAS carried out for each phenotype axis (A-F). Variants associated with phenotype axes 1 and 4 surpassed genome-wide significance



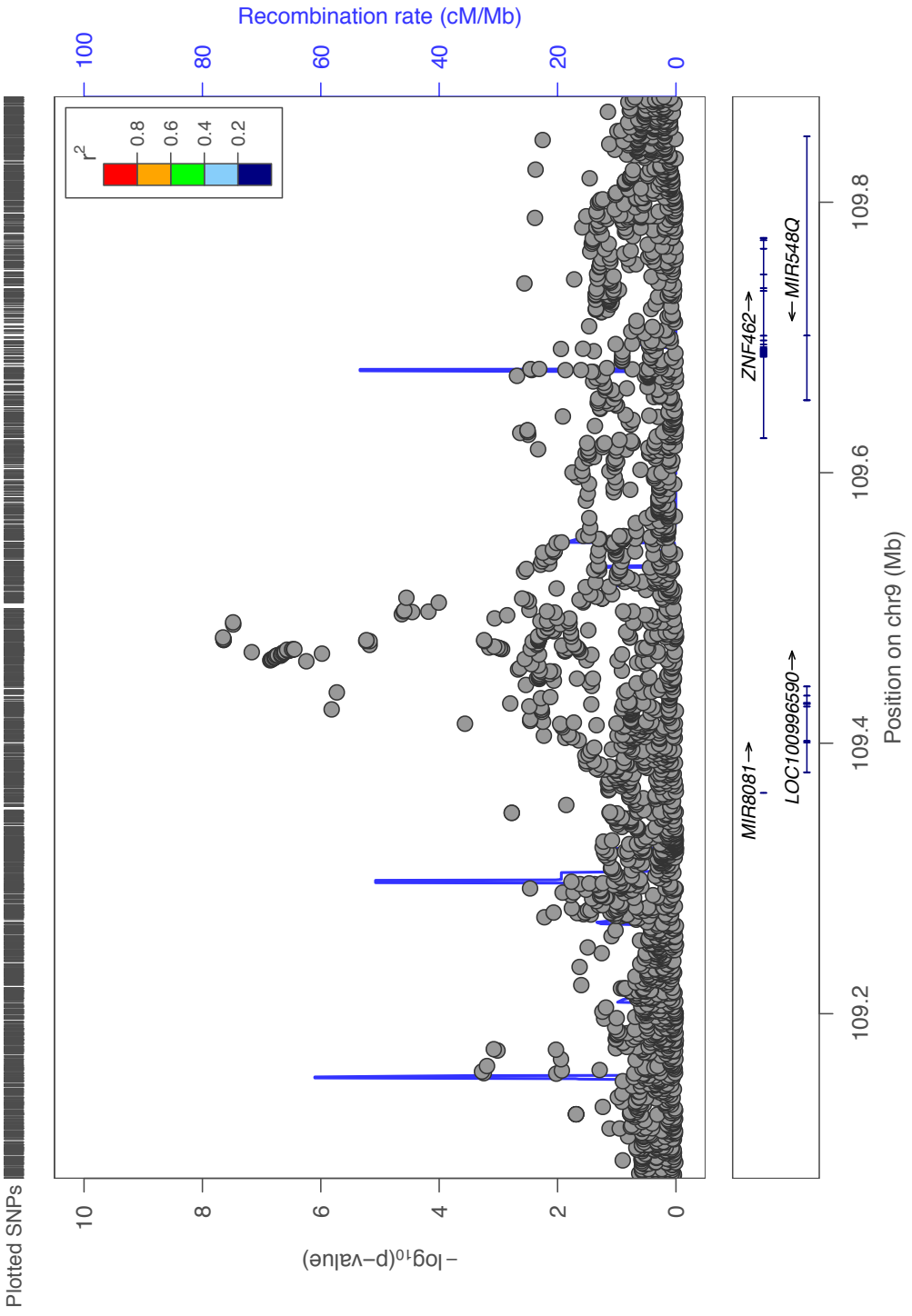


Figure 5.7: Visualized association peak of SNPs associated with phenotypic axis 4

observed phenotypes correlated with its associated phenotypic axis, to identify clinically relevant genotype-phenotype links. Results are summarised in Table 5.8.

Two SNPs were associated with the first component, both of which affected a range of motor and non-motor phenotypes. 3:194252539 (no RS ID exists, the closest SNP is 2 base pairs away at rs768869505) was associated with an improvement in anxiety and depression phenotypes in addition to fatigue and pain. Motor function was also improved, in particular bradykinesia. Rs55754428 was associated with effects in the opposite direction, causing worsening of anxiety and depression as well as motor phenotypes associated with rigidity.

The second component was associated with five variants in total, three of which were clustered around the same genomic locus. All were linked to a worsening of their associated phenotypes. The individual variants rs55819636 and rs77151048 were both strongly linked to pain in addition to general mood and mentality. Rs77151048 also affected the severity of clinical anxiety and depression. Three SNPs were closely located on chromosome 8, the most strongly associated of which was rs73696270. These variants were all linked to the same phenotypes: increased anxiety and depression and greater problems with activities of day-to-day life such as eating and dressing.

The third axis represented six observed phenotypes encompassing specific aspects of motor function. Eight linked variants clustered together in an association peak around rs115184951 and one single variant - rs12435357 – was associated independently. All variants within the peak significantly affected the severity of tremor, particularly of the arms, but no other subtype of motor

Table 5.8: Summary of genotype-phenotype associations elucidated through use of the phenotype axes

Phenotype axis	SNP	Phenotype	P value	Beta
1	3:194252539	Leeds depression total	0.000383	-0.24
		BDI total	0.000045	-0.286
		Purdue total	0.000433	0.216
		Flamingo time	0.00111	0.185
		UPDRS I	0.000000521	-0.34
		UPDRS II	0.00000053	-0.338
		UPDRS III	0.0000832	-0.26
		UPDRS fatigue	0.0000235	-0.264
		UPDRS pain	0.00198	-0.197
		UPDRS bradykinesia	0.00000708	-0.302
		UPDRS arms	0.000108	-0.26
		UPDRS legs	0.0000754	-0.267
	rs55754428	Leeds anxiety total	0.00176	0.48
		BDI total	0.0000713	0.649
		UPDRS II	0.0000136	0.692
		UPDRS rigidity	0.00169	0.498
2	rs73696270 (lead)	Leeds anxiety total	0.000230817	0.859
		Leeds depression total	0.0000867	0.968
		UPDRS II	0.00273	0.727
	rs55819636	UPDRS I	0.00289	1.48
		UPDRS pain	0.0002289	1.65
	rs77151048	Leeds anxiety total	0.000218	0.355
		Leeds depression total	0.00337	0.29
		BDI total	0.000268	0.368
		UPDRS I	0.0000349	0.407
		UPDRS pain	0.00272	0.275
3	rs115184951 (lead)	UPDRS III	0.00233	0.327
		UPDRS tremor	0.00285	0.321
		UPDRS arms	0.000513	0.376
	rs12435357	UPDRS III	0.0000378	0.374
		UPDRS rigidity	0.000177	0.343
		UPDRS bradykinesia	0.00000192	0.439
		UPDRS arms	0.0000417	0.377
		UPDRS legs	0.0000047	0.422
4	rs80346154	UPDRS apathy	0.00552	0.726
	rs76433669	UPDRS tremor	0.0000959	-0.187
6	rs7200759	UPDRS tremor	0.00931	-0.132

dysfunction. Conversely rs12435357 did not associate with tremor but was strongly linked to increased severity of all other phenotypes represented by this axis, including rigidity and bradykinesia of both the arms and legs.

42 variants were associated with the fourth axis, consisting of one independent SNP and 41 SNPs within a single GWAS peak around rs80346154. The single variant rs76433669 was associated with a decrease in tremor severity independently of all other motor and non-motor features. Variants forming the association peak were not linked with any of the phenotypes correlated with r^2 greater than 0.3 with this axis. Consequently the correlation threshold was lowered and all observed phenotypes obtaining a correlation coefficient with this axis greater than 0.2 were analysed. This elucidated a significant association with apathy, whereby these SNPs were linked to an increase in phenotype severity (p=0.00465).

A single variant rs7200759 was associated with the sixth axis. This axis represented several motor phenotypes however the associated SNP was linked only to a decrease in tremor severity. No SNPs were associated with the fifth axis.

5.3.4) Elastic net regression

Elastic net regression was used to produce models of phenotype severity as a product of weighted genotype dosages, in an attempt to predict likely disease course from genetic data. It was not computationally feasible to input all genotyped SNPs into the regression models, therefore logistic regression P values below 0.001 were used to filter variants for some evidence of association.

This resulted in an average of 11400 SNPs being input to the elastic net model for each axis, ranging between 9800 and 13200.

90% of the population was used as a discovery cohort to define genetic models for each phenotype axis. Cross-validation was used to calculate optimum values for the fixed model parameters α and λ, which control the penalty size and shrinkage respectively. Bootstrapping was subsequently performed to identify the most informative SNPs, which were then used to create the final model. This was evaluated in the remaining 10% of samples by comparing observed and predicted scores using the Pearson correlation test implemented in R.

5.3.4.1) Elastic net, ridge and lasso regression

When elastic net regression was performed the final models contained a large number of SNPs in close physical proximity, all of which had corresponding small coefficients. On examination it was discovered that this was because cross-validation had defined a low value of α close to zero and subsequently a procedure very similar to ridge regression. This shrunk the coefficients of highly correlated predictors close to each other, which resulted in the inclusion of a high proportion of redundant variants in the generated models [356].

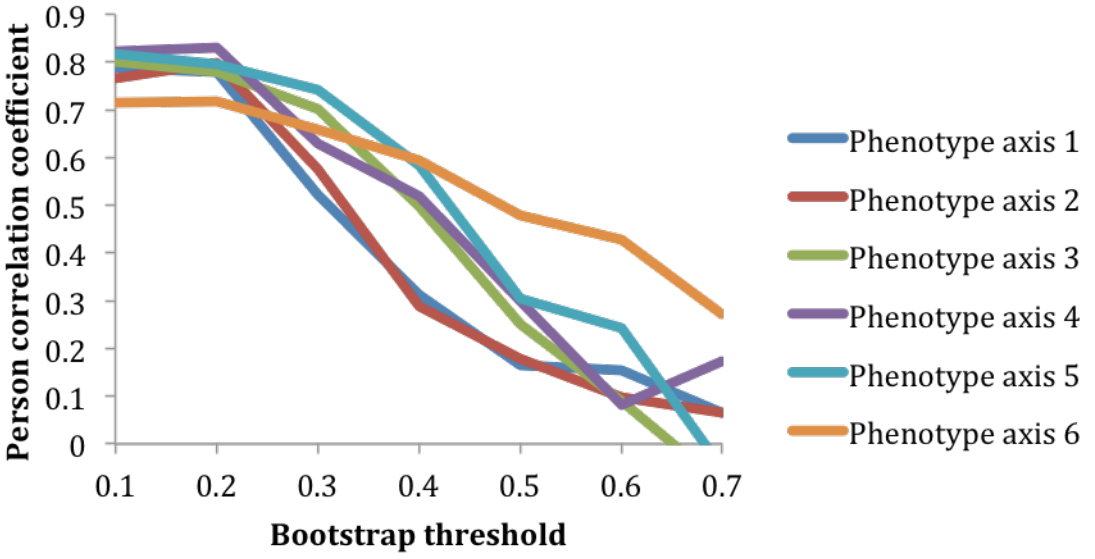
The characteristics of lasso regression are opposite: in the presence of highly correlated predictors it chooses one and assigns the remainder zero coefficients [357]. This is a more desirable characteristic for this study as it constitutes a form of variable selection, so uses only the most informative, independent SNPs to construct the final model. Effect sizes of associated genomic regions are also easier to interpret. Consequently in all further analyses α=1 was specified so that lasso regression was performed.

5.3.4.2) Bootstrapping

1000 bootstrap iterations were performed by holding out a random subset of individuals and performing lasso regression on those remaining. SNPs included in more than a defined proportion of the models produced by this process were used to form the final genetic model. However there was no obvious choice of this threshold due to previously filtering SNPs based on their likely association with the phenotype axis. Consequently genetic models were generated for a range of values and the ability of each to predict phenotype severity evaluated in the independent replication set.

Correlation between observed phenotype score and that predicted by the genetic model was evaluated by the Pearson correlation coefficient at bootstrap thresholds between 10% and 90% at 10% intervals. This was plotted for each phenotype axis individually in Figure 5.9, showing that as more SNPs were included the performance of the models improved until they reached a plateau.

Figure 5.9: Pearson correlation coefficient values between observed phenotype scores and those predicted from genotype using different bootstrapping thresholds for each axis. As more SNPs are included the models increase in accuracy.



The beginning of this plateau, a threshold of 0.2, was chosen as this provided the most accurate genetic models without including non-informative and redundant SNPs.

For these models high correlations were observed between known phenotype scores and those predicted from genotype in the independent samples excluded before the model generation phase (Figure 5.10). Pearson correlation coefficients were above 0.7 and correlation P values were highly significant for all six phenotypic axes. This indicated that phenotype severity was highly influenced by genetic factors.

An additional validation step was carried out using samples that had not been involved in any stage of analysis thus far, so were independent of the generation of phenotype axes, quantitative trait GWAS and lasso regression. For each sample their observed phenotypic axis scores were calculated from their clinically measured phenotype data. The genetic model for each axis was then applied to genotype dosages to give an estimate of phenotypic severity attributable to individual genetic variation.

In these samples there was no correlation between observed phenotype score and that predicted from genotype. Pearson correlation coefficients did not exceed 0.16 and P values for correlation were insignificant for all phenotype axes. Models corresponding to alternative bootstrapping thresholds also performed poorly and correlation coefficients did not exceed 0.23 (Table 5.11). Furthermore a number of coefficients were negative, implying genetic scores were less informative of phenotype severity than merely taking the sample mean.

Figure 5.10: observed phenotype scores and those predicted from genotype in the independent sample set for each of the six phenotype axes. This demonstrates that genotype is can be used to accurately predict phenotype severity in these individuals

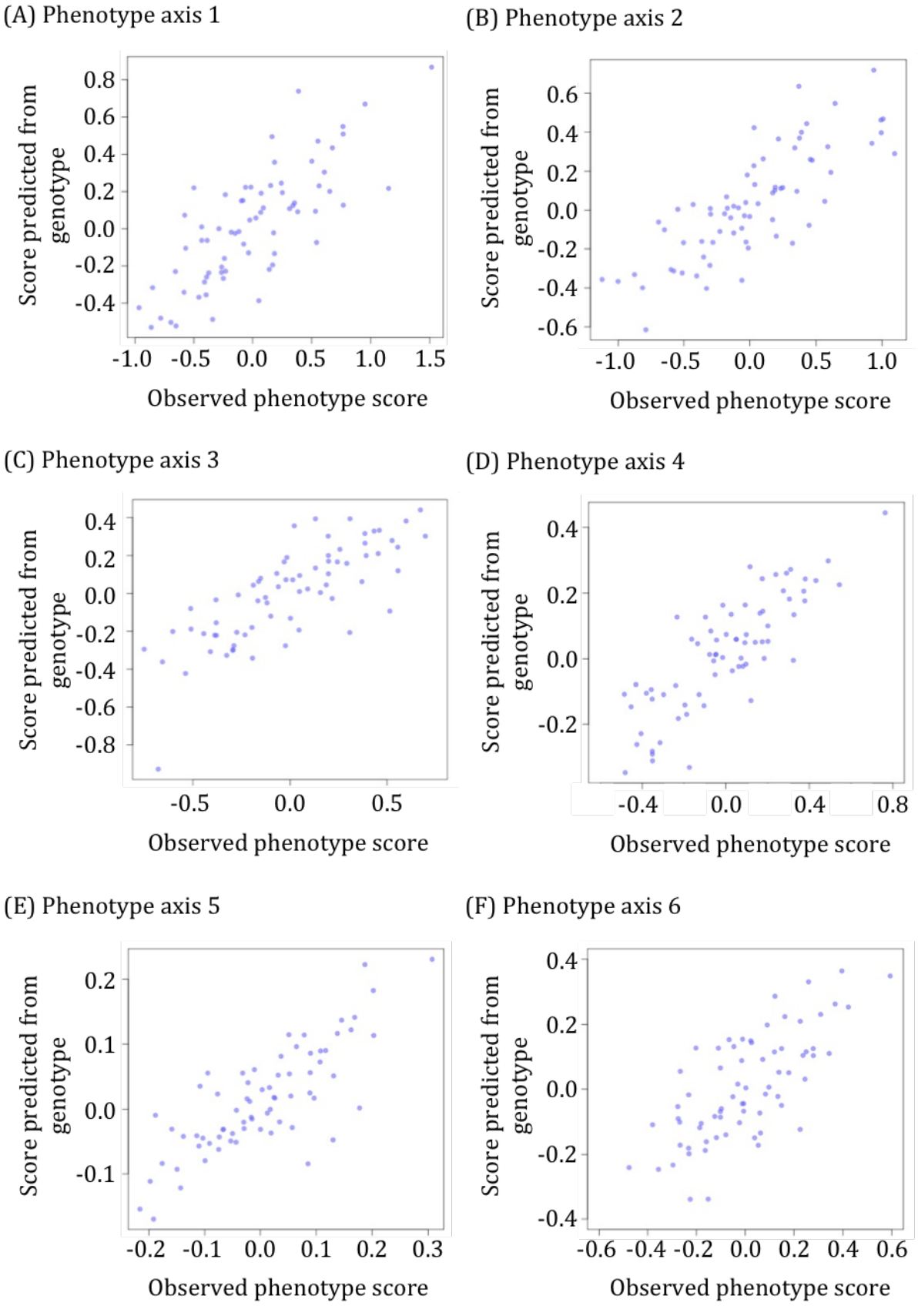


Table 5.11: Pearson correlation coefficients between observed phenotype scores and those predicted from genotype in an additional independent dataset show that none of the genetic models accurately reflect phenotype severity

Phenotype axis	Bootstrap threshold						
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**
1	0.011	0.0056	0.034	0.023	0.0029	0.12	0.023
2	0.13	0.16	0.23	0.19	0.23	0.040	0.096
3	-0.022	-0.072	-0.029	-0.027	0.10	0.055	0.22
4	0.070	-0.042	0.14	0.0075	0.018	-0.0058	0.055
5	0.071	-0.044	-0.095	-0.097	-0.067	-0.066	-0.076
6	-0.10	-0.12	-0.15	-0.045	-0.0044	-0.050	-0.18

5.3.4.3) Predicting phenotypic extremes

Small and intermediate variability in phenotype severity is likely to be noisy and may also have a significant environmental component. Consequently only extremes of genetic load might impart a discernable effect on phenotype. As an alternative to predicting the entire spectrum of severity the ability of genetic models to identify extremely low and high phenotype scores was investigated.

Receiver Operator Characteristic (ROC) curves were produced to quantify how well samples with mild and severe phenotypes were separated by their genetic scores. Values of the highest and lowest percentiles of phenotype scores were calculated from the reference dataset for each axis to represent phenotypic extremes. Samples independent of all analyses were then used to generate ROC curves reflecting how well individuals within these extremes could be identified.

Genetic scores for all axes were ineffective at separating individuals in the lowest and highest 20^{th} and 25^{th} percentiles of phenotypic severity. Area Under Curve (AUC) values indicate that the level of identification achievable was only marginally greater than random chance. Separation of individuals in the highest

and lowest 10th percentiles was slightly better and AUC above 0.7 was observed for axes 2 and 3 (Table 5.12). However this was still not high enough to be clinically useful. Furthermore the minimum AUC value of 0.5 was observed for axis 6, indicating the performance of genetic models for different axes was inconsistent.

This pattern of inconsistency was reflected across models created from all bootstrapping thresholds. There was also no correlation between model performance and the inclusion of more or less variants. Consequently no single threshold provided the best solution for all phenotype axes. This indicated that the small amount of phenotypic variance explained by some genetic models was likely due to chance rather than biological relevance. Correspondingly AUC values did not reach a clinically useful value for any model.

Despite initially promising results these genetic models were therefore not useful predictors of phenotypic severity in the general PD population. This lack of replicability indicated a systematic bias within the original analysis. Consequently investigations were carried out to identify from where this was originating.

Table 5.12: AUC values reflecting the ability of the genetic models to separate individuals in the lowest and highest percentiles of phenotypic severity

Phenotype axis	25th percentile	20th percentile	10th percentile
1	0.519	0.538	0.575
2	0.562	0.512	0.725
3	0.526	0.473	0.758
4	0.564	0.586	0.539
5	0.544	0.548	0.667
6	0.612	0.669	0.5

5.3.4.4) Robustness of the phenotype axes produced by PHENIX

The phenotypic axes generated by PHENIX were an abstract representation of phenotype and were therefore not guaranteed to relate to biological underpinnings. Results pertaining to observed phenotypes in section 3.3 would indicate that in this case they were clinically relevant. Nonetheless the effects of several potentially modifying factors on the final axes generated by PHENIX were investigated. Considerable variation of the axes under such changes would have been indicative that they did not well represent the underlying biology.

In order to determine how stable the axes were to changes in the underlying data PHENIX was run on different population subsets. If the axes were biologically relevant they should have been robust to changes in sample size given it remained sufficiently large to detect inter-phenotype relationships. Whilst keeping all other variables identical PHENIX was applied to a randomly selected 80%, 90% and 100% of the population. The phenotype axes varied only minimally due to sample size, both in the observed scores for each individual and the correlation of each component with observed phenotypes. Therefore PHENIX did not appear to be overfitting the phenotypic axes to samples that remained in the analysis.

The effect of kinship matrix was also investigated. As the axes were genetically informed the method by which relatedness was quantified could have affected their structure. The PLINK *z-genome* function was used to estimate the proportion of alleles identical by descent (IBD) pairwise for all individuals. GEMMA was also used and matrices calculated for both standardised and

centred genotypes. All methods were carried out with the full SNP set and a pruned SNP set formed using the PLINK *indep-pairwise* function.

In total six kinship matrices were created and used identically to generate phenotype axes in PHENIX. For all kinship matrices the axes remained almost identical to those originally defined. The method by which relatedness was quantified was therefore not a cause of bias in this analysis.

The PHENIX method incorporates a stochastic process which is defined by a random component. This can produce artefacts, but if the phenotypic axes represented underlying biology then they should have been robust to this variability and remained constant between different iterations. Throughout the analysis thus far the start point for random number generation had been specified to ensure tractability. Next, PHENIX was run ten times on identical input data but specifying different random number seeds. Although some components were flipped, so positive individual scores became negative and vice versa, the group of observed phenotypes that each axis represented remained constant. The robustness of the phenotypic axes to a random component in their generation indicated that they reflected genuine biological patterns in the phenotype structure.

The PHENIX model also contained an error term to account for random small variation in phenotype measurements. This could not be accounted for when projecting new samples onto existing phenotype axes, which could have affected the accuracy of phenotype scores for additional individuals. In order to examine the effect of this term, phenotype scores were calculated from observed measurements for individuals whose scores were originally assigned as part of

the initial PHENIX analysis. The difference between these and their original scores quantified the size of effect the additional error term imparted. The newly calculated scores deviated only minimally from their true values for all phenotype axes. Consequently it was concluded that the error term imparted no significant effect on the calculation of phenotype scores for additional individuals.

The phenotypic components generated by PHENIX were therefore robust to changes in population size, kinship matrix, random number seed and error term. This indicated that they were highly likely to represent genuine biological patterns. Consequently this stage of the analysis was unlikely to be the cause of poor genetic models.

In order to confirm this a group of samples were included in the PHENIX stage only and removed before the quantitative trait GWAS and lasso regression steps. These individuals were therefore used in the production of the phenotype axes but were completely independent of any genetic analysis. Consequently if the variable performance of genetic models was attributable to PHENIX then the severity of phenotypes among these samples should have correlated well with their genetic scores.

Similarly to samples excluded from all analyses, the genetic scores for these individuals did not correlate with their phenotypic severity. Pearson correlation coefficients were close to zero for most axes and P values for correlation were not significant regardless of the bootstrapping threshold applied (Table 5.13). Consequently it was concluded that the phenotypic axes identified by PHENIX were not the cause of bias among the genetic models.

Table 5.13: Pearson correlation coefficients between observed phenotype scores and those predicted from the genetic models for individuals removed between the PHENIX and quantitative trait GWAS steps. These values show that the models are poor predictors of phenotypic severity in these individuals.

Phenotype axis	Bootstrapping threshold					
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**
1	0.252	0.187	0.104	0.0109	0.0443	-0.0711
2	0.0139	0.117	-0.0762	-0.0748	-0.134	0.00115
3	0.0261	-0.00168	0.041	0.00119	0.06	0.031
4	0.149	0.0994	0.0995	0.141	0.148	0.189
5	-0.0948	-0.0161	-0.176	-0.0491	0.0058	0.0507
6	0.115	0.0548	0.168	0.131	0.266	-0.00315

5.3.5.5) Robustness of the quantitative trait analysis filter

A quantitative trait GWAS was performed for each phenotype axis in order to filter out most non-relevant variation. Only those SNPs demonstrating some evidence of association, defined by a P value less than 0.001, were carried forward to the regression model. However GWAS results are highly linked to the statistical power of the discovery sample, so the small cohort used in this analysis may have resulted in the exclusion of some causal SNPs by this filter. This was impossible to test without knowing which SNPs were pathogenic, so the robustness of the particular SNPs selected was investigated as a surrogate. Excessive variability among these SNPs would have indicated a high probability that causal variants could have been excluded by chance and conversely that many benign variants may have been included in the lasso regression.

To test the robustness of SNPs associated with phenotype severity the quantitative trait GWAS was carried out again having randomly removed an additional 10% of samples. On average 800 fewer SNPs had a P value less than 0.001 in this analysis than that using the whole population, likely attributable to

reduced statistical power. Only 48-58% of the variants surpassing this threshold were common between the analyses performed using the two different sample sizes (Figure 5.14). Filtering based on GWAS P values was therefore susceptible to sample-specific bias and may have contributed to the poor performance of the genetic models produced. However lasso regression with all SNPs was not computationally feasible so this step could not be discarded completely.

A less stringent P value threshold of 0.01 was attempted for the first two phenotype axes, to increase the probability of including causal variants and those of small effect. It was hoped that this might allow the development of more universally applicable final models. When the bootstrapping procedure was performed few SNPs were present in more than 50% of iterations and none above 60%. This indicated a high number of superfluous SNPs not involved in moderating phenotype severity. Nonetheless final genetic models were created for several bootstrapping thresholds as before.

These were then applied to individuals excluded before the regression procedure. The correlation between genetic score and phenotype severity was considerably lower than that observed using the previous P value threshold of 0.001 and correlation coefficients did not exceed 0.2 (Table 5.15a). As expected samples excluded from both PHENIX and genetic analysis also showed poor correlation between genetic and phenotypic scores (Table 5.15b). Relaxation of the P value threshold therefore did not alleviate the potential problems associated with its use.

Figure 5.14: the number of SNPs passing the P value threshold from quantitative trait analysis for each phenotype axis (A-F), calculated from the original sample (green) and with an additional 10% of individuals removed (purple). Overlap of variants passing this filter (blue) is relatively low.

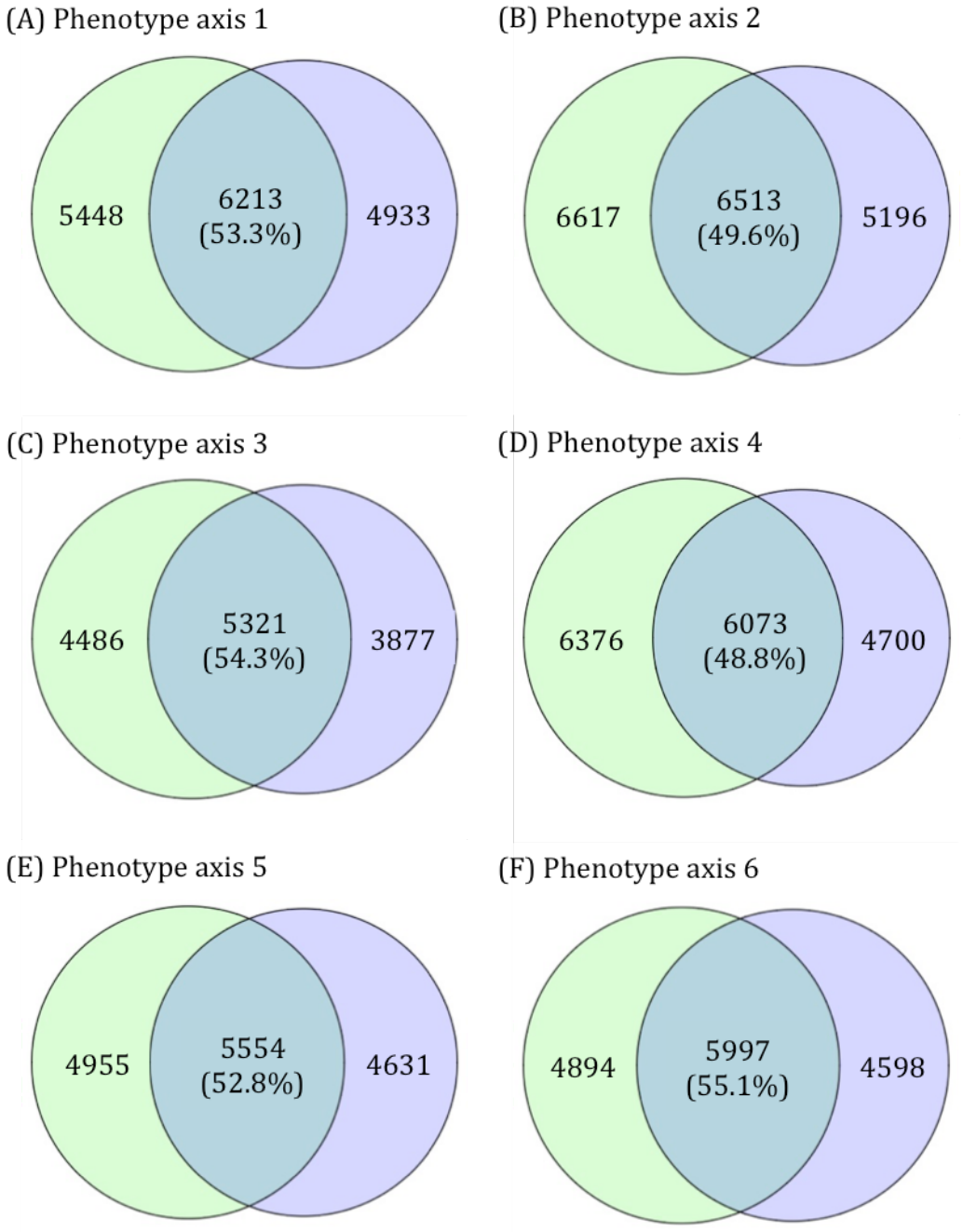


Table 5.15: Pearson correlation coefficients between observed phenotype scores and those calculated from genetic models descending from a less stringent P value threshold ($p<0.01$) in samples removed before the lasso regression (A) and before the PHENIX step (B)

(A)

Phenotype axis	Bootstrapping threshold				
	0.1	**0.2**	**0.3**	**0.4**	**0.5**
1	0.0664	0.108	-0.0733	-0.00629	0.0415
2	0.178	0.14	0.113	0.0104	-

(B)

Phenotype axis	Bootstrapping threshold				
	0.1	**0.2**	**0.3**	**0.4**	**0.5**
1	0.062	0.12	0.199	0.139	0.00338
2	0.0966	0.314	0.135	0.11	-

5.3.5.6) Polygenic risk scoring

An alternative cause of bias among the genetic models could have originated from lasso regression over-fitting to the training sample. Polygenic risk scores were explored as an alternative measure of genetic risk to try and avoid this problem. They were calculated independently for each phenotype axis as the sum of effect alleles each weighted by their log odds ratio. Only SNPs with association P values less than a pre-defined threshold were used. A range of values was explored for this threshold ranging 0.0000001-0.5. Models were tested using the subset of individuals excluded from the beginning of analysis.

As before both the Pearson correlation coefficient and ROC curve analysis were used to determine the performance of these models on continuous and

binary scales respectively. Table 5.16 demonstrates that for all phenotype axes the correlation between observed phenotype scores and those predicted from genotype was low for most P value thresholds. However there was a marked increase in correlation as the threshold became more liberal and the best-fitted model for most axes was produced using SNPs with P values up to 0.5.

Polygenic scores were then investigated as predictors of phenotypic extremes. ROC curves were generated to examine how well these scores could separate individuals within the lower and upper 10^{th}, 20^{th} and 25^{th} percentiles of phenotypic severity. For P value thresholds below 0.01 the AUC was highly variable and did not show a consistent trend of improvement or decline as more SNPs were included, aligning with the correlation analysis (Table 5.17). However models generated using thresholds of 0.1 and 0.5 both performed relatively well when phenotypic extremes were represented by the 10^{th} percentiles, with AUC values exceeding 0.75 for several components (Figure 5.18).

5.4) Discussion

PD is a highly heterogeneous disorder and as such the study of phenotypic variation could be crucial for understanding diverse molecular mechanisms that might be masked under a traditional case-control framework. This analysis quantified the heritability of individual phenotypes among PD patients, identifying those for which genetic and environmental components were important. A novel method of quantifying phenotypic variation was developed that provided a continuous measure of severity, whilst also reflecting

Table 5.16: Pearson correlation coefficients between observed phenotype scores and polygenic risk scores calculated using different P value thresholds

Phenotype axis	P value threshold							
	$1x10^{-7}$	$1x10^{-6}$	$1x10^{-5}$	$1x10^{-4}$	$1x10^{-3}$	0.01	0.1	0.5
1	0.204	0.00679	0.136	0.0155	0.0961	-0.00262	0.102	0.0941
2	-	0.0776	-0.0198	-0.0259	-0.0195	0.0896	0.221	0.221
3	0.126	-0.0577	-0.1	-0.164	-0.217	-0.066	0.109	0.152
4	-0.144	-0.163	-0.255	-0.157	-0.0115	0.111	0.15	0.234
5	-	-	-0.0761	0.0238	0.0343	0.0789	0.0658	0.0992
6	-	-0.0886	-0.0917	-0.0505	-0.0733	-0.0125	0.17	0.153

Table 5.17: AUC values representing the ability of polygenic risk scores to differentiate between individuals in the lowest and highest 10th (A), 20th (B) and 25th (C) percentile of phenotype severity

(A)

Phenotype axis	P value threshold							
	$1x10^{-7}$	$1x10^{-6}$	$1x10^{-5}$	$1x10^{-4}$	$1x10^{-3}$	0.01	0.1	0.5
1	0.625	0.6	0.7	0.475	0.625	0.525	0.6	0.55
2	-	0.6	0.525	0.75	0.575	0.45	0.75	0.8
3	0.788	0.758	0.546	0.667	0.97	0.849	0.879	0.758
4	0.544	0.484	0.791	0.56	0.44	0.626	0.637	0.681
5	-	-	0.667	0.7	0.6	0.667	0.867	1
6	-	0.619	0.762	0.714	0.452	0.405	0.762	0.667

(B)

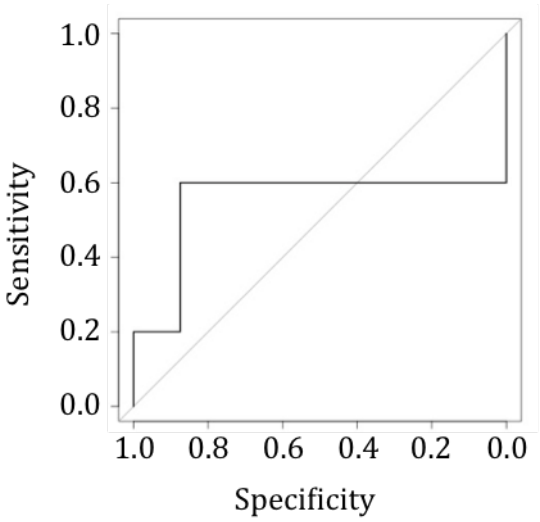
Phenotype axis	P value threshold							
	$1x10^{-7}$	$1x10^{-6}$	$1x10^{-5}$	$1x10^{-4}$	$1x10^{-3}$	0.01	0.1	0.5
1	0.676	0.543	0.614	0.7	0.662	0.519	0.467	0.543
2	-	0.5	0.615	0.597	0.535	0.418	0.724	0.721
3	0.627	0.67	0.509	0.612	0.719	0.621	0.607	0.674
4	0.568	0.564	0.682	0.605	0.591	0.568	0.559	0.627
5	-	-	0.619	0.393	0.571	0.571	0.607	0.667
6	-	0.546	0.604	0.714	0.513	0.578	0.571	0.584

(C)

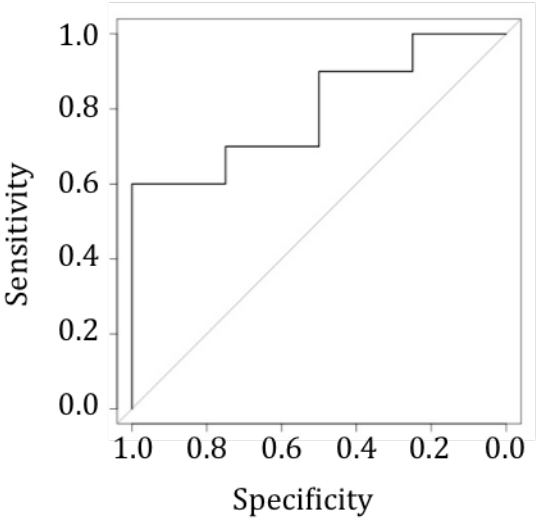
Phenotype axis	P value threshold							
	$1x10^{-7}$	$1x10^{-6}$	$1x10^{-5}$	$1x10^{-4}$	$1x10^{-3}$	0.01	0.1	0.5
1	0.672	0.564	0.644	0.661	0.611	0.525	0.519	0.461
2	-	0.545	0.574	0.621	0.557	0.414	0.702	0.686
3	0.556	0.515	0.588	0.661	0.659	0.529	0.655	0.703
4	0.496	0.513	0.693	0.622	0.542	0.566	0.622	0.669
5	-	-	0.5	0.475	0.533	0.572	0.656	0.703
6	-	0.54	0.518	0.67	0.554	0.585	0.54	0.594

Figure 5.18: ROC curves showing the ability of polygenic risk scores calculated from SNPs with P value less than 0.5 to identify individuals in the lowest and highest 10th percentile of phenotype severity for each phenotype axis (A-F)

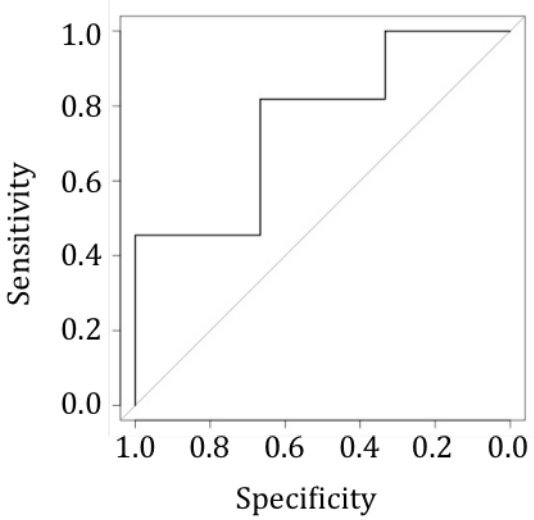
(A) Phenotype axis 1



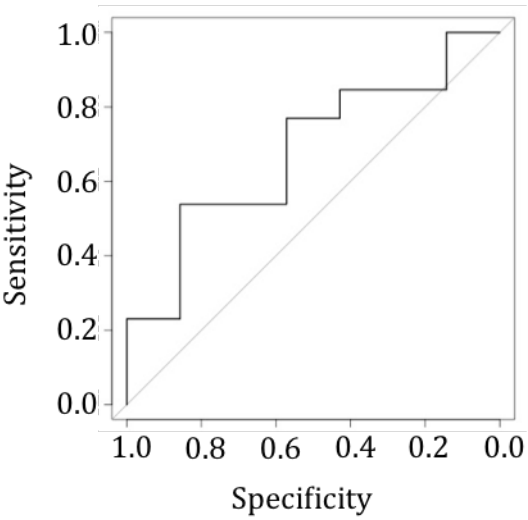
(B) Phenotype axis 2



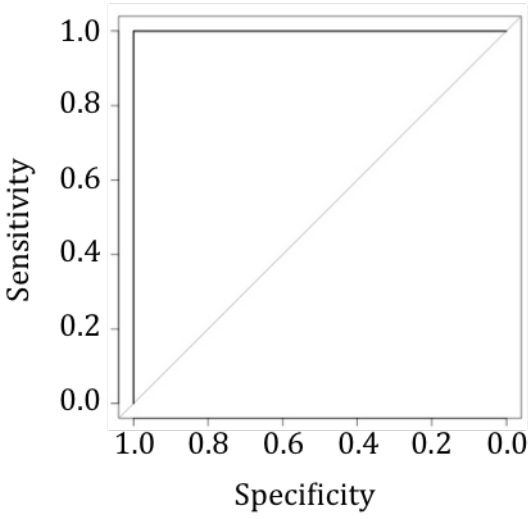
(C) Phenotype axis 3



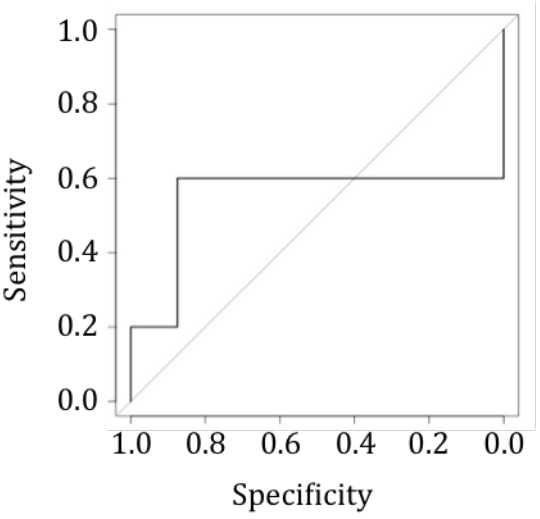
(D) Phenotype axis 4



(E) Phenotype axis 5



(F) Phenotype axis 6



relationships between phenotypes. This facilitated the first genome-wide analysis of phenotypic variability within PD, which identified 59 variants in 10 genomic regions significantly affecting the severity of specific observed phenotypes. Several genetic scoring systems were trialled, but were unable to provide a useful means by which individuals with severe phenotypes could be identified.

5.4.1) Heritability estimates

Heritability of PD is estimated at around 27%, however few studies have investigated the heritability of individual phenotypes [358]. In this analysis the genetic component of observed phenotypes was estimated using PHENIX. Although shrinkage methods employed in this model result in underestimation of the true genetic component, these values can still be interpreted as useful minimums.

Surprisingly age of onset showed low heritability in this study. The segregation of pathogenic PD variants with early-, mid- and late-onset disease indicates that there should be a significant genetic component to this phenotype. However a previous study has shown that age of onset is only moderated by a heritable component among individuals with pathogenic mutations [359]. Siblings concordant for PD but with no known genetic cause showed no significant heritability for this attribute [359]. There were few individuals with known pathogenic mutations in this cohort, aligning with this result. Consequently age of onset in sporadic PD may largely be an environmentally influenced trait.

General motor symptom severity was one of the most heritable aspects of PD, however among subtypes of motor dysfunction heritability was variable and generally less. The severity of motor phenotypes is highly correlated with the extent of neurodegeneration [360, 361]. This result could therefore reflect a genetic component directly affecting the rate of neuronal decline, which is associated with the severity of motor impairment as a secondary effect. The particular aspects of motor dysfunction in which it manifests may then be somewhat independent of this and more influenced by environmental factors.

Anxiety and depression were moderately heritable in this analysis. Although clinically distinct these phenotypes are highly comorbid. This and similar degrees of heritability could indicate shared genetic mechanisms underlying both. No studies have examined heritability of anxiety or depression specifically within PD, but estimates are between 29-42% in the general population [362, 363], supporting a genetic basis for their severity. These disorders are also more prevalent among relatives of PD patients [364], indicating that the genetic variation associated with these phenotypes may overlap with that mediating disease onset.

Cognitive decline was largely independent of genetic factors in this analysis. Few studies have investigated the heritability of cognitive phenotypes within PD, however analyses for a related disorder, dementia with Lewy bodies, also show low heritability and therefore support this result [365, 366]. Cognitive impairment among PD patients is associated with alterations to functional connectivity between brain regions [367]. The severity of cognitive impairment is highly correlated with the extent of these alterations, and as more connections

deteriorate and more brain regions are affected cognitive impairment worsens [367]. Low heritability of this trait indicates that these structural changes and the resulting risk of dementia are not substantially controlled by genetic variation, but are mainly influenced by external factors.

Sleep traits were associated with minimal genetic influence in this analysis. However studies of monozygotic twin pairs indicate a significant heritable component of multiple REM and non-REM sleep traits and daytime sleepiness [368-370]. The low heritability of RBD and daytime sleep features in this analysis may indicate that additional factors modify genetic predispositions to sleep disturbance among PD patients.

Phenotype severity therefore shows highly variable heritability in this analysis. Some phenotypes may be almost as heritable as PD onset overall, yet others appear to be largely independent of genetic factors. Only common SNPs are considered in this analysis due to constraints on sample size and statistical power. However rare variants could also explain a significant proportion of phenotypic variance. Some of the remaining variance not attributed to genetic factors in this analysis may therefore be linked to unmeasured rare variants, rather than environmental factors.

This could be an important consideration in future study as underlying molecular causes will be more easily identified from phenotypes with a greater genetic influence. Highly heritable phenotypes are also likely to be more amenable to consistent pharmacological intervention as less variability is attributable to unknown environmental factors. However this analysis provides

only estimates of minimum heritability and further study is required to quantify more accurately the true genetic components.

5.4.2) Continuous phenotypes

This analysis developed a novel method of quantifying diverse patient phenotypes on a continuous scale via the use of phenotype axes. This overcomes many of the limitations associated with the clustering methods previously used to classify PD heterogeneity. Clustering methods can fail to classify outlying individuals who then cannot be used in further analysis, however phenotypic axis scores can be calculated for every individual regardless of how rare their particular phenotypic presentation might be. Continuous variables also have greater statistical power than categorical variables to detect what are likely to be very small genetic effects. Finally phenotype axes are more biologically relevant, as they capture correlations between covarying phenotypes and quantify severity on a continuous scale that more accurately reflects patient variation. Consequently the whole spectrum of phenotypes displayed by each individual can be fully represented and interrogated.

A number of known comorbidities are represented among the phenotype axes. Anxiety and depression are highly correlated in PD patients, both of which are represented by axes 1 and 2 [371]. Rigidity and bradykinesia are also linked, possibly due to shared physiology [372], and vary in the same direction along axis 3. Finally tremor severity is thought to be relatively independent of most other phenotypes, and is represented in almost complete isolation by axis 4 and opposing other phenotypes in axes 5 and 6 [373]. Agreement of the phenotype

axes with previously observed correlations further evidences their relevance to underlying biological themes.

Several phenotype axes also align with patient clusters identified during previous analyses of PD heterogeneity. Almost every analysis performed to date has defined a tremor-dominant group of patients and this is comparable to axis 4, which reflects tremor severity almost completely independently of all other symptoms [247-251]. Axes 5 and 6 isolate posture and gait phenotypes from many other components, but increased severity of both is associated with an opposing reduction in tremor severity, supporting the separate classification of tremor-dominant and postural-instability gait disorder patients initially proposed by Jankovic et al. [247]. Many analyses have found a phenotype group associated with severe motor and non-motor symptoms [248-250, 252]. This could represent an extreme end of axis 1 as it is associated with worsening anxiety, depression and motor phenotypes, although cognitive impairment improves. The phenotype axes therefore reflect the findings of previous classification systems, whilst surpassing them in clinical relevance and statistical power for downstream analysis.

Quantitative trait GWAS analysis using these axes identified 59 variants within 10 genomic regions significantly affecting the severity of specific measured phenotypes. This gives novel insight into the molecular mechanisms underlying their onset and severity. Patterns of association also identify phenotypes for which these mechanisms are shared, highlighting those that could be responsive to common treatment.

rs55754428 was associated with anxiety, depression and motor symptoms relating to rigidity and the lower body. This SNP is an intron variant of PTPRD, a gene encoding a protein tyrosine phosphatase that is linked to restless legs syndrome (RLS) [374]. This disorder is a common comorbidity among PD patients and is associated with increased incidence of anxiety and depression [375]. A correlation between the severity of RLS, anxiety and depression is also observed [375]. Epidemiological evidence therefore supports a role of this variant in moderating the severity of linked lower body motor and neuropsychiatric symptoms.

3:194252539 affected the largest number of phenotypes, increasing the severity of whole-body motor and non-motor symptoms including depression, fatigue, pain and bradykinesia. Correspondingly activities of daily life were also impacted. This SNP is located downstream of the pseudogene RNU6-1101P. The function of this gene is unknown so it is unclear how this variant might link to phenotype severity.

The small association peak centred on rs73696270 and the single variant rs77151048 lie in intergenic regions. Both were associated with the severity of anxiety, depression and pain. Anxiety and depression are highly comorbid among PD patients and individuals with both experience greater severity of pain, pain disability and pain interference than those with one or neither [376]. In addition anxiety and depression are more severe among PD patients with pain than controls with pain, indicating that the link between these phenotypes is specific to PD [377]. The variants highlighted in this analysis affect the severity

of all three phenotypes in the same direction and may therefore begin to explain their co-occurrence.

rs55819636 was associated with UPDRS I which concerns mentation, behaviour and mood. It lies between TRAJ53 and TRAJ54, which encode different T cell receptor alpha joining segments. Combinations of these along with additional segments make up different T cell receptors and contribute towards their diversity. These receptors then bind to antigens presented on MHC molecules. MHC molecules are encoded by the HLA region which has been repeatedly linked to PD in GWAS studies. In addition to increasing risk of disease onset, immune-linked SNPs may therefore also moderate aspects of disease progression. However this variant was not associated with any other phenotype measures, indicating that it may cause a more general alteration of disease course rather than the onset of specific phenotypes.

rs12435357 was associated with the severity of whole-body rigidity and bradykinesia independently of tremor. It is an intron variant of RGS6, a regulator of G-protein signalling. This gene is required for the adult maintenance of dopaminergic neurons in the substantia nigra and it's absence is associated with altered morphology and degeneration of these cells [378]. It is also involved in GABA signalling [379]. Mouse knockout models of RGS6 demonstrate impaired motor coordination, specifically abnormal gait that affects forelimbs and hind limbs [379]. Physiologic evidence therefore supports a role for this gene in the whole-body motor phenotypes to which it was linked in this analysis, indicating that the mechanism by which neurodegeneration occurs may affect the particular phenotypes developed by an individual.

The largest association peak was centred on rs80346154 in an intergenic region. These were the only variants to be associated with apathy. This phenotype is distinct from other psychiatric and personality symptoms and correspondingly variants associated with anxiety or depression in this analysis didn't overlap with this region [380]. Apathy is thought to arise from alterations in the mesocorticolimbic dopaminergic pathway that impact incentive processing [381], but how the genetic variants implicated in this analysis might affect this or other pathways is unclear.

Axes 3, 4 and 6 all identified variants affecting the severity of tremor phenotypes. Rs115184951 lies within an intron of BMP8B, which encodes a ligand of the TGF-beta protein superfamily involved in bone morphogenic protein (BMP)-signalling. Rs76433669 is upstream of FXR2, an RNA-binding protein which regulates adult neurogenesis via interaction with Noggin. Rs7200759 is an intronic variant of WWOX, a gene involved in cell signalling, lipid metabolism, gene transcription and cell death.

These genes all converge on the TGF-beta signalling pathway, a process responsible for neuronal differentiation and apoptosis. BMP8 protein binds directly to TGF-beta, activating numerous downstream effectors which are required for the induction of dopaminergic neuron development and the TH-positive phenotype [382]. It also acts directly on dopaminergic neurons to promote their survival either independently or concomitantly with other neurotrophic molecules [383].

FXR2 promotes neurogenesis via interactions with Noggin within this pathway. Noggin activity inhibits BMP-signalling by binding to TGF-beta cytokines [384]. FXR2 inhibits Noggin, allowing controlled neurogenesis to take place [385]. Correspondingly FXR2 deficiency causes increased Noggin expression and suppresses TGF-beta-signalling, resulting in increased proliferation and altered fate specification of developing cells [385].

WWOX promotes differentiation and blocks degeneration also via the TGF-beta pathway. WWOX and Smad4 are recruited to the nucleus by the binding of TGF-beta to Hyal-2 [386]. Here the WWOX-containing complex mediates gene transcription and is also involved with apoptosis. [386] Consequently loss of WOXX causes inhibited neurite outgrowth which impairs neuronal differentiation [387], in addition to increased cell death [388].

All genetic variants linked to tremor converging on the TGF-beta signalling pathway provides compelling evidence for its role in moderating the severity of this phenotype. This direct link between a physiologic cause of neurodegeneration and a specific phenotype within PD adds further evidence to that associated with rs12435357 and rs55819636 that the mechanism by which neurodegeneration occurs could moderate which phenotypes an individual develops. Additive effects were not investigated here, but would provide an interesting avenue for study in an independent cohort to explore possible associations between cumulative mutational burden and tremor severity.

Interestingly genetic variants linked to tremor in this analysis had little or no effect on other phenotypes. Tremor severity may therefore be largely independent of other symptoms. This would align with cluster analyses of PD

heterogeneity whereby a “tremor-dominant” group of individuals is often isolated from the remaining PD population. Patients experiencing severe tremor may therefore be both physiologically and genetically distinct.

None of the phenotype-linked genetic variants in this analysis overlap with any known pathogenic PD regions. This could indicate that mechanisms of disease progression are separate from those involved in onset. However many variants show possible links to pathways directly involved in PD pathology and neurodegeneration, indicating that they could instead represent novel causes of disease that have not yet been identified. This would also align with the association of pathogenic genetic variants with characteristic patterns of phenotypes. The specific molecular causes of disease onset could therefore determine the particular phenotypes an individual develops and moderate their severity.

The genetic associations identified through this analysis demonstrate the value of collapsing numerous phenotypic measurements into representative axes of variation. If all SNPs had been tested for association with all phenotypes the required multiple testing corrections would have been prohibitive of any significant finding. This method allowed a hypothesis-free approach to be maintained whilst avoiding millions of redundant tests. Consequently statistical power was maximised without limiting which associations could be discovered.

5.4.3) Genetic models of phenotype severity

Having identified a genetic component moderating phenotype severity this analysis next attempted to create genetic models of phenotypic axis score. Both lasso regression and polygenic scoring were employed with mixed results. This

gives insight into the likely genetic architecture of phenotypic traits and the feasibility of predicting disease course from genotype at the time of onset.

Models derived from lasso regression looked promising in the initial test sample, however these scores were poor predictors of phenotypic severity in a second independent sample. They were also unable to separate individuals with extremely mild and extremely severe phenotypes. Consequently these scores appeared to be unrelated to phenotypic severity in the general PD population. This indicated that they were highly over-fitted to the original training set.

This could have been due to a number of factors within the GWAS filtering or lasso steps. The SNPs selected from GWAS P values were highly inconsistent. Small sample size can both inflate the P value of neutral SNPs by chance and underestimate the significance of genuine associations through lack of statistical power. Consequently the set of SNPs inputted to lasso regression may have contained a high proportion of neutral variants whilst missing some effect SNPs. Relaxing the P value threshold for inclusion should have alleviated the latter problem but the performance of the genetic scores was not improved, indicating that a high proportion of neutral variants is likely to be the confounding factor.

Over-fitting may also have occurred due to the ratio of predictor variables to samples. There were over 10-fold more SNPs than patient samples in this analysis, so neutral SNPs could have explained a large proportion of phenotypic variance by random chance alone. This problem would have been exacerbated by the input of more neutral SNPs into the regression and would explain why more lenient P value thresholds resulted in worse genetic predictors. It is therefore

likely that over-fitting of the lasso regression was the main cause of failure in these models.

Polygenic scores were calculated in an attempt to avoid this over-fitting. Models were generated for a range of P value thresholds. Previous analyses show that polygenic scores generated using only genome-wide significant variants usually perform poorly, whereas the inclusion of additional SNPs under a more lenient P value threshold improves performance. Similarly this analysis shows a trend between more SNPs forming the model and a greater proportion of phenotypic variance explained. This indicates a highly polygenic architecture defining phenotypic severity.

Although polygenic scores produced better models than lasso regression, the correlation between observed and predicted phenotype score for the whole population was still only moderate. Segregation between individuals with the most mild and severe phenotypes based on genetic score was relatively good. AUC values of over 0.75 are considered to be clinically useful for screening at-risk individuals and this was observed for half of the phenotypic axes [389]. For axis 5 the maximum possible AUC value of 1 was observed, indicating that polygenic scores separated perfectly patients with mild and severe phenotypes, although this value would likely decrease slightly in a larger replication cohort. These results indicate that although environmental factors seem to largely determine intermediate variation in phenotypes, genetic load can be used to stratify the population into low and high severity categories for some phenotypic axes.

However the AUC statistic can mask the true usefulness of genetic scores for predicting phenotype severity. A relatively high AUC of 0.8 was observed for component 2. However in order to achieve a specificity of 1, sensitivity is only 0.6. This means that to define a cut-off for genetic scores that selects only individuals with severe phenotypes and no others, just 60% will actually be identified. Conversely a cut-off that identifies all individuals with severe phenotypes will incorrectly classify 80% of individuals with mild phenotypes into this category. Specificity and sensitivity will be further worsened by the inclusion of individuals with intermediate phenotype scores, rather than just the population extremes tested in this analysis. Consequently using genetic scores to predict phenotype severity, even with a classifier that statistically performs relatively well, results in a large number of false-positive and false-negative classifications that would render these scores almost useless in a clinical setting.

As two extremes of a continuum are being compared rather than a binary variable, two separate thresholds could be implemented to improve these errors: one above which individuals are likely to suffer from severe phenotypes and one below which individuals are likely to suffer from mild phenotypes. This should provide better control over specificity and sensitivity. However there was notable overlap between genetic scores of individuals in the highest and lowest 10^{th} percentiles of phenotypic severity. Consequently the scores of these individuals are likely to overlap greatly with those of the rest of the population displaying intermediate phenotypes. The chance of defining thresholds that usefully separate individuals with particularly mild and severe phenotypes from the entire PD population is therefore low.

5.4.4) Study limitations

As mentioned previously sample size is a substantial limiting factor in this analysis. This is likely to have caused the over-fitting of genetic models as the ratio of samples to variables was very low. However due to the time constraints and logistics associated with the collection of such in-depth phenotypic data this is not easily rectified.

In this study data was not collected for environmental variables. Factors such as caffeine intake significantly affect PD risk, so could conceivably affect disease progression also. Furthermore SNP heritability may only account for half of total heritability [390]. Consequently this analysis could explore only a fraction of the total factors affecting phenotype. The increasing popularity of whole-genome sequencing may partially alleviate this problem in future, but the inaccuracy of quantifying environmental factors makes this a difficult component to incorporate in any analysis.

The relatively confined geographical area from which this cohort was drawn is likely to have hindered the development of widely generalizable genetic models. This is because nation- or world-wide genetic variance simply was not present in the discovery dataset. Although pathogenic alleles are usually universally detrimental their frequencies can vary widely between populations. Alleles of smaller and less significant effect vary even more so as they are under less selective pressure. Consequently highly polygenic traits such as PD sub-phenotype severity are likely to be moderated by different sets of variants among diverse populations. It is expected that the results of this analysis should generalise well to the white British population, however without replication sub-

population specific effects cannot be excluded. It is unknown how far beyond the white British population these results may be applicable.

The investigation of genetic risk scores requires one cohort for calculating genetic associations and one cohort for testing. The best arrangement for this consists of two completely independent datasets as the division of a one into a training and test set has comparatively less statistical power [391]. The set-up of this analysis was therefore sub-optimal for investigating genetic scores as predictors of phenotype, as one cohort was split to provide the necessary subsets. At the time this was carried out no other dataset existed with comparable phenotypic data so this was unavoidable.

5.4.5) Conclusion

This analysis forms the first hypothesis-free, genome-wide investigation into the effect of genotype on clinical phenotype. Through the generation of phenotype axes it is shown that the severity of multiple phenotypes is correlated and therefore that different aspects of disease progression are linked. Some of these phenotypes are affected by the same genetic variants, implicating common molecular mechanisms underlying their severity.

In total ten genetic regions were associated with a range of motor and non-motor phenotypes. This provides the first evidence that non-pathogenic variation can affect phenotypic severity in PD, providing novel insight into mechanisms that may underlie heterogeneous disease progression. Furthermore links between these variants and neurodegeneration indicate that phenotype onset may be moderated by the specific molecular cause of PD.

Although genetic scores could not stratify patients with particularly mild or severe phenotypes to a clinically useful degree, they provide some evidence that extremely low or high genetic load might contribute to phenotype severity. This indicates a possible role for additive effects and implies a complex genetic architecture underlying these traits. It is therefore likely that the variants identified in this analysis constitute only a minority of the total factors influencing phenotype severity. Further research is required to elucidate the many additional components affecting disease progression.

These results could have a significant impact on pharmacological intervention. The identification of more precise molecular mechanisms underlying disease subtypes could allow the development of drugs that target specific pathways. This would maximise efficacy whilst reducing the off-target effects associated with more generic treatments. Furthermore those drugs can then be administered to the patients for whom they will be most effective, reducing the time and cost associated with a trial-and-error approach to PD treatment. It is therefore hoped that this work will provide a platform for further study on PD genotype-phenotype links, in order to aid the understanding of the disease process and promote the development and targeting of more specific treatments.

6) Concluding remarks

Before this work was carried out, most research concerning genetic causes of PD consisted of familial segregation or genome-wide association studies comparing case and control. Around 30 causal and risk variants had been identified, giving insight into several molecular mechanisms involved in PD pathology. However there remained a significant proportion of heritability unexplained. Furthermore, genetic causes of heterogeneous disease progression were almost completely unknown. Phenotypic variability among patients had been explored via the identification of phenotype clusters but little analysis had been carried out investigating genetic differences between them. Consequently the genetic causes and corresponding molecular mechanisms underlying sporadic PD onset and progression were for the most part unknown. This work attempted to alleviate some of this deficit and extend the understanding of the disease process by exploring the influence of genetic variation on PD onset and progression.

It is shown that variants implicated by GWAS may mediate disease risk through nearby eQTLs and the corresponding changes in expression of cis- and trans-genes. Variation within four GWAS regions was associated with expression changes in multiple genes linked to PD phenotypes, including PDE10A, NPY2R, FMO3, ARSA, LRRK2, CISD1 and PPARGC1A. The identified trans-effects had not previously been associated with these regions and therefore provide novel means by which disease onset may be conveyed. By affecting multiple genes the effects of single genetic risk variants may be multi-faceted, which could begin to

explain the diversity of the cellular processes altered in PD pathology. This may also contribute towards the development of diverse symptoms and could provide links between correlated aspects of disease progression.

Two GWAS regions contained cis-eQTLs of LRRK2 and the genotype associated with highest additive increase in expression was enriched among two independent PD case cohorts. Over-expression of LRRK2 is associated with cellular PD phenotypes in model organisms but this work provides the first indication that it may also increase PD risk in the general population. This demonstrates that LRRK2 genetic risk factors are not limited to those that modify protein structure, but also include those affecting wild-type gene expression.

In addition to implicating variants acting in isolation, this work suggests that both SNP and CNV variants converge on common molecular pathways to moderate disease onset. Biological mechanisms enriched among PD-associated variants include *nuclear heterochromatin* and *nucleotide binding* among SNPs and *abnormal mitochondrial morphology* among CNVs. Genes whose mouse orthologue's disruption was associated with abnormal exploration of a new environment were also over-represented among CNVs, but this behavioural annotation demonstrates only that mouse and human genes cause similar neurological phenotypes and gives no insight into molecular underpinnings.

Alterations to heterochromatin structure are associated with normal aging. Acceleration of this process has previously been linked to neurodegeneration and Alzheimer's Disease, but this analysis highlights a novel link to PD. Heterochromatin conformation represses gene transcription, so

associated genetic variants may cause structural changes that result in erroneous gene expression; increasing cell stress and degeneration. This analysis has shown that aberrant expression of wild-type LRRK2 may increase PD risk and an effect of dosage on PD risk is also observed for other genes such as SNCA. Consequently genetic variants affecting nuclear heterochromatin may increase the risk of neurodegeneration generally, whereas the comparative risk of PD and other disorders may be mediated by the particular genes whose transcription is affected.

Altered nucleotide binding is responsible for the toxicity of several pathogenic mutations, but this analysis shows it may also have a role in sporadic PD. Nucleotides perform several diverse functions so may moderate PD risk in a number of ways. However centrally these molecules are involved in intracellular signalling and energy metabolism. Alterations in energetic pathways are highly associated with mitochondrial phenotypes, which are known to be a primary cause of PD pathology. Genetic mutations affecting nucleotide binding could therefore contribute to the bioenergetic alterations observed in both sporadic and familial PD. Evidence also suggests that CNV variation may affect mitochondrial morphology and consequently energy production, although lack of replication indicates these variants are likely to cause a minority of PD cases. These results provide evidence that the bioenergetic defects associated with PD onset may arise from numerous genetic mutations which are not of genome-wide significance individually, and that although multiple structural mutations may contribute to this the majority are likely to be single nucleotide variants.

To explore causes of phenotypic heterogeneity the first genome-wide association study of phenotypic subgroups was carried out. Several SNPs demonstrated strong association with patient clusters defined by tremor-dominant phenotypes and severe motor and non-motor phenotypes. All of these variants are associated with neuronal function, so although not quite reaching genome-wide significance they are highly likely to represent genuine effect SNPs. This provides evidence that genetic variants may be at least partially responsible for determining phenotypic progression.

This analysis also showed that genetic variants associated with the severe motor and non-motor phenotype cluster converged on guanyl-nucleotide exchange factor activity. This provides the first evidence that perturbations of specific molecular pathways may cause the onset of characteristic groups of phenotypes. Interestingly this annotation is also a subset of the nucleotide binding term implicated in overall disease onset. It is therefore proposed that broadly similar molecular mechanisms underlie PD onset for the whole population, but that specific processes beneath that umbrella cause the development of particular phenotypes.

Genetic variants were also implicated in mediating the severity of individual phenotypes. This provides novel insight into the molecular mechanisms underlying their development and identifies linked phenotypes for which these mechanisms are shared. No variants overlap with known pathogenic PD genes but several are linked to the same pathways, such as immunity and neuronal differentiation. This provides additional support for phenotype severity being moderated by the specific mechanisms causing PD onset.

Interestingly, tremor severity appears to be largely independent of other phenotypes in both degree of severity and genetic causes. Previous analyses of phenotypic heterogeneity in PD have generally identified a tremor-dominant subgroup, and in this analysis phenotype axis 4 represented tremor severity almost exclusively. Genetic analysis identified three regions significantly associated with tremor severity, yet none of these affected other phenotypes. All these variants converge on the TGF-beta signalling pathway, a process responsible for neuronal differentiation and apoptosis. Both phenotypic and genotypic evidence therefore point toward tremor phenotypes being largely independent of other phenotypes and mediated by distinct molecular mechanisms.

Overall it is proposed that these results indicate that similar molecular mechanisms cause disease onset for the whole PD population, but alterations of specific pathways are responsible for the onset of particular phenotypes. Unfortunately there was no opportunity to confirm these findings in an independent cohort, but if replicated this could identify a crucial component to be considered in the symptomatic treatment of PD. At present Levodopa treats the majority of PD symptoms relatively well, but effects are inconsistent and usually last for less than a decade. Results from this analysis indicate that more specific targeted treatments administered to patients displaying particular phenotypes could achieve much greater efficacy, likely with fewer undesirable side-effects. This could provide the first step towards the coveted goal of personalised medicine.

In order for this to be realised, a robust and relevant method of quantifying patient phenotype is required. Many studies in the past few decades have attempted to define phenotypic subgroups to classify patient heterogeneity. In this analysis a novel method of classification was developed that quantified phenotype severity on a continuous scale. This overcame most of the limitations associated with previous methods, whilst better representing correlations between phenotypes and the spectrum of patient severity observed in the clinic. The most consistent findings of previous analyses were reflected in the results of this hypothesis-free approach, including the relative independence of tremor severity, the separation of posture and gait phenotypes and the segregation of individuals with severe motor and non-motor symptoms. However this method provides phenotypic measures that surpass binary traits in biological relevance and statistical power, paving the way for identification of genetic variants and molecular pathways affecting phenotype severity. It is hoped that in future this will aid the development and targeting of more specific symptomatic treatment.

Overall this analysis has implicated several novel genetic mechanisms in mediating the onset and progression of PD. It has provided evidence for alternative methods by which the effects of known risk variants are conveyed and identified common pathways on which previously unassociated genetic variants converge to cause PD onset. Furthermore the first evidence is provided for the mediation of phenotype onset by non-pathogenic variants and distinct molecular pathways, showing that the process by which neurodegeneration occurs could be crucial in determining the development and severity of phenotypes during disease course. A novel phenotype classification system was

developed that provides a more accurate and biologically relevant method of quantifying patient heterogeneity by incorporating continuous measures of severity and correlations between phenotypes, providing a superior method by which these genetic findings can be confirmed and extended. It is hoped that the results obtained in this study will inspire future work exploring genetic and molecular causes of phenotypic heterogeneity in order to develop more effective and targeted PD treatments.

7) Bibliography

1. Gustavsson, A., et al., *Cost of disorders of the brain in Europe 2010.* European Neuropsychopharmacology, 2011. **21**(10): p. 718-779.
2. Fearnley, J.M. and A.J. Lees, *AGING AND PARKINSONS-DISEASE - SUBSTANTIA-NIGRA REGIONAL SELECTIVITY.* Brain, 1991. **114**: p. 2283-2301.
3. de la Fuente-Fernandez, R., et al., *Age-Specific Progression of Nigrostriatal Dysfunction in Parkinson's Disease.* Annals of Neurology, 2011. **69**(5): p. 803-810.
4. Oerlemans, W.G.H. and A.W. de Weerd, *The prevalence of sleep disorders in patients with Parkinson's disease A self-reported, community-based survey.* Sleep Medicine, 2002. **3**(2): p. 147-149.
5. Riedel, O., et al., *Cognitive impairment in 873 patients with idiopathic Parkinson's disease - Results from the German Study on epidemiology of Parkinson's disease with dementia (GEPAD).* Journal of Neurology, 2008. **255**(2): p. 255-264.
6. Postuma, R.B., et al., *Parkinson risk in idiopathic REM sleep behavior disorder Preparing for neuroprotective trials.* Neurology, 2015. **84**(11): p. 1104-1113.
7. McCormack, A.L., et al., *Environmental risk factors and Parkinson's disease: Selective degeneration of nigral dopaminergic neurons caused by the herbicide paraquat.* Neurobiology of Disease, 2002. **10**(2): p. 119-127.
8. Bowenkamp, K.E., et al., *6-Hydroxydopamine induces the loss of the dopaminergic phenotype in substantia nigra neurons of the rat - A possible mechanism for restoration of the nigrostriatal circuit mediated by glial cell line-derived neurotrophic factor.* Experimental Brain Research, 1996. **111**(1): p. 1-7.
9. Bus, J.S., S.D. Aust, and J.E. Gibson, *PARAQUAT TOXICITY - PROPOSED MECHANISM OF ACTION INVOLVING LIPID PEROXIDATION.* Environmental Health Perspectives, 1976. **16**(AUG): p. 139-146.
10. Rodriguez-Pallares, J., et al., *Mechanism of 6-hydroxydopamine neurotoxicity: the role of NADPH oxidase and microglial activation in 6-hydroxydopamine-induced degeneration of dopaminergic neurons.* Journal of Neurochemistry, 2007. **103**(1): p. 145-156.
11. Riederer, P., et al., *TRANSITION-METALS, FERRITIN, GLUTATHIONE, AND ASCORBIC-ACID IN PARKINSONIAN BRAINS.* Journal of Neurochemistry, 1989. **52**(2): p. 515-520.
12. Sofic, E., et al., *INCREASED IRON(III) AND TOTAL IRON CONTENT IN POST-MORTEM SUBSTANTIA NIGRA OF PARKINSONIAN BRAIN.* Journal of Neural Transmission, 1988. **74**(3): p. 199-205.
13. Saggu, H., et al., *A SELECTIVE INCREASE IN PARTICULATE SUPEROXIDE-DISMUTASE ACTIVITY IN PARKINSONIAN SUBSTANTIA NIGRA.* Journal of Neurochemistry, 1989. **53**(3): p. 692-697.
14. Sian, J., et al., *ALTERATIONS IN GLUTATHIONE LEVELS IN PARKINSONS-DISEASE AND OTHER NEURODEGENERATIVE DISORDERS AFFECTING BASAL GANGLIA.* Annals of Neurology, 1994. **36**(3): p. 348-355.
15. Dexter, D.T., et al., *INCREASED LEVELS OF LIPID HYDROPEROXIDES IN THE PARKINSONIAN SUBSTANTIA-NIGRA - AN HPLC AND ESR STUDY.* Movement Disorders, 1994. **9**(1): p. 92-97.
16. Alam, Z.I., et al., *Oxidative DNA damage in the parkinsonian brain: An apparent selective increase in 8-hydroxyguanine levels in substantia nigra.* Journal of Neurochemistry, 1997. **69**(3): p. 1196-1203.

17. Alam, Z.I., et al., *A generalised increase in protein carbonyls in the brain in Parkinson's but not incidental Lewy body disease.* Journal of Neurochemistry, 1997. **69**(3): p. 1326-1329.
18. Dexter, D.T. and P. Jenner, *Parkinson disease: from pathology to molecular disease mechanisms.* Free Radical Biology and Medicine, 2013. **62**: p. 132-144.
19. Dagda, R.K., et al., *Loss of PINK1 Function Promotes Mitophagy through Effects on Oxidative Stress and Mitochondrial Fission.* Journal of Biological Chemistry, 2009. **284**(20): p. 13843-13855.
20. Santos, D., et al., *The Impact of Mitochondrial Fusion and Fission Modulation in Sporadic Parkinson's Disease.* Molecular Neurobiology, 2015. **52**(1): p. 573-586.
21. Trimmer, P.A., et al., *Abnormal mitochondrial morphology in sporadic Parkinson's and Alzheimer's disease cybrid cell lines.* Experimental Neurology, 2000. **162**(1): p. 37-50.
22. Exner, N., et al., *Loss-of-function of human PINK1 results in mitochondrial pathology and can be rescued by parkin.* Journal of Neuroscience, 2007. **27**(45): p. 12413-12418.
23. Keeney, P.M., et al., *Parkinson's disease brain mitochondrial complex I has oxidatively damaged subunits and is functionally impaired and misassembled.* Journal of Neuroscience, 2006. **26**(19): p. 5256-5264.
24. Mann, V.M., et al., *BRAIN, SKELETAL-MUSCLE AND PLATELET HOMOGENATE MITOCHONDRIAL-FUNCTION IN PARKINSONS-DISEASE.* Brain, 1992. **115**: p. 333-342.
25. Schapira, A.H.C., J. M.; Dexter, D.; Clark, J. B.; Jenner, P.; Marsden, C.D., *Mitochondrial complex I deficiency in Parkinson's disease.* J Neurochem, 1990. **54**(3): p. 823-827.
26. Mortiboys, H., et al., *Mitochondrial Function and Morphology Are Impaired in parkin-Mutant Fibroblasts.* Annals of Neurology, 2008. **64**(5): p. 555-565.
27. Devi, L., et al., *Mitochondrial import and accumulation of alpha-synuclein impair complex I in human dopaminergic neuronal cultures and Parkinson disease brain.* Journal of Biological Chemistry, 2008. **283**(14): p. 9089-9100.
28. Bajpai, P., et al., *Metabolism of 1-Methyl-4-phenyl-1,2,3,6-tetrahydropyridine by Mitochondrion-targeted Cytochrome P450 2D6 IMPLICATIONS IN PARKINSON DISEASE.* Journal of Biological Chemistry, 2013. **288**(6): p. 4436-4451.
29. Richardson, J.R., et al., *Obligatory role for complex I inhibition in the dopaminergic neurotoxicity of 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP).* Toxicological Sciences, 2007. **95**(1): p. 196-204.
30. Li, N.Y., et al., *Mitochondrial complex I inhibitor rotenone induces apoptosis through enhancing mitochondrial reactive oxygen species production.* Journal of Biological Chemistry, 2003. **278**(10): p. 8516-8525.
31. Kussmaul, L. and J. Hirst, *The mechanism of superoxide production by NADH : ubiquinone oxidoreductase (complex I) from bovine heart mitochondria.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(20): p. 7607-7612.
32. Jin, S.M., et al., *Mitochondrial membrane potential regulates PINK1 import and proteolytic destabilization by PARL.* Journal of Cell Biology, 2010. **191**(5): p. 933-942.
33. Greene, A.W., et al., *Mitochondrial processing peptidase regulates PINK1 processing, import and Parkin recruitment.* Embo Reports, 2012. **13**(4): p. 378-385.
34. Meissner, C., et al., *The mitochondrial intramembrane protease PARL cleaves human Pink1 to regulate Pink1 trafficking.* Journal of Neurochemistry, 2011. **117**(5): p. 856-867.

35. Cookson, M.R., *Parkinsonism Due to Mutations in PINK1, Parkin, and DJ-1 and Oxidative Stress and Mitochondrial Pathways.* Cold Spring Harbor Perspectives in Medicine, 2012. **2**(9).
36. Geisler, S., et al., *PINK1/Parkin-mediated mitophagy is dependent on VDAC1 and p62/SQSTM1.* Nature Cell Biology, 2010. **12**(2): p. 119-U70.
37. Taira, T., et al., *DJ-1 has a role in antioxidative stress to prevent cell death.* Embo Reports, 2004. **5**(2): p. 213-218.
38. Guzman, J.N., et al., *Oxidant stress evoked by pacemaking in dopaminergic neurons is attenuated by DJ-1.* Nature, 2010. **468**(7324): p. 696-U119.
39. De Marco, E.V., et al., *DJ-1 is a Parkinson's disease susceptibility gene in southern Italy.* Clinical Genetics, 2010. **77**(2): p. 183-188.
40. Power, J.H.T.B., O. L.; Chegini, F., *Lewy Bodies and the Mechanisms of Neuronal Cell Death in Parkinson's Disease and Dementia with Lewy Bodies.* Brain Pathology, 2016.
41. Muller, S.K., et al., *Lewy body pathology is associated with mitochondrial DNA damage in Parkinson's disease.* Neurobiology of Aging, 2013. **34**(9): p. 2231-2233.
42. Hansen, C., et al., *alpha-Synuclein propagates from mouse brain to grafted dopaminergic neurons and seeds aggregation in cultured human cells.* Journal of Clinical Investigation, 2011. **121**(2): p. 715-725.
43. Luk, K.C., et al., *Exogenous alpha-synuclein fibrils seed the formation of Lewy body-like intracellular inclusions in cultured cells.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(47): p. 20051-20056.
44. Paik, S.R., et al., *Copper(II)-induced self-oligomerization of alpha-synuclein.* Biochemical Journal, 1999. **340**: p. 821-828.
45. Guo, Y.J. and S. Scarlata, *A Loss in Cellular Protein Partners Promotes alpha-Synuclein Aggregation in Cells Resulting from Oxidative Stress.* Biochemistry, 2013. **52**(22): p. 3913-3920.
46. Lee, H.J., et al., *Formation and removal of alpha-synuclein aggregates in cells exposed to mitochondrial inhibitors.* Journal of Biological Chemistry, 2002. **277**(7): p. 5411-5417.
47. Mirzaei, H., et al., *Identification of rotenone-induced modifications in alpha-synuclein using affinity pull-down and tandem mass spectrometry.* Analytical Chemistry, 2006. **78**(7): p. 2422-2431.
48. Hodara, R., et al., *Functional consequences of alpha-synuclein tyrosine nitration - Diminished binding to lipid vesicles and increased fibril formation.* Journal of Biological Chemistry, 2004. **279**(46): p. 47746-47753.
49. Giasson, B.I., et al., *Oxidative damage linked to neurodegeneration by selective alpha-synuclein nitration in synucleinopathy lesions.* Science, 2000. **290**(5493): p. 985-989.
50. Larsen, K.E., et al., *alpha-synuclein overexpression in PC12 and chromaffin cells impairs catecholamine release by interfering with a late step in exocytosis.* Journal of Neuroscience, 2006. **26**(46): p. 11915-11922.
51. Oaks, A.W., et al., *Synucleins Antagonize Endoplasmic Reticulum Function to Modulate Dopamine Transporter Trafficking.* Plos One, 2013. **8**(8).
52. Gosavi, N., et al., *Golgi fragmentation occurs in the cells with prefibrillar alpha-synuclein aggregates and precedes the formation of fibrillar inclusion.* Journal of Biological Chemistry, 2002. **277**(50): p. 48984-48992.
53. Cooper, A.A., et al., *alpha-synuclein blocks ER-Golgi traffic and Rab1 rescues neuron loss in Parkinson's models.* Science, 2006. **313**(5785): p. 324-328.
54. Smith, W.W., et al., *Endoplasmic reticulum stress and mitochondrial cell death pathways mediate A53T mutant alpha-synuclein-induced toxicity.* Human Molecular Genetics, 2005. **14**(24): p. 3801-3811.

55. Lindersson, E., et al., *Proteasomal inhibition by alpha-synuclein filaments and oligomers.* Journal of Biological Chemistry, 2004. **279**(13): p. 12924-12934.
56. Emmanouilidou, E., L. Stefanis, and K. Vekrellis, *Cell-produced alpha-synuclein oligomers are targeted to, and impair, the 26S proteasome.* Neurobiology of Aging, 2010. **31**(6): p. 953-968.
57. Stefanis, L., et al., *Expression of A53T mutant but not wild-type alpha-synuclein in PC12 cells induces alterations of the ubiquitin-dependent degradation system, loss of dopamine release, and autophagic cell death.* Journal of Neuroscience, 2001. **21**(24): p. 9549-9560.
58. Snyder, H., et al., *Aggregated and monomeric alpha-synuclein bind to the S6 ' proteasomal protein and inhibit proteasomal function.* Journal of Biological Chemistry, 2003. **278**(14): p. 11753-11759.
59. Nakamura, K., et al., *Direct Membrane Association Drives Mitochondrial Fission by the Parkinson Disease-associated Protein alpha-Synuclein.* Journal of Biological Chemistry, 2011. **286**(23): p. 20710-20726.
60. Kamp, F., et al., *Inhibition of mitochondrial fusion by alpha-synuclein is rescued by PINK1, Parkin and DJ-1.* Embo Journal, 2010. **29**(20): p. 3571-3589.
61. Choubey, V., et al., *Mutant A53T alpha-Synuclein Induces Neuronal Death by Increasing Mitochondrial Autophagy.* Journal of Biological Chemistry, 2011. **286**(12): p. 10814-10824.
62. Xu, J., et al., *Dopamine-dependent neurotoxicity of alpha-synuclein: A mechanism for selective neurodegeneration in Parkinson disease.* Nature Medicine, 2002. **8**(6): p. 600-606.
63. Shimura, H., et al., *Ubiquitination of a new form of alpha-synuclein by parkin from human brain: Implications for Parkinson's disease.* Science, 2001. **293**(5528): p. 263-269.
64. Leroy, E., et al., *The ubiquitin pathway in Parkinson's disease.* Nature, 1998. **395**(6701): p. 451-452.
65. Elstner, M., et al., *Expression analysis of dopaminergic neurons in Parkinson's disease and aging links transcriptional dysregulation of energy metabolism to cell death.* Acta Neuropathologica, 2011. **122**(1): p. 75-86.
66. Grunblatt, E., et al., *Gene expression profiling of parkinsonian substantia nigra pars compacta; alterations in ubiquitin-proteasome, heat shock protein, iron and oxidative stress regulated proteins, cell adhesion/cellular matrix and vesicle trafficking genes.* Journal of Neural Transmission, 2004. **111**(12): p. 1543-1573.
67. Zhang, Y.L., et al., *Transcriptional analysis of multiple brain regions in Parkinson's disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms.* American Journal of Medical Genetics Part B-Neuropsychiatric Genetics, 2005. **137B**(1): p. 5-16.
68. McNaught, K.S., et al., *Altered proteasomal function in sporadic Parkinson's disease.* Experimental Neurology, 2003. **179**(1): p. 38-46.
69. McNaught, K.S.P. and P. Jenner, *Proteasomal function is impaired in substantia nigra in Parkinson's disease.* Neuroscience Letters, 2001. **297**(3): p. 191-194.
70. Sun, F., et al., *Proteasome inhibitor MG-132 induces dopaminergic degeneration in cell culture and animal models.* Neurotoxicology, 2006. **27**(5): p. 807-815.
71. Li, X.P., et al., *A Mechanistic Study of Proteasome Inhibition-Induced Iron Misregulation in Dopamine Neuron Degeneration.* Neurosignals, 2012. **20**(4): p. 223-236.
72. Xie, W.J., et al., *Proteasome inhibition modeling nigral neuron degeneration in Parkinson's disease.* Journal of Neurochemistry, 2010. **115**(1): p. 188-199.

73. Cuervo, A.M., et al., *Impaired degradation of mutant alpha-synuclein by chaperone-mediated autophagy.* Science, 2004. **305**(5688): p. 1292-1295.
74. Alvarez-Erviti, L., et al., *Chaperone-Mediated Autophagy Markers in Parkinson Disease Brains.* Archives of Neurology, 2010. **67**(12): p. 1464-1472.
75. Martinez-Vicente, M., et al., *Dopamine-modified alpha-synuclein blocks chaperone-mediated autophagy.* Journal of Clinical Investigation, 2008. **118**(2): p. 777-788.
76. Lynch-Day, M.A., et al., *The Role of Autophagy in Parkinson's Disease.* Cold Spring Harbor Perspectives in Medicine, 2012. **2**(4).
77. Usenovic, M.T., E.; Mazzulli, J.R.; Taylor, J.P.; Krainc, D., *Deficiency of ATP13A2 leads to lysosomal dysfunction, alpha-synuclein accumulation, and neurotoxicity.* Journal of Neuroscience, 2012. **32**: p. 4240-4246.
78. Dehay, B.R., A.; Martinez-Vicente, M.; Perier, C.; Canron, M.H.; Doudnikoff, E.; Vital, A.; Vila, M.; Klein, C.; Bezard, E., *Loss of P-type ATPase ATP13A2/PARK9 function induces general lysosomal deficiency and leads to Parkinson disease neurodegeneration.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**: p. 9611-9616.
79. Sidransky, E., et al., *Multicenter Analysis of Glucocerebrosidase Mutations in Parkinson's Disease.* New England Journal of Medicine, 2009. **361**(17): p. 1651-1661.
80. Mazzulli, J.R., et al., *Gaucher Disease Glucocerebrosidase and alpha-Synuclein Form a Bidirectional Pathogenic Loop in Synucleinopathies.* Cell, 2011. **146**(1): p. 37-52.
81. Neumann, J., et al., *Glucocerebrosidase mutations in clinical and pathologically proven Parkinson's disease.* Brain, 2009. **132**: p. 1783-1794.
82. Gegg, M.E., et al., *Glucocerebrosidase deficiency in substantia nigra of parkinson disease brains.* Annals of Neurology, 2012. **72**(3): p. 455-463.
83. Anglade, P., et al., *Apoptosis and autophagy in nigral neurons of patients with Parkinson's disease.* Histology and Histopathology, 1997. **12**(1): p. 25-31.
84. Hamza, T.H., et al., *Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease.* Nature Genetics, 2010. **42**(9): p. 781-U75.
85. McGeer, P.L., et al., *REACTIVE MICROGLIA ARE POSITIVE FOR HLA-DR IN THE SUBSTANTIA NIGRA OF PARKINSONS AND ALZHEIMERS-DISEASE BRAINS.* Neurology, 1988. **38**(8): p. 1285-1291.
86. Shavali, S., C. Combs, and M. Ebadi, *Reactive macrophages increase oxidative stress and alpha-synuclein nitration during death of dopaminergic neuronal cells in co-culture: Relevance to Parkinson's disease.* Neurochemical Research, 2006. **31**(1): p. 85-94.
87. Scalzo, P., et al., *Increased serum levels of soluble tumor necrosis factor-alpha receptor-1 in patients with Parkinson's disease.* Journal of Neuroimmunology, 2009. **216**(1-2): p. 122-125.
88. Scalzo, P., et al., *Serum levels of interleukin-6 are elevated in patients with Parkinson's disease and correlate with physical performance.* Neuroscience Letters, 2010. **468**(1): p. 56-58.
89. Mogi, M., et al., *TUMOR-NECROSIS-FACTOR-ALPHA (TNF-ALPHA) INCREASES BOTH IN THE BRAIN AND IN THE CEREBROSPINAL-FLUID FROM PARKINSONIAN-PATIENTS.* Neuroscience Letters, 1994. **165**(1-2): p. 208-210.
90. Mogi, M., et al., *Interleukin (IL)-1 beta, IL-2, IL-4, IL-6 and transforming growth factor-alpha levels are elevated in ventricular cerebrospinal fluid in juvenile parkinsonism and Parkinson's disease.* Neuroscience Letters, 1996. **211**(1): p. 13-16.
91. BlumDegen, D., et al., *Interleukin-1 beta and interleukin-6 are elevated in the cerebrospinal fluid of Alzheimer's and de novo Parkinson's disease patients.* Neuroscience Letters, 1995. **202**(1-2): p. 17-20.

92. Tang, P., et al., *Correlation between Serum RANTES Levels and the Severity of Parkinson's Disease.* Oxidative Medicine and Cellular Longevity, 2014: p. 1-4.
93. Rentzos, M., et al., *Circulating interleukin-15 and RANTES chemokine in Parkinson's disease.* Acta Neurologica Scandinavica, 2007. **116**(6): p. 374-379.
94. Simunovic, F., et al., *Gene expression profiling of substantia nigra dopamine neurons: further insights into Parkinson's disease pathology.* Brain, 2009. **132**: p. 1795-1809.
95. Bellani, S., et al., *The regulation of synaptic function by alpha-synuclein.* Communicative & integrative biology, 2010. **3**(2): p. 106-9.
96. Lotharius, J. and P. Brundin, *Pathogenesis of Parkinson's disease: Dopamine, vesicles and alpha-synuclein.* Nature Reviews Neuroscience, 2002. **3**(12): p. 932-942.
97. Liu, S.M., et al., *alpha-synuclein produces a long-lasting increase in neurotransmitter release.* Embo Journal, 2004. **23**(22): p. 4506-4516.
98. Zhang, L., et al., *Semi-quantitative analysis of alpha-synuclein in subcellular pools of rat brain neurons: An immunogold electron microscopic study using a C-terminal specific monoclonal antibody.* Brain Research, 2008. **1244**: p. 40-52.
99. Liu, G., et al., *alpha-Synuclein is differentially expressed in mitochondria from different rat brain regions and dose-dependently down-regulates complex I activity.* Neuroscience Letters, 2009. **454**(3): p. 187-192.
100. Nalls, M.A., et al., *Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease.* Nature Genetics, 2014. **46**(9): p. 989-+.
101. Nuytemans, K., et al., *Genetic Etiology of Parkinson Disease Associated with Mutations in the SNCA, PARK2, PINK1, PARK7, and LRRK2 Genes: A Mutation Update.* Human Mutation, 2010. **31**(7): p. 763-780.
102. Trinh, J., I. Guella, and M.J. Farrer, *Disease Penetrance of Late-Onset Parkinsonism A Meta-analysis.* Jama Neurology, 2014. **71**(12): p. 1535-1539.
103. Polymeropoulos, M.H., et al., *Mapping of a gene for Parkinson's disease to chromosome 4q21-q23.* Science, 1996. **274**(5290): p. 1197-1199.
104. Ritz, B., et al., *alpha-Synuclein Genetic Variants Predict Faster Motor Symptom Progression in Idiopathic Parkinson Disease.* Plos One, 2012. **7**(5).
105. Klein, C. and A. Westenberger, *Genetics of Parkinson's Disease.* Cold Spring Harbor Perspectives in Medicine, 2012. **2**(1).
106. Nishioka, K., et al., *Clinical heterogeneity of alpha-synuclein gene duplication in Parkinson's disease.* Annals of Neurology, 2006. **59**(2): p. 298-309.
107. Ross, O.A., et al., *Genomic investigation of alpha-synuclein multiplication and parkinsonism.* Annals of Neurology, 2008. **63**(6): p. 743-750.
108. Fuchs, J., et al., *Phenotypic variation in a large Swedish pedigree due to SNCA duplication and triplication.* Neurology, 2007. **68**(12): p. 916-922.
109. Miller, D.W., et al., *alpha-Synuclein in blood and brain from familial Parkinson disease with SNCA locus triplication.* Neurology, 2004. **62**(10): p. 1835-1838.
110. Singleton, A.B., et al., *alpha-synuclein locus triplication causes Parkinson's disease.* Science, 2003. **302**(5646): p. 841-841.
111. Wang, X., et al., *LRRK2 regulates mitochondrial dynamics and function through direct interaction with DLP1.* Human Molecular Genetics, 2012. **21**(9): p. 1931-1944.
112. Niu, J., et al., *Leucine-rich repeat kinase 2 disturbs mitochondrial dynamics via Dynamin-like protein.* Journal of Neurochemistry, 2012. **122**(3): p. 650-658.
113. Stafa, K., et al., *Functional interaction of Parkinsons disease-associated LRRK2 with members of the dynamin GTPase superfamily.* Human Molecular Genetics, 2014. **23**(8): p. 2055-2077.

114. Ryan, B.J., et al., *Mitochondrial dysfunction and mitophagy in Parkinson's: from familial to sporadic disease.* Trends in Biochemical Sciences, 2015. **40**(4): p. 200-210.
115. Guerreiro, P.S., Huang, Y., Gysbers, A., Cheng, D., Gai, W. P., Outeiro, T. F., & Halliday, G. M., *LRRK2 interactions with alpha-synuclein in Parkinson's disease brains and in cell models.* Journal of Molecular Medicine (Berlin, Germany), 2013. **91**(4): p. 513-522.
116. Qing, H., et al., *Lrrk2 phosphorylates alpha synuclein at serine 129: Parkinson disease implications.* Biochemical and Biophysical Research Communications, 2009. **387**(1): p. 149-152.
117. Smith, W.W., et al., *Leucine-rich repeat kinase 2 (LRRK2) interacts with parkin, and mutant LRRK2 induces neuronal degeneration.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(51): p. 18676-18681.
118. Venderova, K., et al., *Leucine-rich repeat kinase 2 interacts with Parkin, DJ-1 and PINK-1 in a Drosophila melanogaster model of Parkinson's disease.* Human Molecular Genetics, 2009. **18**(22): p. 4390-4404.
119. Guedes, L.C., et al., *Worldwide frequency of G2019S LRRK2 mutation in Parkinson's disease: A systematic review.* Parkinsonism & Related Disorders, 2010. **16**(4): p. 237-242.
120. Lesage, S., et al., *LRRK2 G2019S as a cause of Parkinson's disease in North African Arabs.* New England Journal of Medicine, 2006. **354**(4): p. 422-423.
121. Ozelius, L.J., et al., *LRRK2 G2019S as a cause of Parkinson's disease in Ashkenazi Jews.* New England Journal of Medicine, 2006. **354**(4): p. 424-425.
122. Gloeckner, C.J., et al., *The Parkinson disease causing LRRK2 mutation I2020T is associated with increased kinase activity.* Human Molecular Genetics, 2006. **15**(2): p. 223-232.
123. West, A.B., et al., *Parkinson's disease-associated mutations in leucine-rich repeat kinase 2 augment kinase activity.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(46): p. 16842-16847.
124. Lin, M.T. and M.F. Beal, *Mitochondrial dysfunction and oxidative stress in neurodegenerative diseases.* Nature, 2006. **443**(7113): p. 787-795.
125. Heo, H.Y., et al., *LRRK2 enhances oxidative stress-induced neurotoxicity via its kinase activity.* Experimental Cell Research, 2010. **316**(4): p. 649-656.
126. Nguyen, H.N., et al., *LRRK2 mutant iPSC-derived DA neurons demonstrate increased susceptibility to oxidative stress.* Cell stem cell, 2011. **8**(3): p. 267-80.
127. Deng, J., et al., *Structure of the ROC domain from the Parkinson's disease-associated leucine-rich repeat kinase 2 reveals a dimeric GTPase.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(5): p. 1499-1504.
128. Biosa, A., et al., *GTPase activity regulates kinase activity and cellular phenotypes of Parkinson's disease-associated LRRK2.* Human Molecular Genetics, 2013. **22**(6): p. 1140-1156.
129. Ito, G., et al., *GTP binding is essential to the protein kinase activity of LRRK2, a causative gene product for familial Parkinson's disease.* Biochemistry, 2007. **46**(5): p. 1380-1388.
130. Jaleel, M., et al., *LRRK2 phosphorylates moesin at threonine-558: characterization of how Parkinson's disease mutants affect kinase activity.* Biochemical Journal, 2007. **405**: p. 307-317.
131. Nichols, R.J., et al., *14-3-3 binding to LRRK2 is disrupted by multiple Parkinson's disease-associated mutations and regulates cytoplasmic localization.* Biochemical Journal, 2010. **430**: p. 393-404.
132. Xiong, Y.L., et al., *GTPase Activity Plays a Key Role in the Pathobiology of LRRK2.* Plos Genetics, 2010. **6**(4).

133. Haugarvoll, K., et al., *Lrrk2 R1441C parkinsonism is clinically similar to sporadic Parkinson disease.* Neurology, 2008. **70**(16): p. 1456-1460.
134. Mabel Gatto, E., et al., *The LRRK2 G2019S mutation in a series of Argentinean patients with Parkinson's disease: Clinical and demographic characteristics.* Neuroscience Letters, 2013. **537**: p. 1-5.
135. Zimprich, A., et al., *Mutations in LRRK2 cause autosomal-dominant Parkinsonism with pleomorphic pathology.* Neuron, 2004. **44**(4): p. 601-607.
136. Khan, N.L., et al., *Mutations in the gene LRRK2 encoding dardarin (PARK8) cause familial Parkinson's disease: clinical, pathological, olfactory and functional imaging and genetic data.* Brain, 2005. **128**: p. 2786-2796.
137. Ross, O.A., et al., *Lrrk2 and Lewy body disease.* Annals of Neurology, 2006. **59**(2): p. 388-393.
138. Kalia, L.V., et al., *Clinical Correlations With Lewy Body Pathology in LRRK2-Related Parkinson Disease.* Jama Neurology, 2015. **72**(1): p. 100-105.
139. Poulopoulos, M., O.A. Levy, and R.N. Alcalay, *The neuropathology of genetic Parkinson's disease.* Movement Disorders, 2012. **27**(7): p. 831-842.
140. Hasegawa, K., et al., *Familial parkinsonism: Study of original Sagamihara PARK8 (I2020T) kindred with variable clinicopathologic outcomes.* Parkinsonism & Related Disorders, 2009. **15**(4): p. 300-306.
141. Marti-Masso, J.-F., et al., *Neuropathology of Parkinson's Disease with the R1441G Mutation in LRRK2.* Movement Disorders, 2009. **24**(13): p. 1998-2001.
142. Buettner, S., et al., *Synphilin-1 Enhances alpha-Synuclein Aggregation in Yeast and Contributes to Cellular Stress and Cell Death in a Sir2-Dependent Manner.* Plos One, 2010. **5**(10).
143. Nuber, S., et al., *Transgenic overexpression of the alpha-synuclein interacting protein synphilin-1 leads to behavioral and neuropathological alterations in mice.* Neurogenetics, 2010. **11**(1): p. 107-120.
144. Ko, H.S., et al., *Accumulation of the authentic parkin substrate aminoacyl-tRNA synthetase cofactor, p38/JTV-1, leads to catecholaminergic cell death.* Journal of Neuroscience, 2005. **25**(35): p. 7968-7978.
145. Corti, O., et al., *The p38 subunit of the aminoacyl-tRNA synthetase complex is a Parkin substrate: linking protein biosynthesis and neurodegeneration.* Human Molecular Genetics, 2003. **12**(12): p. 1427-1437.
146. Wakabayashi, K., et al., *Synphilin-1 is present in Lewy bodies in Parkinson's disease.* Annals of Neurology, 2000. **47**(4): p. 521-523.
147. Chung, K.K.K., et al., *Parkin ubiquitinates the alpha-synuclein-interacting protein, synphilin-1: implications for Lewy-body formation in Parkinson disease.* Nature Medicine, 2001. **7**(10): p. 1144-1150.
148. Huynh, D.P., et al., *Parkin is an E3 ubiquitin-ligase for normal and mutant ataxin-2 and prevents ataxin-2-induced cell death.* Experimental Neurology, 2007. **203**(2): p. 531-541.
149. Imai, Y., et al., *An unfolded putative transmembrane polypeptide, which can lead to endoplasmic reticulum stress, is a substrate of parkin.* Cell, 2001. **105**(7): p. 891-902.
150. Ng, C.-H., et al., *Parkin Protects against LRRK2 G2019S Mutant-Induced Dopaminergic Neurodegeneration in Drosophila.* Journal of Neuroscience, 2009. **29**(36): p. 11257-11262.
151. Khandelwal, P.J., et al., *Parkinson-related parkin reduces alpha-Synuclein phosphorylation in a gene transfer model.* Molecular Neurodegeneration, 2010. **5**.
152. Petrucelli, L., et al., *Parkin protects against the toxicity associated with mutant alpha-synuclein: Proteasome dysfunction selectively affects catecholaminergic neurons.* Neuron, 2002. **36**(6): p. 1007-1019.

153. Yasuda, T., et al., *Neuronal specificity of alpha-synuclein toxicity and effect of parkin co-expression in primates.* Neuroscience, 2007. **144**(2): p. 743-753.
154. Yang, Y.F., et al., *Parkin suppresses dopaminergic neuron-selective neurotoxicity induced by Pael-R in Drosophila.* Neuron, 2003. **37**(6): p. 911-924.
155. Song, P., et al., *Parkin promotes proteasomal degradation of p62: implication of selective vulnerability of neuronal cells in the pathogenesis of Parkinson's disease.* Protein & Cell, 2016. **7**(2): p. 114-129.
156. Shimura, H., et al., *Familial Parkinson disease gene product, parkin, is a ubiquitin-protein ligase.* Nature Genetics, 2000. **25**(3): p. 302-305.
157. Hampe, C., et al., *Biochemical analysis of Parkinson's disease-causing variants of Parkin, an E3 ubiquitin-protein ligase with monoubiquitylation capacity.* Human Molecular Genetics, 2006. **15**(13): p. 2059-2075.
158. Wang, C., et al., *Alterations in the solubility and intracellular localization of parkin by several familial Parkinson's disease-linked point mutations.* Journal of Neurochemistry, 2005. **93**(2): p. 422-431.
159. Dawson, T.M. and V.L. Dawson, *The Role of Parkin in Familial and Sporadic Parkinson's Disease.* Movement Disorders, 2010. **25**(3): p. S32-S39.
160. Sun, M., et al., *Influence of heterozygosity for Parkin mutation on onset age in familial Parkinson disease The GenePD study.* Archives of Neurology, 2006. **63**(6): p. 826-832.
161. Doherty, K.M., et al., *Parkin Disease A Clinicopathologic Entity?* Jama Neurology, 2013. **70**(5): p. 571-579.
162. Chang, F.C.F., et al., *"Dancing feet dyskinesias": A clue to parkin gene mutations.* Movement Disorders, 2012. **27**(4): p. 587-588.
163. Khan, N.L., et al., *Parkin disease: a phenotypic study of a large case series.* Brain, 2003. **126**: p. 1279-1292.
164. Lohmann, E., et al., *How much phenotypic variation can be attributed to parkin genotype?* Annals of Neurology, 2003. **54**(2): p. 176-185.
165. Alcalay, R.N., et al., *Cognitive and Motor Function in Long-Duration PARKIN-Associated Parkinson Disease.* Jama Neurology, 2014. **71**(1): p. 62-67.
166. Kitada, T., et al., *Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism.* Nature, 1998. **392**(6676): p. 605-608.
167. Sharp, M.E., et al., *Parkinson's disease with Lewy bodies associated with a heterozygous PARKIN dosage mutation.* Movement Disorders, 2014. **29**(4): p. 566-568.
168. Miyakawa, S., et al., *Lewy body pathology in a patient with a homozygous Parkin deletion.* Movement Disorders, 2013. **28**(3): p. 388-391.
169. Farrer, M., et al., *Lewy bodies and parkinsonism in families with parkin mutations.* Annals of Neurology, 2001. **50**(3): p. 293-300.
170. Sasaki, S., et al., *Parkin-positive autosomal recessive juvenile parkinsonism with alpha-synuclein-positive inclusions.* Neurology, 2004. **63**(4): p. 678-682.
171. McLelland, G.-L., et al., *Parkin and PINK1 function in a vesicular trafficking pathway regulating mitochondrial quality control.* Embo Journal, 2014. **33**(4): p. 282-295.
172. Petit, A., et al., *Wild-type PINK1 prevents basal and induced neuronal apoptosis, a protective effect abrogated by Parkinson disease-related mutations.* Journal of Biological Chemistry, 2005. **280**(40): p. 34025-34032.
173. Wood-Kaczmar, A., et al., *PINK1 Is Necessary for Long Term Survival and Mitochondrial Function in Human Dopaminergic Neurons.* Plos One, 2008. **3**(6).
174. Haque, M.E., et al., *Cytoplasmic Pink1 activity protects neurons from dopaminergic neurotoxin MPTP.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(5): p. 1716-1721.

175. Song, S., et al., *Characterization of PINK1 (PTEN-induced Putative Kinase 1) Mutations Associated with Parkinson Disease in Mammalian Cells and Drosophila.* Journal of Biological Chemistry, 2013. **288**(8): p. 5660-5672.
176. Samaranch, L., et al., *PINK1-linked parkinsonism is associated with Lewy body pathology.* Brain, 2010. **133**: p. 1128-1142.
177. Ephraty, L., et al., *Neuropsychiatric and cognitive features in autosomal-recessive early Parkinsonism due to PINK1 mutations.* Movement Disorders, 2007. **22**(4): p. 566-569.
178. Takanashi, M., Y. Li, and N. Hattori, *Absence of Lewy pathology associated with PINK1 homozygous mutation.* Neurology, 2016. **86**(23): p. 2212-3.
179. Sidransky, E. and G. Lopez, *The link between the GBA gene and parkinsonism.* Lancet Neurology, 2012. **11**(11): p. 986-998.
180. Alcalay, R.N., et al., *Glucocerebrosidase activity in Parkinson's disease with and without GBA mutations.* Brain, 2015. **138**: p. 2648-2658.
181. Schapira, A.H.V., *Glucocerebrosidase and Parkinson disease: Recent advances.* Molecular and Cellular Neuroscience, 2015. **66**: p. 37-42.
182. Alvarez-Erviti, L., et al., *Lysosomal dysfunction increases exosome-mediated alpha-synuclein release and transmission.* Neurobiology of Disease, 2011. **42**(3): p. 360-367.
183. Bae, E.-J., et al., *Glucocerebrosidase depletion enhances cell-to-cell transmission of alpha-synuclein.* Nature Communications, 2014. **5**.
184. Goker-Alpan, O., et al., *Glucocerebrosidase is present in alpha-synuclein inclusions in Lewy body disorders.* Acta Neuropathologica, 2010. **120**(5): p. 641-649.
185. Velayati, A., W.H. Yu, and E. Sidransky, *The Role of Glucocerebrosidase Mutations in Parkinson Disease and Lewy Body Disorders.* Current Neurology and Neuroscience Reports, 2010. **10**(3): p. 190-198.
186. Davis, A.A., et al., *Variants in GBA, SNCA, and MAPT influence Parkinson disease risk, age at onset, and progression.* Neurobiology of Aging, 2016. **37**.
187. Shendelman, S., et al., *DJ-1 is a redox-dependent molecular chaperone that inhibits alpha-synuclein aggregate formation.* Plos Biology, 2004. **2**(11): p. 1764-1773.
188. Zhou, W.B., et al., *The oxidation state of DJ-1 regulates its chaperone activity toward alpha-synuclein.* Journal of Molecular Biology, 2006. **356**(4): p. 1036-1048.
189. Lev, N., et al., *DJ-1 protects against dopamine toxicity.* Journal of Neural Transmission, 2009. **116**(2): p. 151-160.
190. Baulac, S., et al., *Increased DJ-1 expression under oxidative stress and in Alzheimer's disease brains.* Molecular Neurodegeneration, 2009. **4**.
191. Joselin, A.P., et al., *ROS-dependent regulation of Parkin and DJ-1 localization during oxidative stress in neurons.* Human Molecular Genetics, 2012. **21**(22): p. 4888-4903.
192. Ariga, H., et al., *Neuroprotective function of DJ-1 in Parkinson's disease.* Oxidative medicine and cellular longevity, 2013. **2013**: p. 683920-683920.
193. Junn, E., et al., *Interaction of DJ-1 with Daxx inhibits apoptosis signal-regulating kinase 1 activity and cell death.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(27): p. 9691-9696.
194. Pankratz, N., et al., *Mutations in DJ-1 are rare in familial Parkinson disease.* Neuroscience Letters, 2006. **408**(3): p. 209-213.
195. Bonifati, V., et al., *Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism.* Science, 2003. **299**(5604): p. 256-259.
196. Clark, L.N., et al., *Analysis of an early-onset Parkinson's disease cohort for DJ-1 mutations.* Movement Disorders, 2004. **19**(7): p. 796-800.
197. Hague, S., et al., *Early-onset Parkinson's disease caused by a compound heterozygous DJ-1 mutation.* Annals of Neurology, 2003. **54**(2): p. 271-274.

198. Mullett, S.J. and D.A. Hinkle, *DJ-1 deficiency in astrocytes selectively enhances mitochondrial Complex I inhibitor-induced neurotoxicity.* Journal of Neurochemistry, 2011. **117**(3): p. 375-387.
199. Kahle, P.J., J. Waak, and T. Gasser, *DJ-1 and prevention of oxidative stress in Parkinson's disease and other age-related disorders.* Free Radical Biology and Medicine, 2009. **47**(10): p. 1354-1361.
200. Schulte, C. and T. Gasser, *Genetic basis of Parkinson's disease: inheritance, penetrance, and expression.* The application of clinical genetics, 2011. **4**: p. 67-80.
201. Taipa, R., et al., *DJ-1 linked parkinsonism (PARK7) is associated with Lewy body pathology.* Brain, 2016. **139**: p. 1680-1687.
202. Abraham, G. and M. Inouye, *Genomic risk prediction of complex human disease and its clinical application.* Current Opinion in Genetics & Development, 2015. **33**: p. 10-16.
203. Manolio, T.A., *Bringing genome-wide association findings into clinical use.* Nature Reviews Genetics, 2013. **14**(8): p. 549-558.
204. Nalls, M.A., et al., *Diagnosis of Parkinson's disease on the basis of clinical and genetic classification: a population-based modelling study.* Lancet Neurology, 2015. **14**(10): p. 1002-1009.
205. Hall, T.O., et al., *Risk prediction for complex diseases: application to Parkinson disease.* Genetics in Medicine, 2013. **15**(5): p. 361-367.
206. Escott-Price, V., et al., *Polygenic Risk of Parkinson Disease Is Correlated with Disease Age at Onset.* Annals of Neurology, 2015. **77**(4): p. 582-591.
207. Zang, L.Y. and H.P. Misra, *GENERATION OF REACTIVE OXYGEN SPECIES DURING THE MONOAMINE OXIDASE-CATALYZED OXIDATION OF THE NEUROTOXICANT, 1-METHYL-4-PHENYL-1,2,3,6-TETRAHYDROPYRIDINE.* Journal of Biological Chemistry, 1993. **268**(22): p. 16504-16512.
208. Nicklas, W.J., I. Vyas, and R.E. Heikkila, *INHIBITION OF NADH-LINKED OXIDATION IN BRAIN MITOCHONDRIA BY 1-METHYL-4-PHENYL-PYRIDINE, A METABOLITE OF THE NEUROTOXIN, 1-METHYL-4-PHENYL-1,2,5,6-TETRAHYDROPYRIDINE.* Life Sciences, 1985. **36**(26): p. 2503-2508.
209. Ramsay, R.R., J.I. Salach, and T.P. Singer, *UPTAKE OF THE NEUROTOXIN 1-METHYL-4-PHENYLPYRIDINE (MPP+) BY MITOCHONDRIA AND ITS RELATION TO THE INHIBITION OF THE MITOCHONDRIAL OXIDATION OF NAD+-LINKED SUBSTRATES BY MPP+.* Biochemical and Biophysical Research Communications, 1986. **134**(2): p. 743-748.
210. Langston, J.W., et al., *Evidence of active nerve cell degeneration in the substantia nigra of humans years after 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine exposure.* Annals of Neurology, 1999. **46**(4): p. 598-605.
211. Tanner, C.M., et al., *Rotenone, Paraquat, and Parkinson's Disease.* Environmental Health Perspectives, 2011. **119**(6): p. 866-872.
212. Palmer, G.H., D. J.; Tisdale, H.; Singer, T. P.; Beinert, H., *Studies on the respiratory chain-linked reduced nicotinamide adenine dinucleotide dehydrogenase. XIV. Location of the sites of inhibition of rotenone, barbiturates, and piericidin by means of electron paramagnetic resonance spectroscopy.* Journal of Biological Chemistry, 1968. **243**(4): p. 844-7.
213. Purisai, M.G., et al., *Microglial activation as a priming event leading to paraquat-induced dopaminergic cell degeneration.* Neurobiology of Disease, 2007. **25**(2): p. 392-400.
214. Cannon, J.R., et al., *A highly reproducible rotenone model of Parkinson's disease.* Neurobiology of Disease, 2009. **34**(2): p. 279-290.
215. Marella, M., et al., *Protection by the NDI1 Gene against Neurodegeneration in a Rotenone Rat Model of Parkinson's Disease.* Plos One, 2008. **3**(1).

216. Manning-Bog, A.B., et al., *The herbicide paraquat causes up-regulation and aggregation of alpha-synuclein in mice - Paraquat and alpha-synuclein.* Journal of Biological Chemistry, 2002. **277**(3): p. 1641-1644.
217. Checkoway, H., et al., *Parkinson's disease risks associated with cigarette smoking, alcohol consumption, and caffeine intake.* American Journal of Epidemiology, 2002. **155**(8): p. 732-738.
218. Allam, M.F., et al., *Smoking and Parkinson's disease: Systematic review of prospective studies.* Movement Disorders, 2004. **19**(6): p. 614-621.
219. Fowler, J.S., et al., *Inhibition of monoamine oxidase B in the brains of smokers.* Nature, 1996. **379**(6567): p. 733-736.
220. Alves, G., et al., *Cigarette smoking in Parkinson's disease: Influence on disease progression.* Movement Disorders, 2004. **19**(9): p. 1087-1092.
221. Ross, G.W., et al., *Association of coffee and caffeine intake with the risk of Parkinson disease.* Jama-Journal of the American Medical Association, 2000. **283**(20): p. 2674-2679.
222. Ascherio, A., et al., *Prospective study of caffeine consumption and risk of Parkinson's disease in men and women.* Annals of Neurology, 2001. **50**(1): p. 56-63.
223. Saaksjarvi, K., et al., *Prospective study of coffee consumption and risk of Parkinson's disease.* European Journal of Clinical Nutrition, 2008. **62**(7): p. 908-915.
224. Chen, J.F., et al., *Neuroprotection by caffeine and A(2A) adenosine receptor inactivation in a model of Parkinson's disease.* Journal of Neuroscience, 2001. **21**(10): p. art. no.-RC143.
225. Simoes, A.P., et al., *Blockade of adenosine A(2A) receptors prevents interleukin-1 beta-induced exacerbation of neuronal toxicity through a p38 mitogen-activated protein kinase pathway.* Journal of Neuroinflammation, 2012. **9**.
226. Postuma, R.B., et al., *Caffeine for treatment of Parkinson disease A randomized controlled trial.* Neurology, 2012. **79**(7): p. 651-658.
227. Bara-Jimenez, W., et al., *Adenosine A(2A) receptor antagonist treatment of Parkinson's disease.* Neurology, 2003. **61**(3): p. 293-296.
228. Hauser, R.A., et al., *Randomized trial of the adenosine A(2A) receptor antagonist istradefylline in advanced PD.* Neurology, 2003. **61**(3): p. 297-303.
229. Stacy, M., et al., *A 12-week, placebo-controlled study (6002-US-006) of istradefylline in Parkinson disease.* Neurology, 2008. **70**(23): p. 2233-2240.
230. Fernandez, H.H., et al., *Istradefylline as monotherapy for Parkinson disease: Results of the 6002-US-051 trial.* Parkinsonism & Related Disorders, 2010. **16**(1): p. 16-20.
231. Ton, T.G., et al., *Nonsteroidal anti-inflammatory drugs and risk of Parkinson's disease.* Movement Disorders, 2006. **21**(7): p. 964-969.
232. Maharaj, H., D.S. Maharaj, and S. Daya, *Acetylsalicylic acid and acetaminophen protect against MPP+-induced mitochondrial damage and superoxide anion generation.* Life Sciences, 2006. **78**(21): p. 2438-2443.
233. Hirohata, M., et al., *Non-steroidal anti-inflammatory drugs have potent anti-fibrillogenic and fibril-destabilizing effects for alpha-synuclein fibrils in vitro.* Neuropharmacology, 2008. **54**(3): p. 620-627.
234. Di Matteo, V., et al., *Aspirin protects striatal dopaminergic neurons from neurotoxin-induced degeneration: An in vivo microdialysis study.* Brain Research, 2006. **1095**: p. 167-177.
235. Driver, J.A., et al., *Use of non-steroidal anti-inflammatory drugs and risk of Parkinson's disease: nested case-control study.* British Medical Journal, 2011. **342**.
236. Kageyama, T., et al., *The 4F2hc/LAT1 complex transports L-DOPA across the blood-brain barrier.* Brain Research, 2000. **879**(1-2): p. 115-121.

237. Ahlskog, J.E. and M.D. Muenter, *Frequency of levodopa-related dyskinesias and motor fluctuations as estimated from the cumulative literature.* Movement Disorders, 2001. **16**(3): p. 448-458.
238. Hauser, R.A., et al., *Long-term evaluation of bilateral fetal nigral transplantation in Parkinson disease.* Archives of Neurology, 1999. **56**(2): p. 179-187.
239. Freed, C.R., et al., *Transplantation of embryonic dopamine neurons for severe Parkinson's disease.* New England Journal of Medicine, 2001. **344**(10): p. 710-719.
240. Hagell, P., et al., *Sequential bilateral transplantation in Parkinson's disease - Effects of the second graft.* Brain, 1999. **122**: p. 1121-1132.
241. Brundin, P., et al., *Bilateral caudate and putamen grafts of embryonic mesencephalic tissue treated with lazaroids in Parkinson's disease.* Brain, 2000. **123**: p. 1380-1390.
242. Kefalopoulou, Z., et al., *Long-term Clinical Outcome of Fetal Cell Transplantation for Parkinson Disease Two Case Reports.* Jama Neurology, 2014. **71**(1): p. 83-87.
243. Hagell, P., et al., *Dyskinesias following neural transplantation in Parkinson's disease.* Nature Neuroscience, 2002. **5**(7): p. 627-628.
244. Chu, Y.P. and J.H. Kordower, *Lewy body pathology in fetal grafts.* Year in Neurology 2, 2010. **1184**: p. 55-67.
245. Kordower, J.H., et al., *Lewy body-like pathology in long-term embryonic nigral transplants in Parkinson's disease.* Nature Medicine, 2008. **14**(5): p. 504-506.
246. Li, J.Y., et al., *Lewy bodies in grafted neurons in subjects with Parkinson's disease suggest host-to-graft disease propagation.* Nature Medicine, 2008. **14**(5): p. 501-503.
247. Jankovic, J.M., M.; Carter, J.; Gauthier, S.; Goetz, C.; Golbe, L.; Huber, S.; Koller, W.; Olanow, C.; Shoulson, I.; Stern, M.; Tanner, C.; Weiner, W.; Parkinson Study Group, *Variable expression of Parkinson's disease: a base-line analysis of the DATATOP cohort.* Neurology, 1990. **40**(10): p. 1529-34.
248. Michael Lawton, F.B., Michal Rolinski, Claudio Ruffman, Kannan Nithi, and Y.B.-S.a.M.T.M.H. Margaret T May, *Parkinson's disease subtypes in the Oxford Parkinson Disease Centre*

(OPDC) Discovery cohort. Journal of Parkinson's Disease, 2015.
249. Szeto, J.Y.Y., et al., *The relationships between mild cognitive impairment and phenotype in Parkinson's disease.* Movement Disorders, 2015. **30**: p. S350-S350.
250. Ma, L.Y., et al., *Heterogeneity among patients with Parkinson's disease: Cluster analysis and genetic association.* Journal of the Neurological Sciences, 2015. **351**(1-2): p. 41-45.
251. Lewis, S.J.G., et al., *Heterogeneity of Parkinson's disease in the early clinical stages using a data driven approach.* Journal of Neurology Neurosurgery and Psychiatry, 2005. **76**(3): p. 343-348.
252. van Balkom, T.D., et al., *Profiling cognitive and neuropsychiatric heterogeneity in Parkinson's disease.* Parkinsonism & Related Disorders, 2016. **28**: p. 130-136.
253. Honti, F., S. Meader, and C. Webber, *Unbiased Functional Clustering of Gene Variants with a Phenotypic-Linkage Network.* Plos Computational Biology, 2014. **10**(8): p. 7.
254. Robinson, P.N., et al., *The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease.* American Journal of Human Genetics, 2008. **83**(5): p. 610-615.
255. Eppig, J.T., et al., *The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse.* Nucleic Acids Research, 2012. **40**(D1): p. D881-D886.
256. Lee, I., et al., *Prioritizing candidate disease genes by network-based boosting of genome-wide association data.* Genome Research, 2011. **21**(7): p. 1109-1121.

257. Gillis, J. and P. Pavlidis, *The Impact of Multifunctional Genes on "Guilt by Association" Analysis.* Plos One, 2011. **6**(2).
258. Opsahl, T., *Structure and Evolution of Weighted Networks.* University of London (Queen Mary College), London, UK, 2009: p. 104-122.
259. Purcell, S., et al., *PLINK: A tool set for whole-genome association and population-based linkage analyses.* American Journal of Human Genetics, 2007. **81**(3): p. 559-575.
260. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nature Genetics, 2006. **38**(8): p. 904-909.
261. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-2158.
262. Das, S., et al., *Next-generation genotype imputation service and methods.* Nature Genetics, 2016. **48**(10): p. 1284-1287.
263. Altshuler, D.M., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-+.
264. Loh, P.R., et al., *Reference-based phasing using the Haplotype Reference Consortium panel.* Nature Genetics, 2016. **48**(11): p. 1443-1448.
265. Ashburner, M., et al., *Gene Ontology: tool for the unification of biology.* Nature Genetics, 2000. **25**(1): p. 25-29.
266. Su, A.I., et al., *A gene atlas of the mouse and human protein-encoding transcriptomes.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(16): p. 6062-6067.
267. Hunter, S., et al., *InterPro in 2011: new developments in the family and domain prediction database.* Nucleic Acids Research, 2012. **40**(D1): p. D306-D312.
268. Szklarczyk, D., et al., *STRING v10: protein-protein interaction networks, integrated over the tree of life.* Nucleic Acids Research, 2015. **43**(D1): p. D447-D452.
269. Hunt, S.C., et al., *Polymorphisms in the NPY2R Gene Show Significant Associations With BMI That Are Additive to FTO, MC4R, and NPFFR2 Gene Effects.* Obesity, 2011. **19**(11): p. 2241-2247.
270. Coskran, T.M., et al., *Immunohistochemical localization of phosphodiesterase 10A in multiple mammalian species.* Journal of Histochemistry & Cytochemistry, 2006. **54**(11): p. 1205-1213.
271. Nishi, A., et al., *Distinct Roles of PDE4 and PDE10A in the Regulation of cAMP/PKA Signaling in the Striatum.* Journal of Neuroscience, 2008. **28**(42): p. 10460-10471.
272. Niccolini, F., et al., *Loss of phosphodiesterase 10A expression is associated with progression and severity in Parkinson's disease.* Brain, 2015. **138**: p. 3003-3015.
273. Wills, A.M.A., et al., *Association Between Change in Body Mass Index, Unified Parkinson's Disease Rating Scale Scores, and Survival Among Persons With Parkinson Disease Secondary Analysis of Longitudinal Data From NINDS Exploratory Trials in Parkinson Disease Long-term Study 1.* Jama Neurology, 2016. **73**(3): p. 321-328.
274. Bloom, J.M., S; Martinez, M; von Weymarn, L; Bierut, L; Goate, A, *Effects upon in vivo nicotine metabolism reveal functional variation in FMO3 associated with cigarette consumption.* Pharmacogenet Genomics, 2015. **23**(2): p. 62-68.
275. Uehara, S., et al., *Activation and Deactivation of 1-Methyl-4-Phenyl-1,2,3,6-Tetrahydropyridine by Cytochrome P450 Enzymes and Flavin-Containing Monooxygenases in Common Marmosets (Callithrix jacchus).* Drug Metabolism and Disposition, 2015. **43**(5): p. 735-742.
276. Shimizu, M., et al., *Potential for drug interactions mediated by polymorphic flavin-containing monooxygenase 3 in human livers.* Drug Metabolism and Pharmacokinetics, 2015. **30**(1): p. 70-74.

277. Koukouritaki, S.B., et al., *Human hepatic flavin-containing monooxygenases 1 (FMO1) and 3 (FMO3) developmental expression.* Pediatric Research, 2002. **51**(2): p. 236-243.
278. Kusminski, C.M., et al., *MitoNEET-driven alterations in adipocyte mitochondrial activity reveal a crucial adaptive process that preserves insulin sensitivity in obesity.* Nature Medicine, 2012. **18**(10): p. 1539-U144.
279. Zheng, B., et al., *PGC-1 alpha, A Potential Therapeutic Target for Early Intervention in Parkinson's Disease.* Science Translational Medicine, 2010. **2**(52).
280. St-Pierre, J., et al., *Suppression of reactive oxygen species and neurodegeneration by the PGC-1 transcriptional coactivators.* Cell, 2006. **127**(2): p. 397-408.
281. Dabrowska, A., et al., *PGC-1 alpha controls mitochondrial biogenesis and dynamics in lead-induced neurotoxicity.* Aging-Us, 2015. **7**(9): p. 629-647.
282. Beccano-Kelly, D.A., et al., *LRRK2 overexpression alters glutamatergic presynaptic plasticity, striatal dopamine tone, postsynaptic signal transduction, motor activity and memory.* Human Molecular Genetics, 2015. **24**(5): p. 1336-1349.
283. Shin, N., et al., *LRRK2 regulates synaptic vesicle endocytosis.* Experimental Cell Research, 2008. **314**(10): p. 2055-2065.
284. Rudenko, I.N. and M.R. Cookson, *Heterogeneity of Leucine-Rich Repeat Kinase 2 Mutations: Genetics, Mechanisms and Therapeutic Implications.* Neurotherapeutics, 2014. **11**(4): p. 738-750.
285. Guffanti, G., et al., *Increased CNV-Region deletions in mild cognitive impairment (MCI) and Alzheimer's disease (AD) subjects in the ADNI sample.* Genomics, 2013. **102**(2): p. 112-122.
286. Heinzen, E.L., et al., *Genome-Wide Scan of Copy Number Variation in Late-Onset Alzheimer's Disease.* Journal of Alzheimers Disease, 2010. **19**(1): p. 69-77.
287. Swaminathan, S., et al., *Analysis of Copy Number Variation in Alzheimer's Disease: The NIA-LOAD/NCRAD Family Study.* Current Alzheimer Research, 2012. **9**(7): p. 801-814.
288. Chartier-Harlin, M.C., et al., *alpha-synuclein locus duplication as a cause of familial Parkinson's disease.* Lancet, 2004. **364**(9440): p. 1167-1169.
289. Wang, L., et al., *High-Resolution Survey in Familial Parkinson Disease Genes Reveals Multiple Independent Copy Number Variation Events in PARK2.* Human Mutation, 2013. **34**(8): p. 1071-1074.
290. Szewczyk-Krolikowski, K., et al., *The influence of age and gender on motor and non-motor features of early Parkinson's disease: Initial findings from the Oxford Parkinson Disease Center (OPDC) discovery cohort.* Parkinsonism & Related Disorders, 2014. **20**(1): p. 99-105.
291. Valsesia, A., et al., *Identification and validation of copy number variants using SNP genotyping arrays from a large clinical cohort.* Bmc Genomics, 2012. **13**.
292. Leek, J.T., et al., *Tackling the widespread and critical impact of batch effects in high-throughput data.* Nature Reviews Genetics, 2010. **11**(10): p. 733-739.
293. Hong, H., et al., *Assessing batch effects of genotype calling algorithm BRLMM for the Affymetrix GeneChip Human Mapping 500 K array set using 270 HapMap samples.* Bmc Bioinformatics, 2008. **9**.
294. Luo, J., et al., *A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data.* Pharmacogenomics Journal, 2010. **10**(4): p. 278-291.
295. Valsesia, A., et al., *The growing importance of CNVs: new insights for detection and clinical interpretation.* Frontiers in Genetics, 2013. **4**: p. 92-Article No.: 92.

296. Reese, S.E., et al., *A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis.* Bioinformatics, 2013. **29**(22): p. 2877-2883.
297. Wang, K., et al., *PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data.* Genome Research, 2007. **17**(11): p. 1665-1674.
298. Need, A.C., et al., *A Genome-Wide Investigation of SNPs and CNVs in Schizophrenia.* Plos Genetics, 2009. **5**(2).
299. Liu X, C.R., Ye X, et al., *Increased rate of sporadic and recurrent rare genic copy number variants in Parkinson's disease among Ashkenazi Jews.* Molecular Genetics & Genomic Medicine, 2013. **1**(3): p. 142-154.
300. Schenck, C.H., S.R. Bundlie, and M.W. Mahowald, *Delayed emergence of a parkinsonian disorder in 38% of 29 older men initially diagnosed with idiopathic rapid eye movement sleep behavior disorder.* Neurology, 1996. **46**(2): p. 388-393.
301. Schenck, C.H., B.F. Boeve, and M.W. Mahowald, *Delayed emergence of a parkinsonian disorder or dementia in 81% of older men initially diagnosed with idiopathic rapid eye movement sleep behavior disorder: a 16-year update on a previously reported series.* Sleep Medicine, 2013. **14**(8): p. 744-748.
302. Boeve, B., et al., *Synucleinopathy pathology and REM sleep behavior disorder plus dementia or parkinsonism.* Neurology, 2003. **61**(1): p. 40-45.
303. Claassen, D.O., et al., *REM sleep behavior disorder preceding other aspects of synucleinopathies by up to half a century.* Neurology, 2010. **75**(6): p. 494-499.
304. Mitsui, J., et al., *Mechanisms of Genomic Instabilities Underlying Two Common Fragile-Site-Associated Loci, PARK2 and DMD, in Germ Cell and Cancer Cell Lines.* American Journal of Human Genetics, 2010. **87**(1): p. 75-89.
305. Masuho, I., K. Xie, and K.A. Martemyanov, *Macromolecular Composition Dictates Receptor and G Protein Selectivity of Regulator of G Protein Signaling (RGS) 7 and 9-2 Protein Complexes in Living Cells.* Journal of Biological Chemistry, 2013. **288**(35): p. 25129-25142.
306. Anderson, G.R., et al., *R7BP Complexes With RGS9-2 and RGS7 in the Striatum Differentially Control Motor Learning and Locomotor Responses to Cocaine.* Neuropsychopharmacology, 2010. **35**(4): p. 1040-1050.
307. Fransson, A., A. Ruusala, and P. Aspenstrom, *Atypical Rho GTPases have roles in mitochondrial homeostasis and apoptosis.* Journal of Biological Chemistry, 2003. **278**(8): p. 6495-6502.
308. Guo, X.F., et al., *The GTPase dMiro is required for axonal transport of mitochondria to Drosophila synapses.* Neuron, 2005. **47**(3): p. 379-393.
309. Saotome, M., et al., *Bidirectional Ca2+-dependent control of mitochondrial dynamics by the Miro GTPase.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(52): p. 20728-20733.
310. Daniels, V., et al., *Insight into the mode of action of the LRRK2 Y1699C pathogenic mutant.* Journal of Neurochemistry, 2011. **116**(2): p. 304-315.
311. Lewis, P.A., et al., *The R1441C mutation of LRRK2 disrupts GTP hydrolysis.* Biochemical and Biophysical Research Communications, 2007. **357**(3): p. 668-671.
312. Liu, S., et al., *Parkinson's Disease-Associated Kinase PINK1 Regulates Miro Protein Level and Axonal Transport of Mitochondria.* Plos Genetics, 2012. **8**(3).
313. Puschmann, A., *Monogenic Parkinson's disease and parkinsonism: Clinical phenotypes and frequencies of known mutations.* Parkinsonism & Related Disorders, 2013. **19**(4): p. 407-415.
314. MacAskill, A.F. and J.T. Kittler, *Control of mitochondrial transport and localization in neurons.* Trends in Cell Biology, 2010. **20**(2): p. 102-112.

315. Nangaku, M., et al., *KIF1B, A NOVEL MICROTUBULE PLUS END-DIRECTED MONOMERIC MOTOR PROTEIN FOR TRANSPORT OF MITOCHONDRIA.* Cell, 1994. **79**(7): p. 1209-1220.
316. Tanaka, Y., et al., *Targeted disruption of mouse conventional kinesin heavy chain, kif5B, results in abnormal perinuclear clustering of mitochondria.* Cell, 1998. **93**(7): p. 1147-1158.
317. Morris, R.L. and P.J. Hollenbeck, *AXONAL-TRANSPORT OF MITOCHONDRIA ALONG MICROTUBULES AND F-ACTIN IN LIVING VERTEBRATE NEURONS.* Journal of Cell Biology, 1995. **131**(5): p. 1315-1326.
318. Ziegler, C.G.K., et al., *Parkinson's disease-like neuromuscular defects occur in prenyl diphosphate synthase subunit 2 (Pdss2) mutant mice.* Mitochondrion, 2012. **12**(2): p. 248-257.
319. Cheung, Z.H. and N.Y. Ip, *The emerging role of autophagy in Parkinson's disease.* Molecular Brain, 2009. **2**.
320. Hall, H., et al., *Hippocampal Lewy pathology and cholinergic dysfunction are associated with dementia in Parkinson's disease.* Brain, 2014. **137**: p. 2493-2508.
321. Rajput, A.H., et al., *Course in Parkinson disease subtypes A 39-year clinicopathologic study.* Neurology, 2009. **73**(3): p. 206-212.
322. Alcalay, R.N., et al., *Motor Phenotype of LRRK2 G2019S Carriers in Early-Onset Parkinson Disease.* Archives of Neurology, 2009. **66**(12): p. 1517-1522.
323. Kay, D.M., et al., *A comprehensive analysis of deletions, multiplications, and copy number variations in PARK2.* Neurology, 2010. **75**(13): p. 1189-1194.
324. Jarick, I., et al., *Genome-wide analysis of rare copy number variations reveals PARK2 as a candidate gene for attention-deficit/hyperactivity disorder.* Molecular Psychiatry, 2014. **19**(1): p. 115-121.
325. Pankratz, N., et al., *Copy Number Variation in Familial Parkinson Disease.* Plos One, 2011. **6**(8).
326. Pinto, D., et al., *Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants.* Nature Biotechnology, 2011. **29**(6): p. 512-U76.
327. Gibbs, R.A., et al., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-796.
328. Marchini, J., et al., *A new multipoint method for genome-wide association studies by imputation of genotypes.* Nature Genetics, 2007. **39**(7): p. 906-913.
329. Segre, A.V., et al., *Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits.* Plos Genetics, 2010. **6**(8).
330. Veyrieras, J.-B., et al., *High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation.* Plos Genetics, 2008. **4**(10).
331. Miller, J.D., et al., *Human iPSC-Based Modeling of Late-Onset Disease via Progerin-Induced Aging.* Cell Stem Cell, 2013. **13**(6): p. 691-705.
332. Frost, B., et al., *Tau promotes neurodegeneration through global chromatin relaxation.* Nature Neuroscience, 2014. **17**(3): p. 357-U48.
333. Jiang, N., et al., *Dietary and genetic effects on age-related loss of gene silencing reveal epigenetic plasticity of chromatin repression during aging.* Aging-Us, 2013. **5**(11): p. 813-824.
334. West, A.B., et al., *Parkinson's disease-associated mutations in LRRK2 link enhanced GTP-binding and kinase activities to neuronal toxicity.* Human Molecular Genetics, 2007. **16**(2): p. 223-232.
335. Soreq, L., et al., *Deep brain stimulation modulates nonsense-mediated RNA decay in Parkinson's patients leukocytes.* Bmc Genomics, 2013. **14**.

336. Piltonen, M., et al., *VASCULAR ENDOTHELIAL GROWTH FACTOR C ACTS AS A NEUROTROPHIC FACTOR FOR DOPAMINE NEURONS IN VITRO AND IN VIVO.* Neuroscience, 2011. **192**: p. 550-563.
337. Roukens, M.G., et al., *Functional Dissection of the CCBE1 Protein A Crucial Requirement for the Collagen Repeat Domain.* Circulation Research, 2015. **116**(10): p. 1660-1669.
338. Helmich, R.C., et al., *Pallidal Dysfunction Drives a Cerebellothalamic Circuit into Parkinson Tremor.* Annals of Neurology, 2011. **69**(2): p. 269-281.
339. Chen, R.H.C., et al., *alpha-Synuclein Membrane Association Is Regulated by the Rab3a Recycling Machinery and Presynaptic Activity.* Journal of Biological Chemistry, 2013. **288**(11): p. 7438-7449.
340. Gitler, A.D., et al., *The Parkinson's disease protein alpha-synuclein disrupts cellular Rab homeostasis.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(1): p. 145-150.
341. Yang, S.Y., et al., *Association between ST8SIA2 and the Risk of Schizophrenia and Bipolar I Disorder across Diagnostic Boundaries.* Plos One, 2015. **10**(9).
342. Mendiratta, S.S., et al., *A novel alpha-helix in the first fibronectin type III repeat of the neural cell adhesion molecule is critical for N-glycan polysialylation.* Journal of Biological Chemistry, 2006. **281**(47): p. 36052-36059.
343. Sato, C. and K. Kitajima, *Impact of structural aberrancy of polysialic acid and its synthetic enzyme ST8SIA2 in schizophrenia.* Frontiers in Cellular Neuroscience, 2013. **7**.
344. Livesey, F.J. and S.P. Hunt, *Netrin and netrin receptor expression in the embryonic mammalian nervous system suggests roles in retinal, striatal, nigral, and cerebellar development.* Molecular and Cellular Neuroscience, 1997. **8**(6): p. 417-429.
345. Pokinko, M., et al., *Resilience to amphetamine in mouse models of netrin-1 haploinsufficiency: role of mesocortical dopamine.* Psychopharmacology, 2015. **232**(20): p. 3719-3729.
346. Lesnick, T.G., et al., *A genomic pathway approach to a complex disease: Axon guidance and parkinson disease.* Plos Genetics, 2007. **3**(6): p. 984-995.
347. Lin, L., et al., *Axon guidance and synaptic maintenance: preclinical markers for neurodegenerative disease and therapeutics.* Trends in Neurosciences, 2009. **32**(3): p. 142-149.
348. Quilliam, L.A., J.F. Rebhun, and A.F. Castro, *A growing family of guanine nucleotide exchange factors is responsible for activation of Ras-family GTPases.* Progress in Nucleic Acid Research and Molecular Biology, Vol 71, 2002. **71**: p. 391-444.
349. Lai, Y.C., et al., *Phosphoproteomic screening identifies Rab GTPases as novel downstream targets of PINK1.* Embo Journal, 2015. **34**(22): p. 2840-2861.
350. Habig, K., et al., *LRRK2 guides the actin cytoskeleton at growth cones together with ARHGEF7 and Tropomyosin 4.* Biochimica Et Biophysica Acta-Molecular Basis of Disease, 2013. **1832**(12): p. 2352-2367.
351. Haebig, K., et al., *ARHGEF7 (BETA-PIX) Acts as Guanine Nucleotide Exchange Factor for Leucine-Rich Repeat Kinase 2.* Plos One, 2010. **5**(10).
352. Greggio, E., et al., *Kinase activity is required for the toxic effects of mutant LRRK2/dardarin.* Neurobiology of Disease, 2006. **23**(2): p. 329-341.
353. Dahl, A., et al., *A multiple-phenotype imputation method for genetic studies.* Nature Genetics, 2016. **48**(4): p. 466-+.
354. Zhou, X. and M. Stephens, *Genome-wide efficient mixed-model analysis for association studies.* Nature Genetics, 2012. **44**(7): p. 821-U136.

355. Max Kuhn. Contributions from Jed Wing, S.W., Andre Williams, Chris Keefer and Allan Engelhardt (2012). caret: Classification and Regression Training. R package version 5.15-044. http://CRAN.R-project.org/package=caret.
356. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of Statistical Software, 2010. **33**(1): p. 1-22.
357. Ng, S., *Variable Selection in Predictive Regressions.* Handbook of Economic Forecasting, Vol 2b, 2013: p. 753-789.
358. Keller, M.F., et al., *Using genome-wide complex trait analysis to quantify 'missing heritability' in Parkinson's disease.* Human Molecular Genetics, 2012. **21**(22): p. 4996-5009.
359. Hamza, T.H. and H. Payami, *The heritability of risk and age at onset of Parkinson's disease after accounting for known genetic risk factors.* Journal of Human Genetics, 2010. **55**(4): p. 241-243.
360. Marek, K., et al., *I-123 beta-CIT SPECT imaging assessment of the rate of Parkinson's disease progression.* Neurology, 2001. **57**(11): p. 2089-2094.
361. Ribeiro, M.J., et al., *Dopaminergic function and dopamine transporter binding assessed with positron emission tomography in Parkinson disease.* Archives of Neurology, 2002. **59**(4): p. 580-586.
362. Hettema, J.M., M.C. Neale, and K.S. Kendler, *A review and meta-analysis of the genetic epidemiology of anxiety disorders.* American Journal of Psychiatry, 2001. **158**(10): p. 1568-1578.
363. Kendler, K.S., et al., *A Swedish national twin study of lifetime major depression.* American Journal of Psychiatry, 2006. **163**(1): p. 109-114.
364. Arabia, G., et al., *Increased risk of depressive and anxiety disorders in relatives of patients with Parkinson disease.* Archives of General Psychiatry, 2007. **64**(12): p. 1385-1392.
365. Wang, C.S.M., et al., *Twin pairs discordant for neuropathologically confirmed Lewy body dementia.* Journal of Neurology Neurosurgery and Psychiatry, 2009. **80**(5): p. 562-565.
366. Nervi, A., et al., *Familial Aggregation of Dementia With Lewy Bodies.* Archives of Neurology, 2011. **68**(1): p. 90-93.
367. Lopes, R.D., C.; Defebvre, L.; Moonen, A. J.; Duits, A. A.; Hofman, P.; Leentjens, A. F.G.; Dujardin, K, *Cognitive phenotypes in parkinson's disease differ in terms of brain-network organization and connectivity.* Hum. Brain Mapp, 2016.
368. Adamczyk, M., et al., *Genetics of rapid eye movement sleep in humans.* Translational Psychiatry, 2015. **5**.
369. Ambrosius, U., et al., *Heritability of sleep electroencephalogram.* Biological Psychiatry, 2008. **64**(4): p. 344-348.
370. Watson, N.F., et al., *Genetic and environmental influences on insomnia, daytime sleepiness, and obesity in twins.* Sleep, 2006. **29**(5): p. 645-649.
371. Menza, M.A., D.E. Robertsonhoffman, and A.S. Bonapace, *PARKINSONS-DISEASE AND ANXIETY - COMORBIDITY WITH DEPRESSION.* Biological Psychiatry, 1993. **34**(7): p. 465-470.
372. Berardelli, A., et al., *Pathophysiology of bradykinesia in Parkinson's disease.* Brain, 2001. **124**: p. 2131-2146.
373. Louis, E.D., et al., *Clinical correlates of action tremor in Parkinson disease.* Archives of Neurology, 2001. **58**(10): p. 1630-1634.
374. Schormair, B., et al., *PTPRD (protein tyrosine phosphatase receptor type delta) is associated with restless legs syndrome.* Nature Genetics, 2008. **40**(8): p. 946-948.

375. Sevim, S., et al., *Correlation of anxiety and depression symptoms in patients with restless legs syndrome: a population based survey.* Journal of Neurology Neurosurgery and Psychiatry, 2004. **75**(2): p. 226-230.
376. Rana, A.Q., et al., *Association of restless legs syndrome, pain, and mood disorders in parkinson's disease.* International Journal of Neuroscience, 2016. **126**(2): p. 116-120.
377. Rana, A.Q., et al., *Increased likelihood of anxiety and poor sleep quality in Parkinson's disease patients with pain.* Journal of the Neurological Sciences, 2016. **369**: p. 212-215.
378. Bifsha, P., et al., *Rgs6 is Required for Adult Maintenance of Dopaminergic Neurons in the Ventral Substantia Nigra.* Plos Genetics, 2014. **10**(12).
379. Maity, B., et al., *Regulator of G Protein Signaling 6 (RGS6) Protein Ensures Coordination of Motor Movement by Modulating GABA(B) Receptor Signaling.* Journal of Biological Chemistry, 2012. **287**(7): p. 4972-4981.
380. Pluck, G.C. and R.G. Brown, *Apathy in Parkinson's disease.* Journal of Neurology Neurosurgery and Psychiatry, 2002. **73**(6): p. 636-642.
381. Martinez-Horta, S., et al., *Apathy in Parkinson's Disease: Neurophysiological Evidence of Impaired Incentive Processing.* Journal of Neuroscience, 2014. **34**(17): p. 5918-5926.
382. Farkas, L.M., et al., *Transforming growth factor-beta s are essential for the development of midbrain dopaminergic neurons in vitro and in vivo.* Journal of Neuroscience, 2003. **23**(12): p. 5178-5186.
383. Roussa, E., L.M. Farkas, and K. Krieglstein, *TGF-beta promotes survival on mesencephalic dopaminergic neurons in cooperation with Shh and FGF-8.* Neurobiology of Disease, 2004. **16**(2): p. 300-310.
384. Groppe, J., et al., *Structural basis of BMP signaling inhibition by noggin, a novel twelve-membered cystine knot protein.* Journal of Bone and Joint Surgery-American Volume, 2003. **85A**: p. 52-58.
385. Guo, W.X., et al., *RNA-Binding Protein FXR2 Regulates Adult Hippocampal Neurogenesis by Reducing Noggin Expression.* Neuron, 2011. **70**(5): p. 924-938.
386. Hsu, L.J., et al., *Transforming Growth Factor beta 1 Signaling via Interaction with Cell Surface Hyal-2 and Recruitment of WWOX/WOX1.* Journal of Biological Chemistry, 2009. **284**(23): p. 16049-16059.
387. Wang, H.Y., et al., *WW domain-containing oxidoreductase promotes neuronal differentiation via negative regulation of glycogen synthase kinase 3 beta.* Cell Death and Differentiation, 2012. **19**(6): p. 1049-1059.
388. Chang, J.-Y.C., N-S, *WWOX dysfunction induces sequential aggregation of TRAPPC6AΔ, TIAF1, tau and amyloid β, and causes apoptosis.* Cell Death Discovery, 2015. **1**: p. 15003.
389. Janssens, A., et al., *The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases.* Genetics in Medicine, 2007. **9**(8): p. 528-535.
390. Wray, N.R., et al., *Pitfalls of predicting complex traits from SNPs.* Nature Reviews Genetics, 2013. **14**(7): p. 507-515.
391. Dudbridge, F., *Power and Predictive Accuracy of Polygenic Risk Scores.* Plos Genetics, 2013. **9**(3).