Original research

Genetics of validated Parkinson's disease subtypes in the Oxford Discovery and Tracking Parkinson's cohorts

Michael Lawton ,[1] Manuela MX Tan ,[2,3] Yoav Ben-Shlomo,[1] Fahd Baig,[4,5] Thomas Barber,[4,6] Johannes C Klein ,[4,6] Samuel G Evetts,[4,6] Stephanie Millin,[6,7] Naveed Malek,[8] Katherine Grosset,[9] Roger A Barker,[10] Nigel Williams,[11] David J Burn,[12] Thomas Foltynie ,[2] Huw R Morris ,[2,3] Nicholas Wood,[2] Donald G Grosset,[9] Michele Tao-Ming Hu [4,6]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/jnnp-2021-327376).

For numbered affiliations see end of article.

Correspondence to
Michael Lawton, Population Health Sciences, University of Bristol Medical School, Bristol, Bristol, UK; Michael.Lawton@bristol.ac.uk

DGG and MT-MH contributed equally.

DGG and MT-MH are joint senior authors.

Received 14 June 2021
Accepted 25 May 2022
Published Online First 21 June 2022

Check for updates

To cite: Lawton M, Tan MMX, Ben-Shlomo Y, *et al*. *J Neurol Neurosurg Psychiatry* 2022;**93**:952–959.

ABSTRACT

Objectives To explore the genetics of four Parkinson's disease (PD) subtypes that have been previously described in two large cohorts of patients with recently diagnosed PD. These subtypes came from a data-driven cluster analysis of phenotypic variables.

Methods We looked at the frequency of genetic mutations in glucocerebrosidase (GBA) and leucine-rich repeat kinase 2 against our subtypes. Then we calculated Genetic Risk Scores (GRS) for PD, multiple system atrophy, progressive supranuclear palsy, Lewy body dementia, and Alzheimer's disease. These GRSs were regressed against the probability of belonging to a subtype in the two independent cohorts and we calculated q-values as an adjustment for multiple testing across four subtypes. We also carried out a Genome-Wide Association Study (GWAS) of belonging to a subtype.

Results A severe disease subtype had the highest rates of patients carrying GBA mutations while the mild disease subtype had the lowest rates (p=0.009). Using the GRS, we found a severe disease subtype had a reduced genetic risk of PD (p=0.004 and q=0.015). In our GWAS no individual variants met genome wide significance (<5×10e-8) although four variants require further follow-up, meeting a threshold of <1×10e-6.

Conclusions We have found that four previously defined PD subtypes have different genetic determinants which will help to inform future studies looking at underlying disease mechanisms and pathogenesis in these different subtypes of disease.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Data-driven approaches have been used to generate Parkinson's disease subtypes in many studies but little is known about the genetics of these subtypes.

WHAT THIS STUDY ADDS

⇒ We found in previously developed Parkinson's subtypes that a severe disease subtype had the highest rates of glucocerebrosidase mutation carriers and the lowest genetic risk within a Parkinson's Genetic Risk Score.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE AND/OR POLICY

⇒ These results provide some biological validity to our data-driven subtyping approach and will assist in future studies looking at underlying disease mechanisms and pathogenesis.

INTRODUCTION

Parkinson's disease (PD) is a common and progressive neurodegenerative disorder encompassing a wide range of motor and non-motor features. There is considerable heterogeneity within these features in terms of presentation and progression which has led many to believe there are different clinically relevant subtypes of the disease. Data-driven approaches have been applied to many PD cohorts to try and delineate these subtypes, the first was in 1999[1] and three systematic reviews have since been published.[2–4] Other hypothesis driven approaches have also been studied in PD,[2,3] the most commonly studied is the tremor-dominant (TD) versus postural instability gait difficulty (PIGD) motor subtype[5,6] and another of interest is splitting into young-onset versus late-onset PD.[7,8]

We previously derived Parkinson's clinical subtypes in over 2500 early patients with PD recruited from two large cohorts: Oxford Discovery and Tracking Parkinson's.[9,10] These subtypes were derived from the baseline motor and non-motor features using a data-driven approach, which were associated with subsequent motor progression and the medication response. We have recently shown, within the Oxford Discovery cohort, that one of our subtypes had a distinctive biomarker profile with reduced apolipoprotein A1 and increased C reactive protein levels, lending biological validity to our approach.[11]

Considering differences in genetics might help determine any difference in the aetiology of the subtypes while also providing a biological confirmation of data-driven clustering approaches. Here, we report on the genetics of our validated PD subtypes using data from the Oxford Discovery and Tracking Parkinson's cohorts combined. To calculate the genetic risk of PD and related conditions including the atypical parkinsonian disorders and Alzheimer's disease (AD), we identified a Genome Wide Association Study (GWAS) of disease status (an analysis of case/control status) for each of the diseases. We then looked at whether the genetic risk of PD and related disorders was associated with belonging to a particular disease subtype. We also considered two of the most important mutations in PD, glucocerebrosidase (GBA) and leucine-rich repeat kinase 2 (LRRK2), against our subtypes. Finally, we carry out a GWAS study to see whether any individual genetic variants are associated with belonging to a subtype.

A recent GWAS study has been published based on the TD and PIGD motor subtypes which found multiple PD risk alleles that might influence the motor subtype.[12] We have recently published a GWAS study using data from the Oxford Discovery, Tracking Parkinson's and PPMI cohorts to look at motor and cognitive progression which found that APOE ε4 influences progressive cognitive impairment.[13] This study differs to our previous one as its focus is on data-driven PD subtypes.

METHODS

Cohorts

We used data from two large prospective early PD cohorts. The Tracking Parkinson's cohort includes UK-wide centres, recruited between February 2012 and May 2014. Full details of this cohort along with inclusion/exclusion criteria have been published previously.[14] The Oxford Discovery cohort includes patients from 11 hospitals in the Thames Valley region recruited between September 2010 and January 2016.[15] In both cohorts, patients were recruited within 3.5 years of diagnosis, and both studies were funded by Parkinson's UK. Both studies had ethical approval and were undertaken with the understanding and written consent of each subject. Patients are followed up every 18 months collecting a wide range of data in motor, non-motor and cognitive domains. For brevity, we will refer to the Tracking Parkinson's cohort as Tracking.

Patient evaluation

Our data-derived PD subtypes were determined using variables from motor, non-motor and cognitive domains at baseline. Our clustering approach used a factor analysis followed by a k-means cluster analysis where we considered two to five clusters. Individuals were excluded from the cluster analysis if they had been rediagnosed with another condition during follow-up or if they had been given a probability of a diagnosis of PD of <90% at the latest visit as rated by a research neurologist or movement disorder specialist. This was an attempt to exclude those incorrectly diagnosed with PD.

Our first paper on this subject was based on only the Oxford Discovery cohort (with 769 patients) and we found five clusters gave us the optimal solution.[9] In our second paper we used two cohorts where the Tracking cohort (n=1601) was chosen to be the development cohort (as it was larger) and the Oxford Discovery cohort (n=944) was the validation cohort.[10] Here, we identified that four clusters were the optimal solution. Comparing the actual and predicted clusters (from a discriminant analysis model fitted to the Tracking clusters) in Oxford Discovery gave us a kappa statistic of 0.58 indicating moderate agreement, providing evidence our cluster approach was moderately stable across the two cohorts. These four clusters (derived using only baseline data) were shown to be associated with different subsequent motor progression over an average of 3 years follow-up and also with medication response using a levodopa challenge. We also found differences in age, gender, Hoehn and Yahr stage as well as TD/PIGD rates between the clusters which were all factors not included in the cluster analysis. The identified clusters were named (1) fast motor progression with symmetrical motor disease, poor olfaction, cognition and postural hypotension; (2) mild motor and non-motor disease with intermediate motor progression; (3) severe motor disease, poor psychological well-being and poor sleep with an intermediate motor progression and (4) slow motor progression with TD, unilateral disease. When we talk about mild/severe disease we are classifying the cross-sectional associations of data at baseline while fast/slow refers to progression rates after baseline so fast/severe and slow/mild can be thought of as different clusters. In this paper, we describe the genetics of the four subtypes (also referred to as clusters since they were developed using cluster analysis) from our development/validation paper.[10] Within the Oxford Discovery cohort we report on the predicted clusters since any future research on individuals outside of these cohorts would rely on predictions.

Genotyping

In the Tracking Parkinson's cohort, individuals were genotyped using the Illumina HumanCore Exome array with custom content.[14] Within the Oxford Discovery cohort individuals were genotyped on either the Illumina HumanCore Exome-12 V.1.1[16] or the Illumina InfiniumCore Exome-24 V.1.1[17] singl-nucleotide polymorphisms (SNP) arrays. The quality control and imputation of this data has been previously described[13] and is also described in the online supplemental file 1.

In a principal components (PCs) analysis, 20 genetic PCs were generated from a linkage-pruned SNP set (removing SNPs with an r^2 >0.02 in a 1000 kb sliding window shifting 10 SNPs at a time). If an individual was >6 SDs from the mean of one of the first 5 PCs or a clear outlier in a scatter plot they were excluded and then the PCs recalculated and repeated until there were no outliers. The first five PCs were then retained to be included as covariates within the GWAS.

Our main focus was to look at genetic risk of PD but we also wanted to explore whether they might be shared genetic pathways between other neurodegenerative disorders (progressive supranuclear palsy (PSP), multiple system atrophy (MSA), Lewy body dementia (LBD) and AD) and each subtype while also exploring the potential for selection bias where atypical parkinsonian disorders might have been incorrectly diagnosed as PD. To calculate the genetic risk of each condition we identified an external GWAS of disease status (an analysis of case/control status) applied to separate PD, MSA, PSP, LBD and AD cohorts.[18–22] Overlap in genetic pathways and risk has been described previously for LBD, Parkinsons and Alzheimer's.[21] The PD GWAS[19] reports that applying a Genetic Risk Score (GRS) using the genome-wide significant hits explained a minimum of 16% of the genetic liability and led to an AUC of 0.651.

GBA mutations were split into those that are recognised as causing Gaucher's disease (GD) (the most common being L444P and N370S) and those that are not (E326K and T369M) as previously reported from Tracking.[23] For LRRK2, we identified carriers of the G2019S and R1441C mutations, as reported previously from Tracking.[24] In Oxford Discovery, carriers of L444P and R1441C mutations were identified by PCR as previously reported[25] and the other mutations were identified from the Neurochip[26] which is a custom-designed array for the investigation of genetic variation in neurodegenerative diseases and can detect rare variants within the LRRK2 and GBA genes. The Neurochip data underwent similar Quality control to the array data described above and is also described in the online supplemental file 1. In Oxford Discovery, we have carried out Sanger sequencing to confirm the N370S and E326K mutation carriers.

All those who underwent Sanger sequencing had the mutation confirmed, however, two of the N370S carriers have not yet had Sanger sequencing. The numbers with other monogenetic forms of PD such as PRKN, SNCA and PINK1 were too small to draw any conclusions, see discussion.

Statistical analysis

We tabulated the clusters against LRRK2 and GBA status using a Fisher's exact test (since the frequencies are very small in some cells due to the rarity of these mutations) to determine the strength of any association.

We calculated the probability of belonging to a cluster from the discriminant analysis model from our validated subtypes paper.[10] This probability was converted to log odds to give a more suitable continuous score for linear regression (unbounded range and symmetrical).

In an attempt to assess the potential for selection bias we compared age (t-test), gender and cluster assignment (χ^2 test) for those who did and did not have genetic data from the SNP arrays after quality control.

We calculated GRS for PD, PSP, MSA, LBD and AD by multiplying the genome wide significant SNPs ($p<5\times10e\text{-}8$) by their beta coefficients taken from each external GWAS and then standardising the score. This GRS can be interpreted as an estimate of the contribution of genetics to developing one of these diseases.[27] Since the MSA GWAS did not find any genome wide significant SNPs we used those reported at a threshold of $<1\times10e\text{-}6$ to calculate the GRS[20] and in Alzheimer's we used two variants that were from previously reported genome-wide significant loci but did not reach significance in the current GWAS.[22] The number of SNPs from each GRS are reported in online supplemental table 1, which are the number of SNPs reaching the thresholds specified above in each external GWAS that were also available in our genetic data. Then we used linear regression with log odds of belonging to a cluster as the outcome and each GRS as the exposure. This was carried out separately within each cohort and then the results were combined using a fixed effects meta-analysis. We used a false discovery rate method,[28] often called the Benjamini-Hochberg method, to control for multiple comparisons across the four subtypes.[11] Using this method in our GRS analyses we have derived q-values. If our significance threshold was 0.05 we would hope to find q values <0.05. These q-values do not have a simple probabilistic interpretation, it is only important whether they reach the chosen threshold. The authors are aware of problems using corrections to p values[29] and focusing on statistical significance at an arbitrary 0.05 threshold.[30 31] We have tried to not use language like significant and non-significant, instead p values should be viewed by the reader as a continuum where smaller p values represent greater evidence against the null hypothesis and confidence intervals should be examined for the strength of any association. In the results, we have pointed out the direction of some associations and using the derived p values and q-values the reader can decide for themselves the strength of evidence against the null hypothesis. We hope this approach will promote modern thinking that arbitrary p value thresholds are unhelpful.

We carried out a GWAS with linear regression using the logs odds of belonging to a cluster as the outcome. The first five genetic PCs were used as covariates for each regression. Only SNPs with a minor allele frequency (MAF) >0.05 were included. The data were combined using a fixed effects meta-analysis. We also computed the expected power for our sample size[32] for a range of beta and MAFs. The number of SNPs within the GWAS are reported in the online supplemental file 1.

Palindromic SNPs (where the alleles are nucleotides that pair to each other making it difficult to determine the direction of effect) that had an MAF >0.45 were excluded when calculating the GRS and also from the GWAS.

RESULTS

Demographics and potential for selection bias

After all the quality control procedures, we had genetic data on 1467 derived from 1601 (91.6%) individuals from the original Tracking cluster analysis. Average age (67.2 vs 68.0 with p=0.31) and gender rates (34.2% vs 34.3% female with p=0.97) were similar in those with and without genetic data (respectively). Looking within clusters rates of those included varied from 96.8% (cluster 4) to 88.6% (cluster 1) with a p=0.001. For those with genetic data there were 437, 423, 304 and 303 individuals in clusters 1–4, respectively.

In the Oxford Discovery cohort, we had genetic data on 807 individuals, out of 944 (85.5%) individuals from the cluster analysis. Within Oxford Discovery average age (67.4 vs 66.1 with p=0.15) and gender rates (34.3% vs 41.6% female with p=0.099) were similar in those with and without genetic data (respectively). Looking within clusters rates of those included varied from 87.5% (cluster 4) to 83.0% (cluster 3) with a p =value 0.53. For those with genetic data there were 261, 145, 185 and 216 individuals in clusters 1–4, respectively.

Mutation carriers

Table 1 shows the associations between LRRK2 and GBA mutation carriers against the clusters in both cohorts. In the Tracking cohort the third cluster (severe motor disease and poor psychological well-being) had the largest proportion of LRRK2 carriers (1.9%), however, this is not replicated in Oxford Discovery where the third cluster has no carriers. The combined cohort p value of LRRK2 vs the clusters was p=0.35.

Within the Tracking cohort the third disease cluster (severe motor disease and poor psychological well-being) had the greatest proportion of GBA carriers (12.9% across both carrier groups) and the second disease cluster (mild motor and non-motor disease) had the lowest proportion of GBA carriers (6.3%). This trend was also seen in Oxford Discovery cohort (11.3% in cluster 3 vs 6.6% in cluster 2). In the combined cohorts a p value for a difference in GBA carrier rates across the clusters was p=0.036, and when combining the two GBA carrier groups the p value was smaller at p=0.009.

Genetic risk of diseases

Genetic PD risk (see figure 1) is positively associated with belonging to clusters 2 (mild motor and non-motor disease) (pooled p=0.044 and q=0.059) and 4 (slow motor progression), (pooled p=0.021 and q=0.043), while it is negatively associated with belonging to cluster 3 (severe motor disease and poor psychological well-being) (p=0.004 and q=0.015). For the pooled associations a one SD change in the PD GRS was associated with a 0.2 (95% CI 0.00 to 0.39) increase in the log odds of belonging to cluster 2; 0.2 (95% CI 0.03 to 0.37) increase for cluster 4 and a 0.3 (95% CI 0.10 to 0.51) decrease for cluster 3. We also explored a sensitivity analysis where we adjusted for the GBA mutation carrier groups and found very similar results (see online supplemental figure 1).

We can see in figure 2 that within the Oxford Discovery cohort genetic PSP risk is negatively associated with cluster 2

Table 1 Data-derived clusters compared with LRRK2 and GBA mutation status

	LRRK2			GBA		
	Non-carriers	Carriers		Non-carriers	E326K and T369M carriers	GD-causing variants
Tracking Parkinson's cohort						
Cluster 1	469 (99.8%)	1 (0.2%)	Cluster 1	437 (91.8%)	29 (6.1%)	10 (2.1%)
Cluster 2	432 (99.1%)	4 (0.9%)	Cluster 2	413 (93.7%)	20 (4.5%)	8 (1.8%)
Cluster 3	314 (98.1%)	6 (1.9%)	Cluster 3	282 (87.0%)	27 (8.3%)	15 (4.6%)
Cluster 4	304 (99.7%)	1 (0.3%)	Cluster 4	280 (90.9%)	20 (6.5%)	8 (2.6%)
P=0.059			P=0.080			
			P value (GBA variants combined)=0.018			
Oxford discovery cohort						
Cluster 1	280 (99.3%)	2 (0.7%)	Cluster 1	231 (90.9%)	15 (5.9%)	8 (3.2%)
Cluster 2	150 (98.7%)	2 (1.3%)	Cluster 2	127 (93.4%)	8 (5.9%)	1 (0.7%)
Cluster 3	204 (100%)	0	Cluster 3	158 (88.8%)	14 (7.9%)	6 (3.4%)
Cluster 4	221 (99.1%)	2 (0.9%)	Cluster 4	185 (90.7%)	16 (7.8%)	3 (1.5%)
P=0.45			P=0.57			
			P value (GBA variants combined)=0.59			
Combined cohort						
Combined cohort p=0.35			Combined cohort p=0.036			
			Combined cohort p value (GBA variants combined)=0.009			

Note the numbers in this table are slightly different to the numbers in the other analyses since the mutation status did not come from the imputed array data.
GBA, glucocerebrosidase; GD, Gaucher's disease; LRRK2, leucine-rich repeat kinase 2.

(mild motor and non-motor disease) (p=0.006 and q=0.024) and positively associated with cluster 3 (severe motor disease and poor psychological well-being) (p=0.014 and q=0.027). However within the Tracking cohort the association between PSP with cluster 2 (mild motor and non-motor disease) is much smaller (-0.04 vs −0.42) and for cluster 3 (severe motor disease) it is within the opposite direction (-0.12 vs 0.41). When compared with the Oxford Discovery cohort the pooled p values and q values are much larger for both cluster 2 (p=0.046 and q=0.18) and cluster 3 (p=0.38 and q=0.70). Also within figure 2, we can see that genetic MSA risk is negatively associated with belonging to cluster 4 (slow motor progression) (pooled p=0.020 and q=0.079) where a 1 SD change in the GRS was associated with a 0.20 (95% CI 0.03 to 0.37) decrease in the log odds of belonging to cluster 4.

In figure 3, we can see that the associations of the clusters with genetic risk of LBD and AD look very similar (especially for clusters 1, 2 and 4). Cluster 2 (mild motor and non-motor disease) is inversely associated with both LBD and AD in Tracking but not within Oxford Discovery. Within the AD GWAS the APOE genetic variant has an effect size much higher than all the others (OR of 3.32 compared with an average of 1.27 when the direction of effect is coded as positive) so we also explored what would happen when that variant is removed (see online supplemental figure 2). When removing this variant cluster 1 (fast motor progression) is positively associated with AD (pooled p=0.063 and q=0.25) where a one SD change in the GRS was associated with a 0.13 (−0.01 to 0.28) increase in the log odds of belonging to cluster 1.

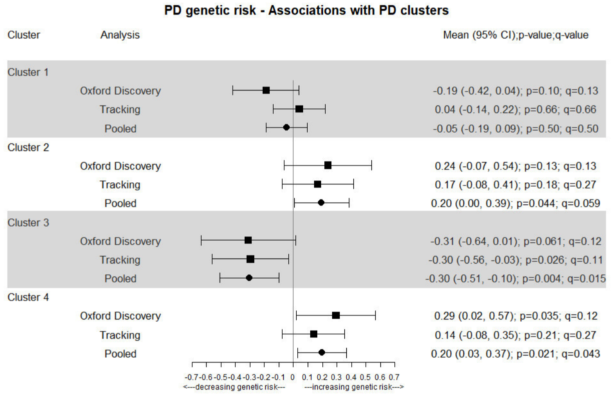


Figure 1 Genetic risk of Parkinson's disease (PD) versus likelihood of belonging to a cluster.

Genome Wide Association Study

There was little evidence of population stratification since within the four GWAS analyses from Tracking, the genomic inflation factor lambda varied from 1.001 to 1.008, while within Oxford Discovery they were all 1.0.

We highlight the power we have to detect a genome wide significant variant given our sample size in online supplemental table 2. Generally our power is small to detect rare variants with high effect sizes or common variants with small effect sizes. Since we found no genome wide significant variants in table 2 we highlight (non-independent) variants that reached a threshold of <1×10e-6, similar to the MSA GWAS study.[20] At this threshold we identified 3 SNPs that were associated with cluster 1. The QQ-plot for this cluster (see online supplemental figure 3) shows a hump at the upper end which implies an excess of genetic variants associated with phenotypic cluster 1 at lower p value levels (0.0001–0.000001). We had one SNP at the reduced threshold for cluster 3 and none for clusters 2 and 4. None of the other QQ-plots (see online supplemental figures 4–6) show evidence of there being an excess of variants associated with any phenotypic cluster. The cohort specific results from table 2 can be found in online supplemental table 3). In the online supplemental file 1 the biological relevance of the identified SNPs are reported along with some network analyses (none of which met a threshold of Bonferroni adjusted-value of 0.05 shown in online supplemental figures 7–10).

DISCUSSION

The associations between GBA and the phenotypic clusters, with a severe disease cluster having the greatest proportion of carriers

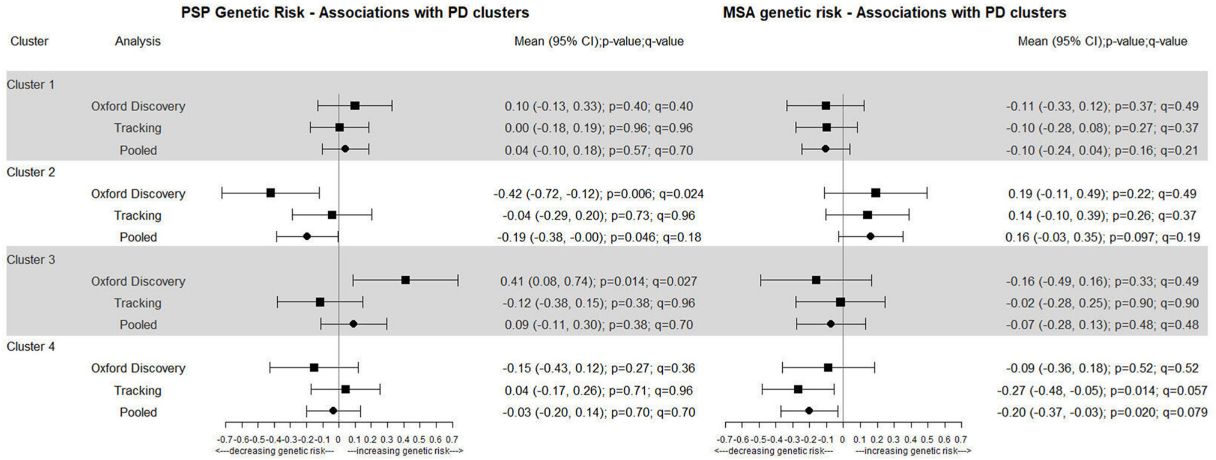


Figure 2 Genetic risk of atypical Parkinson's: progressive supranuclear palsy (PSP) and multiple system atrophy (MSA).

and a mild disease cluster having the smallest proportion, are what would be expected given the observational evidence that GBA mutations are associated with higher Hoehn and Yahr stage and worse cognition.[33–36] GD-causing and GBA risk variants such as E365K (E326K) have also been associated with more rapid motor and cognitive impairment in PD in other studies.[37] This has been hypothesised to relate to lysosomal dysfunction and the more rapid accumulation of pathogenic alpha-synuclein species in patients with carrying GBA variants.[38] However, there are also reports that GBA mutations are associated with earlier disease onset while cluster 3 has the most GBA mutations and a higher than average age at diagnosis and cluster 2 has the least GBA mutations and the lowest average age at diagnosis.[10] This highlights that there is still heterogeneity of disease onset within the clusters and that GBA mutation carriers are only a small proportion (~12%) of even the cluster with the highest carrier rate. We hypothesise that other similar genetic variants are associated with the severe disease cluster that may relate to impaired proteostasis and/or lysosomal dysfunction.

There is also heterogeneity of clinical phenotype within LRRK2 carriers which would make it difficult to correlate them with clusters. One study showed that mutations of the LRRK2 gene are associated with less cognitive impairment compared with iPD[39] while others have failed to confirm this.[40 41] A study of LRRK2 found a slower decline in UPDRS scores[42] and another found no discernible effect on rate of motor disease progression.[43]

There are several possible explanations for the negative association between genetic risk of PD and the third, severe disease cluster. The first is that the individuals in this cluster have a more environmental and less genetically driven disease aetiology. The second is that this cluster is enriched with non-PD cases although the MSA and PSP genetic risk pooled associations do not support this, and it would also require that the PD GWAS studies had no enrichment of other similar conditions. The third is one of selection bias, in that these severe disease cases are less likely to participate in the PD cohorts that supply cases to the PD GWAS study we used, as compared with Oxford Discovery and Tracking cohorts which offered local clinical review for the majority of research participants. This PD GWAS study used data from 17 different datasets.[19] Note that the GRS came from the imputed genetic data which excludes rarer genetic variants such as those within the GBA gene. The severe disease cluster has low genetic risk of PD looking at common variants yet the rare GBA variants have the highest frequency within this cluster.

We have data on other monogenetic forms of Parkinson's (SNCA, PRKN and PINK1) and have published this data from the Tracking cohort.[24] However, the numbers are too small to draw any conclusions against our clusters. Only one individual from the Tracking clusters have a biallelic PINK1 mutation, none

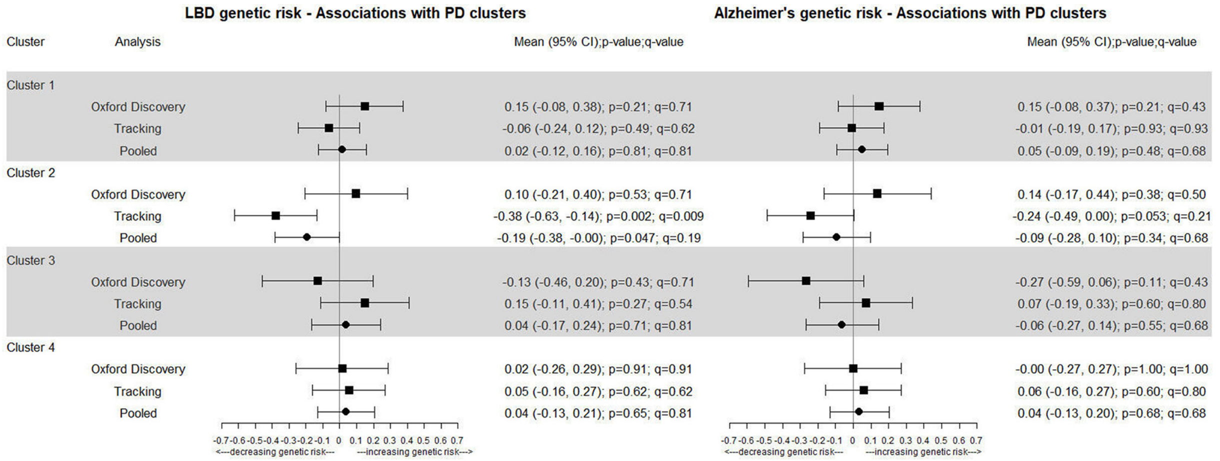


Figure 3 Genetic risk of dementia: Alzheimer's disease and Lewy body dementia (LBD). PD, Parkinson's disease.

Table 2 SNPs meeting a threshold of 1×10e-6 from the genome wide association study meta-analysis for each data-driven cluster

Chr	Position (GRCh37)	Marker	A1	A2	Nearest gene	Beta	SE	P value
Cluster 1								
1	237 734 615	rs151043031	CT	C	RYR2	0.59	0.12	9.986e-07
6	160 698 177	rs316037	G	A	SLC22A2	0.60	0.12	9.867e-07
6	160 699 605	rs5881357	AT	A	SLC22A2	0.60	0.12	6.337e-07
Cluster 2—no SNPs met threshold								
Cluster 3								
1	214 449 747	rs116258323	T	C	SMYD2	1.62	0.33	6.715e-07
Cluster 4—no SNPs met threshold								

A1, effect allele; A2, other allele; Chr, chromosome; SE, SE error; SNPs, single-nucleotide polymorphisms.

had a biallelic PRKN mutation and only one a SNCA mutation. In the Oxford Discovery cohort, we have data within these genes from the Neurochip but again the numbers are too small to draw any conclusions, no one from the cluster analysis had a SNCA or a biallelic PRKN mutation and only one individual had a biallelic PINK1 mutation.

The negative association between genetic risk of PSP and cluster 2 and the positive association with cluster 3 in the Oxford Discovery cohort is what we would expect to see if there was enrichment of PSP cases. That is, PSP cases are more likely to belong to a severe motor disease cluster than a mild motor and non-motor disease cluster. However, this is not backed up by the associations within the larger Tracking cohort. This could represent a chance finding in Oxford Discovery. Alternatively, it could reflect the procedure we used to exclude patients from the analysis, that is dropping those with probability of diagnosis of PD of <90% at the latest clinic visit. In Tracking 367/1975 (18.6%) were dropped, while in Oxford Discovery only 76/1022 (7.4%) were dropped using this criterion (see online supplemental figure 1 in the original paper[10]). Since a greater proportion were dropped in Tracking it is more likely that we have excluded PSP cases from this cohort. The reported PD disease probability would, in all likelihood, be reduced if the clinician documented features consistent with atypical parkinsonism during the clinical review, including the presence of symmetrical motor disease, early onset falls, suboptimal levodopa response, a supranuclear gaze palsy or early autonomic failure.

In previous research, we found cluster 3 was associated with a higher proinflammatory baseline profile (raised CRP, reduced apolipoprotein A1). This is interesting, as it suggests that in PD subtype 3—who have greater rates of cognitive dysfunction—early immune modulation might improve clinical outcomes, for example, by reducing future dementia risk if commenced early enough in the disease process. The lower overall genetic risk of PD and a higher pro-inflammatory profile in this cluster, are consistent with a hypothesis that the aetiology of this cluster is more driven by environmental rather than genetic risk factors.

Although none of our individual variants met the GWAS p value significance threshold the ones that we highlight might be interesting for future follow-up and research. It could be that the variants, or closest genes to these variants, are a reason that a person develops a particular subtype of Parkinson's.

In previous research, we used multinomial logistic regression to look at how blood biomarkers are associated with an individual belonging to one of the clusters.[11] For this genetic analysis, we decided to simplify the analyses by carrying out four separate analyses using the probability of belonging to the cluster as the outcome. This made the GWAS easier to run and interpret with fewer variables to estimate.

The strengths of this study are we have used two large early in the disease course and well-phenotyped PD cohorts. Our subtypes were created using large amounts of phenotypic data incorporating 21 variables across 12 important domains and these subtypes were developed and validated in over 2500 subjects. These subtypes were shown to be associated with both motor progression and medication response in a levodopa challenge. The limitations of this study are that in terms of searching for individual genetic variants it is still too small to find any that reach genome wide significance, assuming that such variants exist. Also there is the possibility of selection bias as rates of those with genetic data varied by cluster within the Tracking cohort. The frequency of PD subtypes in our cohorts may be different to that in the general PD population if belonging to a subtype was related to agreeing to take part in our cohorts or our cohorts failed to identify specific individuals during recruitment. However, to bias our estimates of genetics versus the clusters, it would require that selection into our cohorts was also related to an individual's genetics. Diagnosis of PD will not be perfect and some patients will turn out to have other parkinsonian disorders, although we have attempted to mitigate this by excluding individuals with a diagnostic probability of PD <90% at the latest visit.

There are other subtypes that have been defined by a data-driven cluster analysis on motor and non-motor symptomatic data. Currently, it is difficult to determine whether the cluster definition we have used is more robust or superior to other definitions. However, in a recent systematic review our paper was rated (among 25 other data-driven studies) along with two others as having the highest methodological quality and clinical applicability.[2] What sets our cluster definition apart is our use of an external validation.

Future work is now ongoing to understand the underlying disease pathophysiology driving these different clinical clusters in early PD, and their subsequent progression. This will use a mechanistic approach comparing lysosomal, mitochondrial, inflammatory function, α-synuclein (α-syn) seeding amplification[44] and α-omics profiles across the four PD clinical clusters.

The differences in genetics between these clusters lends biological validity to our data-driven clustering approach while also providing evidence that the different subtypes can inform on underlying disease mechanisms and pathogenesis, as well as informing individual disease trajectories in PD.

Author affiliations
[1]Population Health Sciences, University of Bristol Medical School, Bristol, UK
[2]Department of Clinical and Movement Neurosciences, Queen Square Institute of Neurology, University College London, London, UK
[3]UCL Movement Disorders Centre, University College London, London, UK
[4]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK

[5]Molecular and Clinical Sciences Institute, St. George's University of London, London, UK
[6]Oxford Parkinson's Disease Centre, University of Oxford, Oxford, UK
[7]Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, UK
[8]Department of Neurology, Queen's Hospital, Romford, Essex, UK
[9]Department of Neurology, Institute of Neurological Sciences, Queen Elizabeth University Hospital and University of Glasgow, Glasgow, UK
[10]Cambridge Centre for Brain Repair, University of Cambridge, Cambridge, UK
[11]Psychological Medicine and Clinical Neurosciences, Cardiff University, Cardiff, UK
[12]Faculty of Medical Sciences, Newcastle University, Newcastle, UK

Twitter Thomas Foltynie @foltynie

Acknowledgements The Oxford Discovery study was funded by the Monument Trust Discovery Award from Parkinson's UK and supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre based at Oxford University Hospitals NHS Trust and University of Oxford, and the NIHR Clinical Research Network: Thames Valley and South MidlandsThe Tracking Parkinson's study was funded by Parkinson's UK and supported by the National Institute for Health Research (NIHR) DeNDRoN network, the NIHR Newcastle Biomedical Research Unit based at Newcastle upon Tyne Hospitals NHS Foundation Trust and Newcastle University, and the NIHR funded Biomedical Research Centre in Cambridge (Grant number:146281).

Contributors ML: analysis and interpretation of the data, writing of the manuscript. ML acts as guarantor and accepts full responsibility for the finished work and/or the conduct of the study, had access to the data, and controlled the decision to publish. MT: analysis and interpretation of the data, revision of the manuscript. YB-S: study concept and design, analysis and interpretation of the data, revision of the manuscript. FB: acquisition of data, revision of the manuscript. TB: acquisition of data, revision of the manuscript. JCK: acquisition of data, revision of the manuscript. SGE: analysis and interpretation of data, revision of the manuscript. SM: analysis and interpretation of data, revision of the manuscript. NM: acquisition of data, revision of the manuscript. KG: study concept and design, acquisition of data, revision of the manuscript. RAB: study concept and design, acquisition of data, revision of the manuscript. NW: study concept and design, revision of the manuscript. DJB: study concept and design, acquisition of data, revision of the manuscript. TF: study concept and design, acquisition of data, revision of the manuscript. HRM: study concept and design, acquisition of data, revision of the manuscript. NW: study concept and design, revision of the manuscript. DGG: study concept and design, acquisition of data, revision of the manuscript. MT-MH: study concept and design, acquisition of data, revision of the manuscript.

Funding Both the Oxford Discovery (grant reference J-1403) and Tracking Parkinson's cohorts (grant reference J-1101) were funded by Parkinson's UK

Disclaimer The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Competing interests ML: received fees for advising on a secondary analysis of an RCT sponsored by North Bristol NHS trust. MMXT: reports no disclosures. YB-S: reports no disclosures. FB: reports no disclosures. TB: reports no disclosures. JCK: reports no disclosures. SGE: reports no disclosures. SM: reports no disclosures. NM: reports no disclosures. KG: reports no disclosures. RAB: received grants from Parkinson's UK, NIHR, Cure Parkinson's Trust, Evelyn Trust, Rosetrees Trust, MRC and EU along with payment for advisory board attendance from Oxford Biomedica, Aspen Neuroscience, UCB, BlueRock Therapeutics, Novo Nordisk and LCT, and honoraria from Wiley and Springer. NW: reports no disclosures. DJB: received grants from NIHR, Wellcome Trust, GlaxoSmithKline, Parkinson's UK, and Michael J Fox Foundation. TF: grants from National Institute of Health Research, Michael J Fox Foundation, John Black Charitable Foundation, Cure Parkinson's Trust, Innovate UK, Janet Owens Research Fellowship, Van Andel Research Institute and Defeat MSA. Advisory Boards for Peptron, Voyager Therapeutics, Handl therapeutics, Living Cell Technologies, Bial grants from Welsh Assembly Government, personal fees from Teva, personal fees from Abbvie, personal fees from Teva, personal fees from UCB, personal fees from Boehringer-Ingelheim, personal fees from GSK, non-financial support from Teva, grants from Ipsen Fund, non-financial support from Medtronic, grants from MNDA, grants from PSP Association, grants from CBD Solutions, grants from Drake Foundation, personal fees from Acorda, outside the submitted work; In addition, HRM has a patent. HRM is a coapplicant on a patent application related to C9ORF72 - Method for diagnosing a neurodegenerative disease (PCT/GB2012/052140) pending. NW: reports Funding from Aligning science against Parkinsons (ASAP). He has also received consultancy fees from GSK. DGG: received payment for advisory board attendance from Merz Pharma, Vectura plc, and consultancy fees from the GM clinic. Grant support from Parkinson's UK, the Neurosciences Foundation, and Michael's Movers. MT-MH: received payment for Advisory Board attendance/consultancy for Biogen, Roche, CuraSen Therapeutics, Evidera, Manus Neurodynamica. She received funding/grant support from Parkinson's UK, Oxford NIHR BRC, University of Oxford, CPT, Lab10X, NIHR, Michael J Fox Foundation, H2020 European Union, GE Healthcare and the PSP Association. MT-MH is a coapplicant on a patent application related to smartphone predictions in Parkinson's (PCT/GB2019/052522) pending.

Patient consent for publication Consent obtained directly from patient(s)

Ethics approval This study involves human participants and was approved by the Oxford Discovery cohort was approved by NRES Committee, South Central Oxford A Research Ethics Committee, Reference number 16/SC/0108 The Tracking Parkinsons cohort was approved by West of Scotland Research Ethics Service (WoSRES) reference 11/AL/0163. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Data from the Oxford Discovery cohort is available on request from https://www.dpag.ox.ac.uk/opdc/research/external-collaborations. Data from the Tracking Parkinsons cohort is available on request from https://www.trackingparkinsons.org.uk/about-1/data/.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs

Michael Lawton http://orcid.org/0000-0002-3419-0354
Manuela MX Tan http://orcid.org/0000-0001-5835-669X
Johannes C Klein http://orcid.org/0000-0002-8553-2801
Thomas Foltynie http://orcid.org/0000-0003-0752-1813
Huw R Morris http://orcid.org/0000-0002-5473-3774
Michele Tao-Ming Hu http://orcid.org/0000-0001-6382-5841

REFERENCES

1 Graham JM, Sagar HJ. A data-driven approach to the study of heterogeneity in idiopathic Parkinson's disease: identification of three distinct subtypes. *Mov. Disord.* 1999;14:10–20.
2 Mestre TA, Fereshtehnejad S-M, Berg D, *et al*. Parkinson's Disease Subtypes: Critical Appraisal and Recommendations. *J Parkinsons Dis* 2021;11:395–404.
3 Thenganatt MA, Jankovic J. Parkinson disease subtypes. *JAMA Neurol* 2014;71:499–504.
4 van Rooden SM, Heiser WJ, Kok JN, *et al*. The identification of Parkinson's disease subtypes using cluster analysis: A systematic review. *Mov. Disord.* 2010;25:969–78.
5 Jankovic J, McDermott M, Carter J. Variable expression of Parkinson's disease: a base-line analysis of the DATATOP cohort. *The Parkinson Study Group, Neurology* 1990;40:1529–34.
6 Eisinger RS, Martinez-Ramirez D, Ramirez-Zamora A. Parkinson's disease motor subtype changes during 20 years of follow-up, Parkinsonism Relat. *Disord* 2020;76:104–7.
7 Niemann N, Jankovic J. Juvenile parkinsonism: differential diagnosis, genetics, and treatment, parkinsonism relat. *Disord* 2019;67:74–89.
8 Mehanna R, Jankovic J. Young-onset Parkinson's disease: Its unique features and their impact on quality of life, Parkinsonism Relat. *Disord* 2019;65:39–48.
9 Lawton M, Baig F, Rolinski M, *et al*. Parkinson's Disease Subtypes in the Oxford Parkinson Disease Centre (OPDC) Discovery Cohort. *J Parkinsons Dis* 2015;5:269–79.
10 Lawton M, Ben-Shlomo Y, May MT, *et al*. Developing and validating Parkinson's disease subtypes and their motor and cognitive progression. *J Neurol Neurosurg Psychiatry* 2018;89:1279–87.
11 Lawton M, Baig F, Toulson G, *et al*. Blood Biomarkers With Parkinson's Disease Clusters and Prognosis: The Oxford Discovery Cohort. *Mov Disord* 2020;35:279–87.
12 Alfradique-Dunham I, Al-Ouran R, von Coelln R, *et al*. Genome-Wide association study meta-analysis for Parkinson disease motor subtypes. *Neurol Genet* 2021;7:e557.
13 Tan MMX, Lawton MA, Jabbari E, *et al*. Genome-wide association studies of cognitive and motor progression in Parkinson's Disease. *Mov Disord* 2021;36:424–33.
14 Malek N, Swallow DMA, Grosset KA, *et al*. Tracking Parkinson's: Study Design and Baseline Patient Data. *J Parkinsons Dis* 2015;5:947–59.
15 Szewczyk-Krolikowski K, Tomlinson P, Nithi K. 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 Relat. *Disord* 2013;20:99–105.

16 Illumina. HumanCoreExome 12 v1.1 support files, 2021. Available: https://emea.support.illumina.com/downloads/humancoreexome-12v1-1_product_support_files.html [Accessed 01/10/2021].

17 Illumina. Infinium CoreExome 24 v1.1 support files, 2021. Available: https://emea.support.illumina.com/downloads/infinium-coreexome-24-v1-1-support-files.html [Accessed 01/10/2021].

18 Chen JA, Chen Z, Won H, *et al*. Joint genome-wide association study of progressive supranuclear palsy identifies novel susceptibility loci and genetic correlation to neurodegenerative diseases. *Mol Neurodegener* 2018;13:41.

19 Nalls MA, Blauwendraat C, Vallerga CL, *et al*. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol* 2019;18:1091–102.

20 Sailer A, Scholz SW, Nalls MA, *et al*. A genome-wide association study in multiple system atrophy. *Neurology* 2016;87:1591–8.

21 Chia R, Sabir MS, Bandres-Ciga S. Genome sequencing analysis identifies new loci associated with Lewy body dementia and provides insights into the complex genetic architecture, (2020).

22 Kunkle BW, Grenier-Boley B, Sims R. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing, Nat. *Genet* 2019;51:414–30.

23 Malek N, Weil RS, Bresner C, *et al*. Features of *GBA* -associated Parkinson's disease at presentation in the UK *Tracking Parkinson's* study. *J Neurol Neurosurg Psychiatry* 2018;89:702–9.

24 Tan MMX, Malek N, Lawton MA, *et al*. Genetic analysis of Mendelian mutations in a large UK population-based Parkinson's disease study. *Brain* 2019;142:2828–44.

25 Barber TR, Lawton M, Rolinski M, *et al*. Prodromal parkinsonism and neurodegenerative risk stratification in REM sleep behavior disorder. *Sleep* 2017;40.

26 Blauwendraat C, Faghri F, Pihlstrom L. NeuroChip, an updated version of the NeuroX genotyping platform to rapidly screen for variants associated with neurological diseases. *Neurobiol Aging* 2017;57:247.

27 Igo RP, Kinzy TG, Cooke Bailey JN. Genetic risk scores. *Curr Protoc Hum Genet* 2019;104

28 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B* 1995;57:289–300.

29 Perneger TV. What's wrong with Bonferroni adjustments. *BMJ* 1998;316:1236–8.

30 Sterne JAC, Smith GD. Sifting the evidence - what's wrong with significance tests? *Bmj-Brit Med J* 2001;322:226-+.

31 Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process and Purpose. *Am Stat* 2016;70:129–31.

32 Visscher PM, Wray NR, Zhang Q, *et al*. 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics* 2017;101:5–22.

33 Liu G, Boot B, Locascio JJ. Specifically neuropathic Gaucher's mutations accelerate cognitive decline in Parkinson's. *Ann Neurol* 2016;80:674–85.

34 Winder-Rhodes SE, Evans JR, Ban M, *et al*. Glucocerebrosidase mutations influence the natural history of Parkinson's disease in a community-based incident cohort. *Brain* 2013;136:392–9.

35 Alcalay RN, Caccappolo E, Mejia-Santana H, *et al*. Cognitive performance of GBA mutation carriers with early-onset PD: the CORE-PD study. *Neurology* 2012;78:1434–40.

36 Setó-Salvia N, Pagonabarraga J, Houlden H, *et al*. Glucocerebrosidase mutations confer a greater risk of dementia during Parkinson's disease course. *Mov. Disord.* 2012;27:393–9.

37 Davis MY, Johnson CO, Leverenz JB, *et al*. Association of GBA mutations and the E326K polymorphism with motor and cognitive progression in Parkinson disease. *JAMA Neurol* 2016;73:1217–24.

38 Menozzi E, Schapira AHV. Exploring the Genotype–Phenotype correlation in GBA-Parkinson disease: clinical aspects, biomarkers, and potential modifiers. *Front Neurol* 2021;12:694764.

39 Somme JH, Molano Salazar A, Gonzalez A. Cognitive and behavioral symptoms in Parkinson's disease patients with the G2019S and R1441G mutations of the LRRK2 gene, Parkinsonism Relat. *Disord* 2015;21:494–9.

40 Alcalay RN, Mirelman A, Saunders-Pullman R, *et al*. Parkinson disease phenotype in Ashkenazi jews with and without *LRRK2* G2019S mutations. *Mov Disord.* 2013;28:1966–71.

41 Marras C, Alcalay RN, Caspell-Garcia C, *et al*. Motor and nonmotor heterogeneity of *LRRK2* -related and idiopathic Parkinson's disease. *Mov Disord.* 2016;31:1192–202.

42 Saunders-Pullman R, Mirelman A, Alcalay RN, *et al*. Progression in the LRRK2-Asssociated Parkinson disease population. *JAMA Neurol* 2018;75:312–9.

43 Yahalom G, Orlev Y, Cohen OS, *et al*. Motor progression of Parkinson's disease with the leucine-rich repeat kinase 2 G2019S mutation. *Mov Disord.* 2014;29:1057–60.

44 Poggiolini I, Gupta V, Lawton M. Diagnostic value of cerebrospinal fluid alpha-synuclein seed quantification in synucleinopathies. *Brain* 2021.

Web-appendix

Genetics of validated Parkinson's Disease subtypes in the Oxford Discovery and Tracking Parkinson's cohorts

Illumina Array quality control and Imputation

The same standard quality control procedures on the genotype data were carried out separately within the two cohorts. Individuals were excluded for the following reasons: related individuals as identified by an Identity-By-Descent PIHAT > 0.1; individuals with low overall genotyping rates (<98%); heterozygosity outliers, where data was >2 standard deviations away from the mean; and individuals whose reported sex was not the same as the genetically determined sex. A principle components analysis, after merging with European (CEU) samples from the HapMap reference panel, was carried out and individuals who were >6 standard deviations away from the mean of any of the first 10 principle components (PC) were also excluded. Genetic variants were excluded if they had a low genotyping rate (<99%), minor allele frequency <1% or a Hardy-Weinberg equilibrium p-value < 5x 10^{-6}.

After these quality control steps the genotypes were imputed, separately by cohort, to the 1,000 Genomes Project reference panel (phase 3 release 5)[1] using the Michigan Imputation Server (https://imputationserver.sph.umich.edu). Finally genetic variants were excluded if the imputation quality scores (R2) were < 0.8 and then imputation dosages were converted into hard call genotypes. After the imputation and quality control there were 9,153,714 SNPs in the Oxford Discovery cohort and 8,986,152 SNPs in the Tracking cohort. When we carried out the pruning to generate PC's in the two cohorts there were 25,647 SNPs in the Oxford Discovery cohort and 23,876 SNPs in the Tracking cohort.

Neurochip quality control

Individuals were excluded when: low overall genotyping rates (<97%); heterozygosity outliers, data was >3 standard deviations away from the mean; and individuals whose reported sex was not the same as the genetically determined sex.

Number of SNPs in GWAS study

After excluding variants with a minor allele frequency (MAF) < 0.05 there were 6,625,590 variants remaining in the Oxford Discovery cohort of which another 76,381 were excluded due to being palindromic with a MAF > 0.45. So the final Oxford Discovery cohort GWAS included 6,549,209 variants.

After excluding variants with a minor allele frequency (MAF) < 0.05 there were 6,672,212 variants remaining in the Tracking cohort. Of these 34 were duplicates (with identical Chromosome number and base position) and 76,521 were palindromic. The final Tracking cohort GWAS included 6,595,657 variants.

In the final meta-analysis there were 6,413,412 variants.

Potential biological relevance of GWAS variants

The two SNPs from chromosome 6 (rs316037 and rs5881357) that were associated with belonging to cluster 1 are both significant expression quantitative trait loci (eQTL) for the gene *LPAL2* and also are significant splicing quantitative trait loci (sQTL) for the gene *SLC22A3* according to GTEx Portal on 22/09/2021 (dbGaP Accession phs000424.v8.p2). *SLC22A3* has been shown in humans to be related to choline metabolism in cancer (KEGG T01001: 6581 (genome.jp)) and *LPAL2* is related to lipoprotein (a) in humans (LPAL2 lipoprotein(a) like 2, pseudogene [Homo sapiens (human)] - Gene - NCBI (nih.gov)). The SNP rs5881357 is in the same genomic locus as SNPs that have been reported to be related to BMI [2], various blood biomarkers such as apolipoprotein B levels and lipoprotein (a) levels [3] and mean spheric corpuscular volume [4]. There is also some evidence that the *SLC22A3-LPAL2-LPA* gene cluster contributes to risk and severity of coronary artery disease [5, 6].

The SNP from chromosome 1 (rs116258323) that was associated with belonging to cluster 3 is a significant cis-eQTL for the gene *SMYD2* [7]. This gene has been shown in humans to be involved in lysine degradation and metabolic pathways (KEGG T01001: 56950 (genome.jp)) and is a promising candidate for the treatment of cardiovascular disease and cancer [8].

Looking at the 1000 genomes project (European population) the minor allele frequencies (MAF) for these SNPS were 0.207 (rs316037), 0.217 (rs5881357) and 0.068 (rs116258323). The MAF was not reported for rs151043031. These values are very similar to the MAFs in our data, see supplementary table 2, which might suggest that although these SNPs are not associated with Parkinson's as a whole they might be associated with specific subtypes.

Network analyses

Functional mapping and annotation of GWAS results were carried out using the freely available internet platform FUMA (https://fuma.ctglab.nl/) using standard settings [9]. Within FUMA, summary statistics from our four GWAS results were analysed using Multi-marker Analysis of GenoMic Annotation (MAGMA) gene property tests to compare enrichment of the average gene expression within different tissues. The resulting tissue expression enrichment for each cluster are within supplementary figures 7 to 10, the upper graph shows a tissue expression analysis on 30 general tissue types and the lower graph 53 specific tissue types. After Bonferroni correction there was no evidence of association between our GWAS results for any cluster and average gene expression per tissue.

REFERENCES

[1] C. Genomes Project, A. Auton, L.D. Brooks, et al., A global reference for human genetic variation, Nature 526(7571) (2015) 68-74.
[2] A.E. Locke, B. Kahali, S.I. Berndt, et al., Genetic studies of body mass index yield new insights for obesity biology, Nature 518(7538) (2015) 197-206.
[3] N. Sinnott-Armstrong, Y. Tanigawa, D. Amar, et al., Genetics of 35 blood and urine biomarkers in the UK Biobank, Nat. Genet. 53(2) (2021) 185-194.
[4] D. Vuckovic, E.L. Bao, P. Akbari, et al., The Polygenic and Monogenic Basis of Blood Traits and Diseases, Cell 182(5) (2020) 1214-1231 e11.
[5] L. Wang, J. Chen, Y. Zeng, et al., Functional Variant in the SLC22A3-LPAL2-LPA Gene Cluster Contributes to the Severity of Coronary Artery Disease, Arterioscler. Thromb. Vasc. Biol. 36(9) (2016) 1989-96.
[6] D.A. Tregouet, I.R. Konig, J. Erdmann, et al., Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease, Nat. Genet. 41(3) (2009) 283-5.
[7] U. Vosa, A. Claringbould, H.J. Westra, et al., Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression, Nat. Genet. 53(9) (2021) 1300-1310.
[8] X. Yi, X.J. Jiang, Z.M. Fang, Histone methyltransferase SMYD2: ubiquitous regulator of disease, Clin. Epigenetics 11(1) (2019) 112.
[9] K. Watanabe, E. Taskesen, A. van Bochoven, et al., Functional mapping and annotation of genetic associations with FUMA, Nat Commun 8(1) (2017) 1826.

Supplementary table 1. Number of SNPs included in each genetic risk score after excluding SNPs in quality control/imputation and those that were palindromic with a minor allele frequency > 0.45.

Analysis (total SNPs reported)	**Tracking cohort**	**Discovery cohort**
PD (90)	86	85
PSP (5)	5	5
MSA (23)	21	21
LBD (5)	5	5
AD (23)	22	21

PD = Parkinson’s Disease; PSP = progressive supranuclear palsy; MSA = multiple system atrophy; AD = Alzheimer’s disease; LBD = Lewy body dementia (LBD)

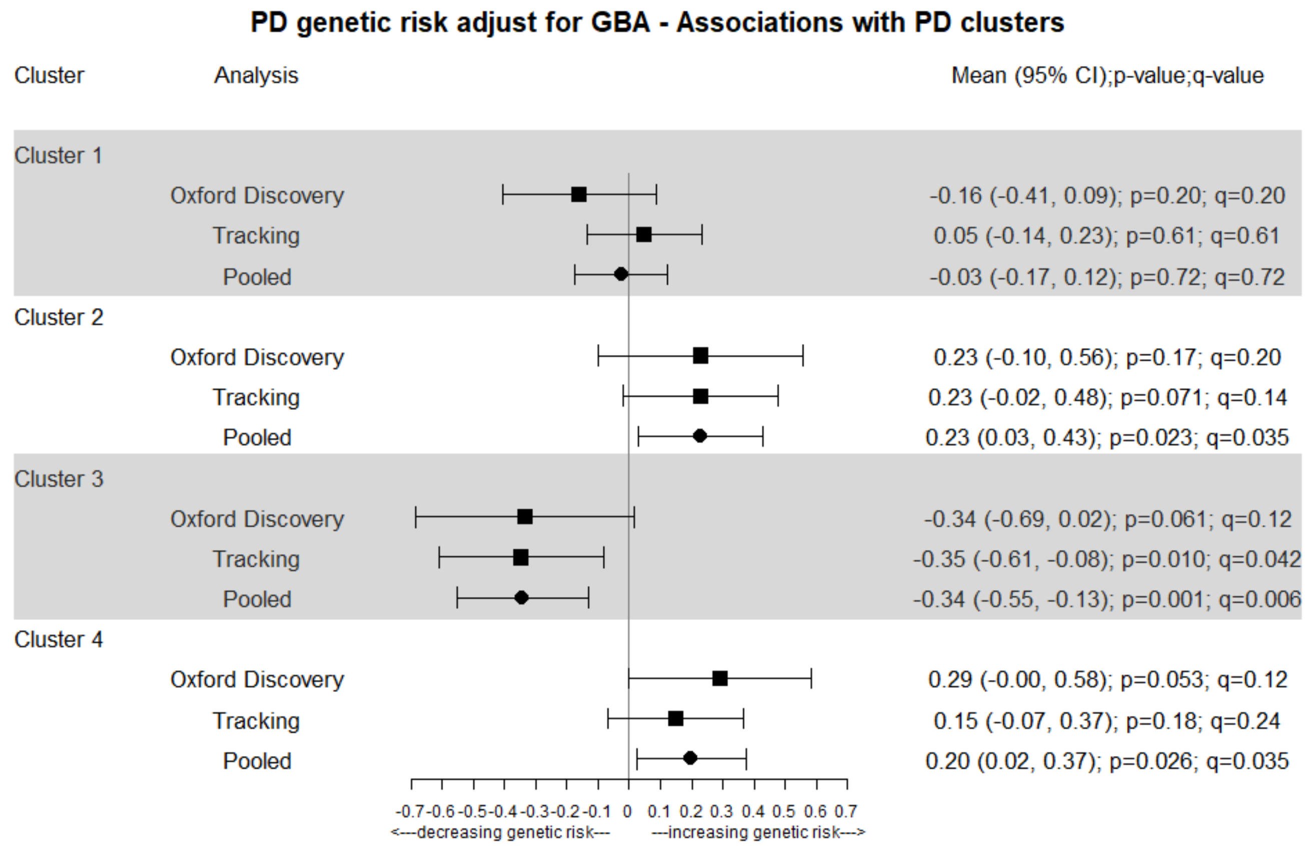
Supplementary table 2. Power calculations for a range of beta's and minor allele frequencies (MAF) for a sample size of 2274 (combined Discovery and Tracking) and p-value < 5x10e-8 . The beta's are in s.d. units for the outcome and the s.d of the outcome was 3.4, 4.6, 4.9 and 4.1 for clusters 1 to 4 respectively. For context the highest beta in s.d. units for cluster 1 was ~0.17 with a MAF of ~0.23 and the highest beta in s.d units for cluster 3 was ~0.33 with a MAF of ~0.05.

Beta\MAF	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.1	0.0%	0.0%	0.1%	0.3%	0.6%	0.9%	1.3%	1.6%	1.8%	1.9%
0.15	0.1%	0.8%	3.4%	8.2%	14.6%	21.4%	27.4%	32.0%	34.8%	35.8%
0.2	0.6%	8.2%	27.1%	49.1%	66.8%	78.3%	85.1%	88.9%	90.8%	91.3%
0.25	3.9%	35.8%	73.2%	91.3%	97.4%	99.1%	99.7%	99.8%	99.9%	99.9%
0.3	15.3%	74.8%	96.8%	99.7%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
0.35	39.1%	95.6%	99.9%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
0.4	68.2%	99.7%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

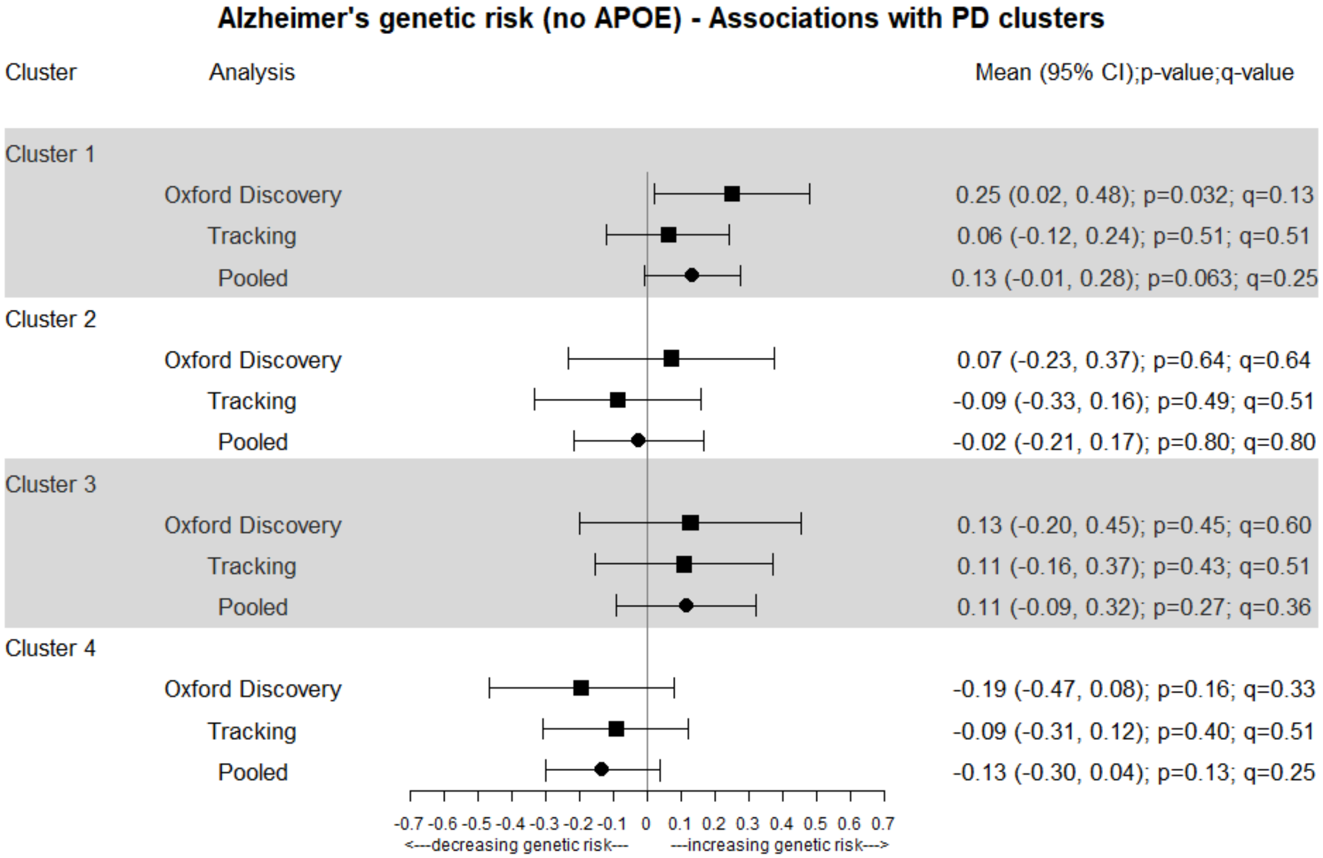
Supplementary Table 3. SNPs meeting a threshold of 1x10e-6 from the genome wide association study meta-analysis – cohort specific results.

CLUSTER 1 - TRACKING COHORT									
Chr	**Position (GRCh37)**	**Marker**	**A1**	**A2**	**N**	**A1 frequency**	**Beta**	**SE**	**p -value**
1	237734615	rs151043031	CT	C	1467	0.231	0.70	0.15	6.30e-06
6	160698177	rs316037	G	A	1467	0.217	0.54	0.16	0.000760
6	160699605	rs5881357	AT	A	1467	0.221	0.52	0.16	0.000875
CLUSTER 1 - DISCOVERY COHORT									
Chr	**Position (GRCh37)**	**Marker**	**A1**	**A2**	**N**	**A1 frequency**	**Beta**	**SE**	**p -value**
1	237734615	rs151043031	CT	C	807	0.230	0.42	0.19	0.0308
6	160698177	rs316037	G	A	807	0.214	0.70	0.19	0.000333
6	160699605	rs5881357	AT	A	807	0.224	0.72	0.19	0.000168
CLUSTER 3 - TRACKING COHORT									
Chr	**Position (GRCh37)**	**Marker**	**A1**	**A2**	**N**	**A1 frequency**	**Beta**	**SE**	**p -value**
1	214449747	rs116258323	T	C	1467	0.050	1.40	0.43	.00120
CLUSTER 3 - DISCOVERY COHORT									
Chr	**Position (GRCh37)**	**Marker**	**A1**	**A2**	**N**	**A1 frequency**	**Beta**	**SE**	**p -value**
1	214449747	rs116258323	T	C	807	0.058	1.92	0.50	.000129

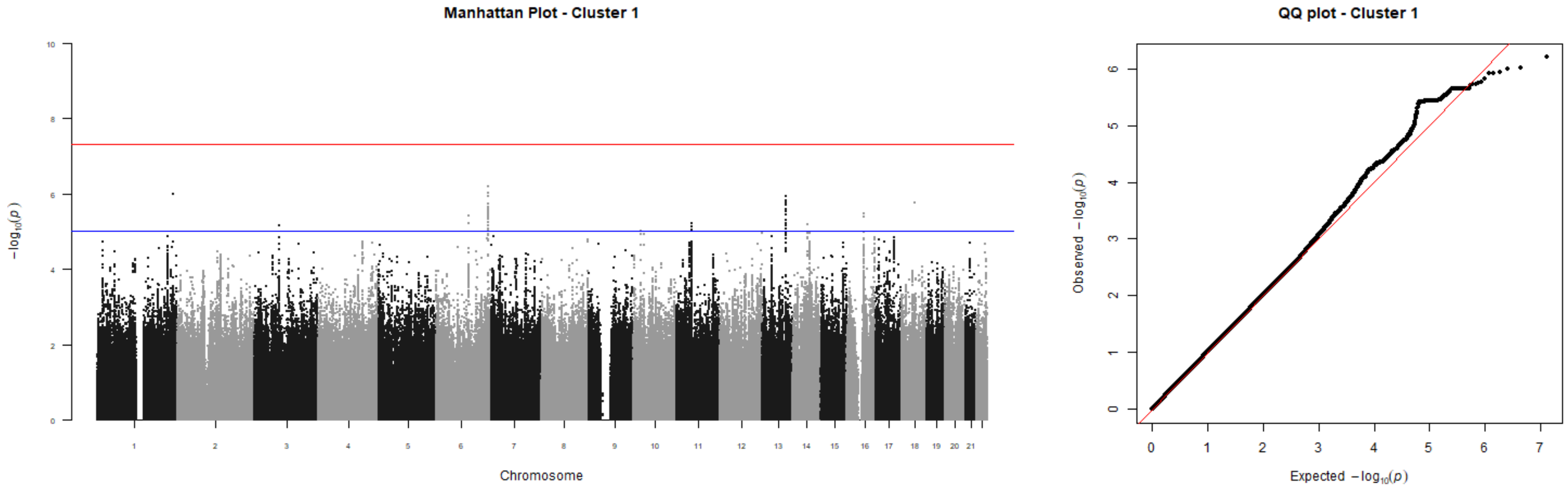
Supplementary figure 1. Genetic risk of Parkinson's when adjusting for GBA mutation carriers – sensitivity analysis



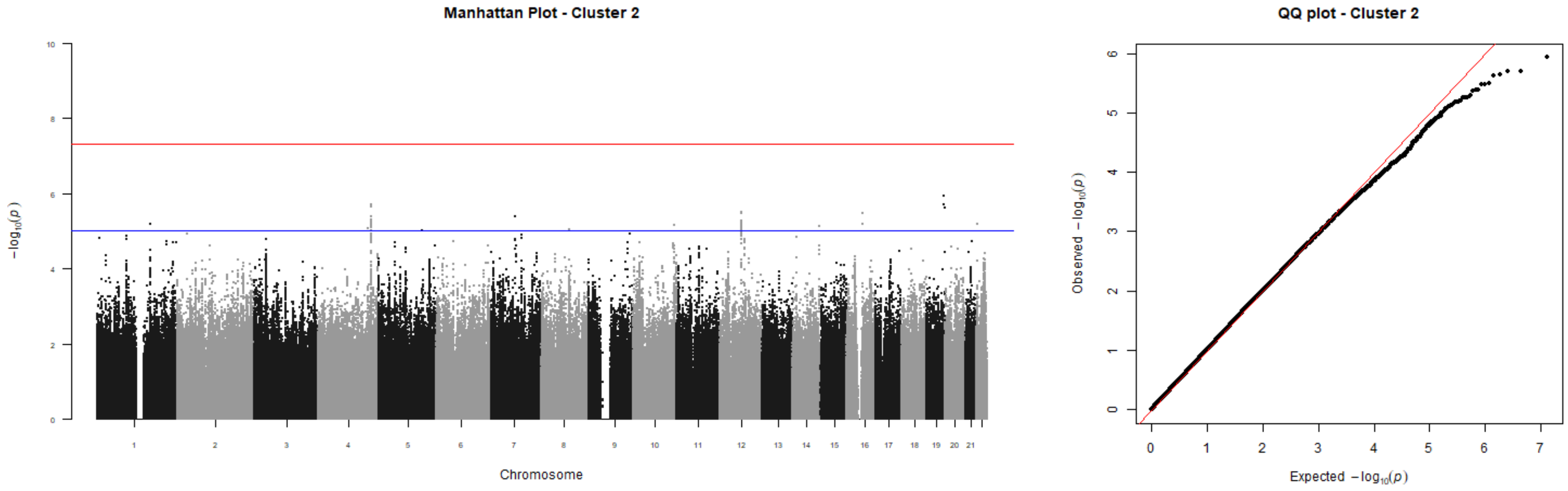
Supplementary figure 2. Genetic risk of Alzheimer's when removing the APOE genetic variant – sensitivity analysis



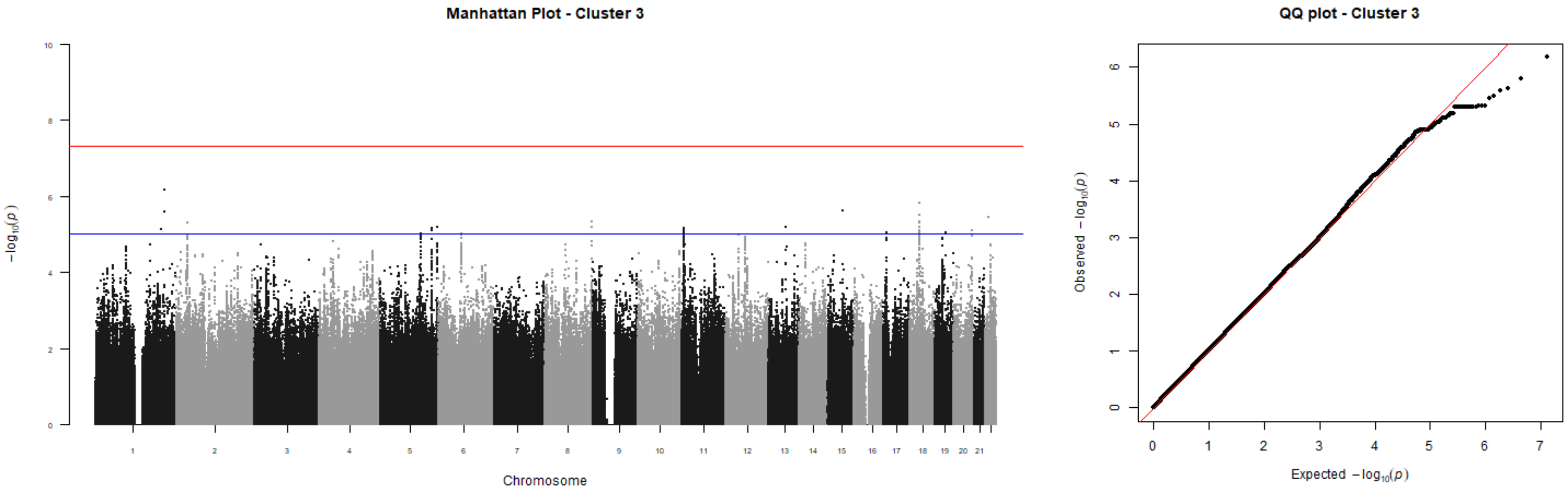
Supplementary figure 3. Manhattan Plot and QQ-plot for cluster 1 genome wide association study



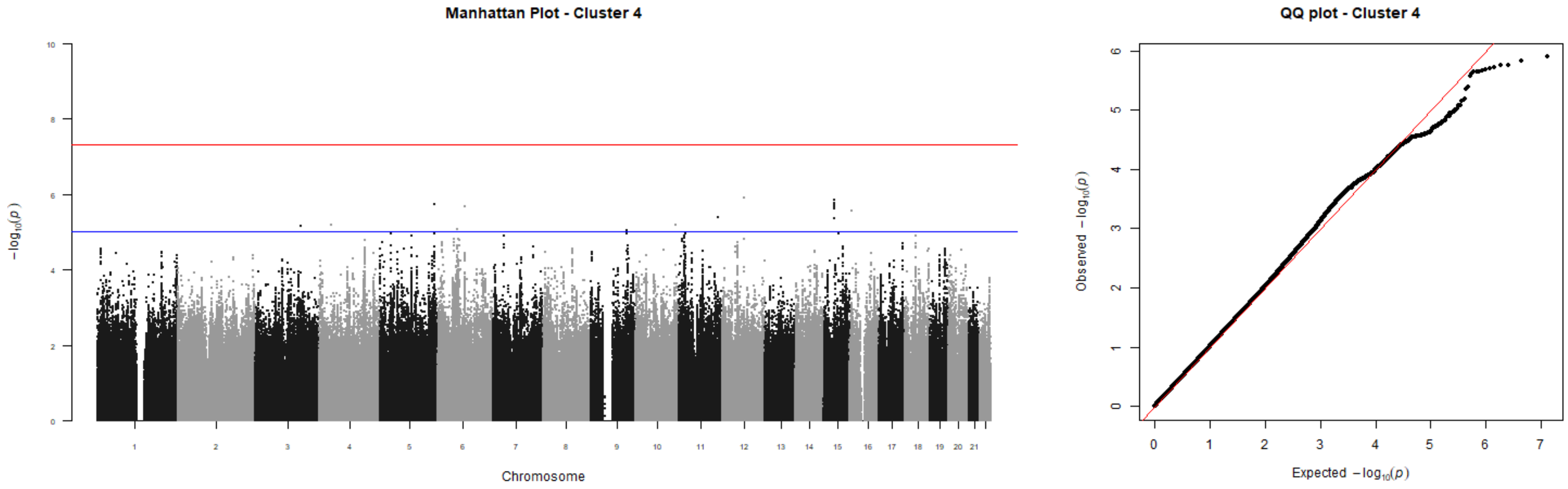
Supplementary figure 4. Manhattan Plot and QQ-plot for cluster 2 genome wide association study



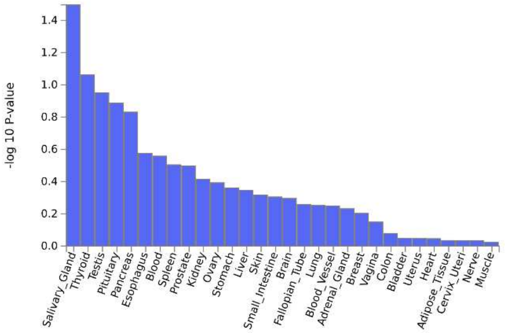
Supplementary figure 5. Manhattan Plot and QQ-plot for cluster 3 genome wide association study

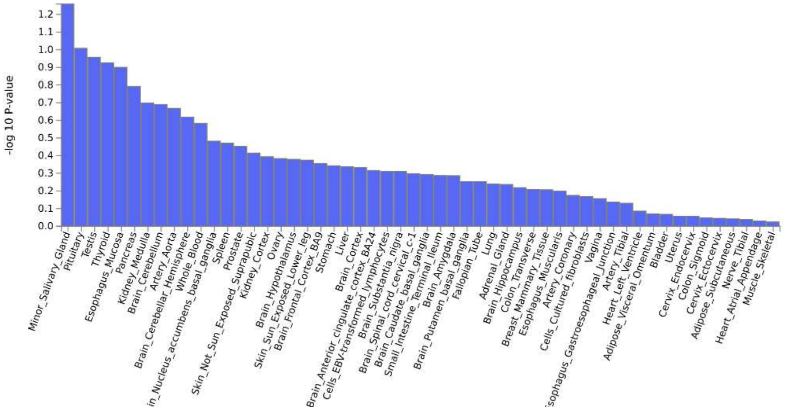


Supplementary figure 6. Manhattan Plot and QQ-plot for cluster 4 genome wide association study

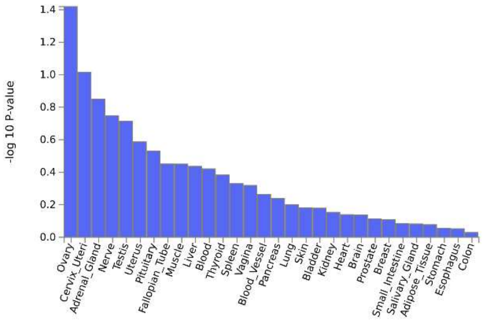


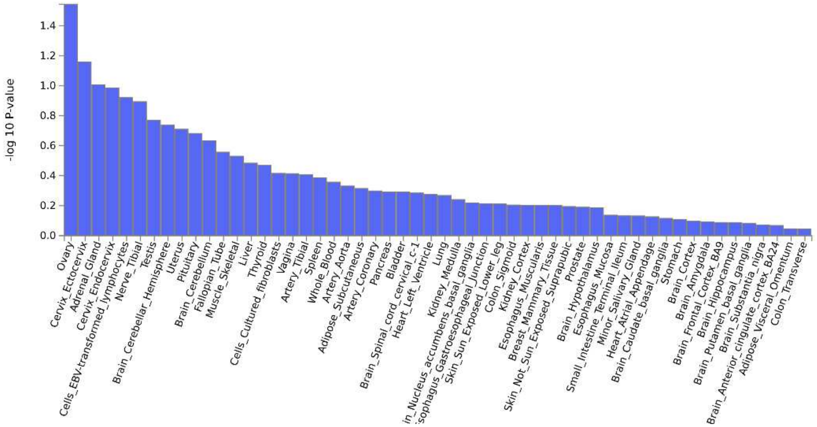
Supplementary figure 7. Results of FUMA analysis for MAGMA tissue expression analysis for cluster 1. Red bars would indicate a significance threshold of <0.05 for a Bonferroni adjusted p-value.



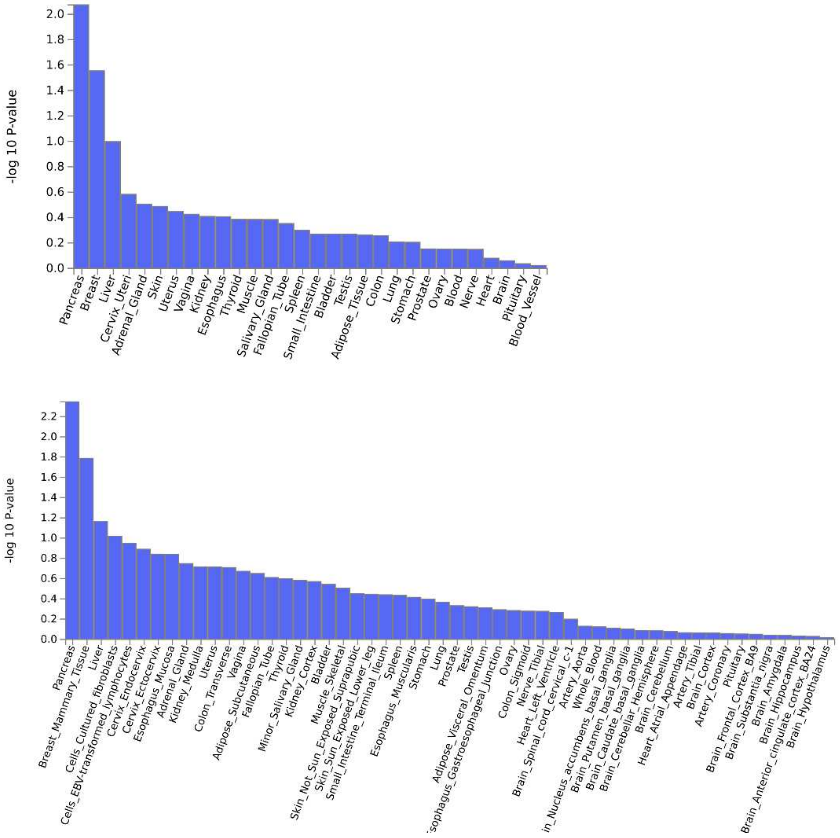


Supplementary figure 8. Results of FUMA analysis for MAGMA tissue expression analysis for cluster 2. Red bars would indicate a significance threshold of <0.05 for a Bonferroni adjusted p-value.

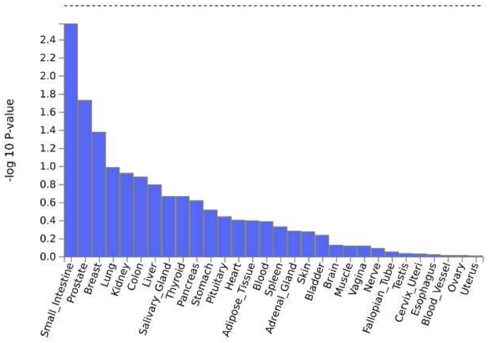


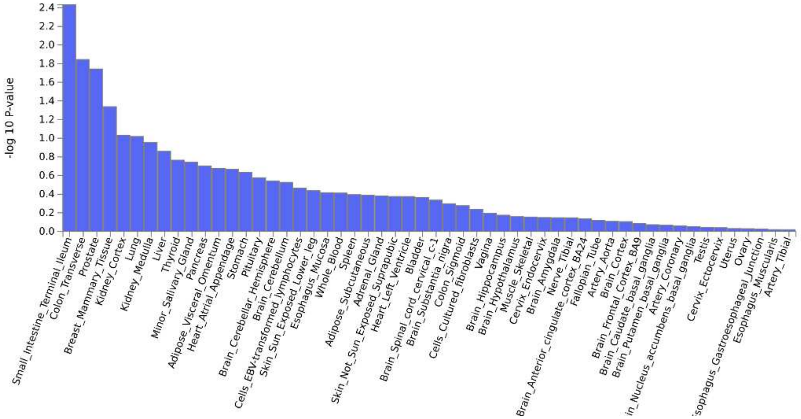


Supplementary figure 9. Results of FUMA analysis for MAGMA tissue expression analysis for cluster 3. Red bars would indicate a significance threshold of <0.05 for a Bonferroni adjusted p-value.



Supplementary figure 10. Results of FUMA analysis for MAGMA tissue expression analysis for cluster 4. Red bars would indicate a significance threshold of <0.05 for a Bonferroni adjusted p-value.


-log 10 P-value
Small_Intestine
Prostate
Breast
Lung
Kidney
Colon
Liver
Salivary_Gland
Thyroid
Pancreas
Stomach
Pituitary
Heart
Adipose_Tissue
Blood
Spleen
Adrenal_Gland
Skin
Bladder
Brain
Muscle
Vagina
Nerve
Fallopian_Tube
Testis
Cervix_Uteri
Esophagus
Blood_Vessel
Ovary
Uterus


-log 10 P-value
Small_Intestine_Terminal_Ileum
Colon_Transverse
Prostate
Breast_Mammary_Tissue
Kidney_Cortex
Lung
Kidney_Medulla
Liver
Thyroid
Minor_Salivary_Gland
Pancreas
Adipose_Visceral_Omentum
Heart_Atrial_Appendage
Stomach
Pituitary
Brain_Cerebellar_Hemisphere
Brain_Cerebellum
Cells_EBV-transformed_lymphocytes
Skin_Sun_Exposed_Lower_leg
Esophagus_Mucosa
Whole_Blood
Spleen
Adipose_Subcutaneous
Adrenal_Gland
Skin_Not_Sun_Exposed_Suprapubic
Heart_Left_Ventricle
Bladder
Brain_Spinal_cord_cervical_c-1
Brain_Substantia_nigra
Colon_Sigmoid
Cells_Cultured_fibroblasts
Vagina
Brain_Hippocampus
Brain_Hypothalamus
Muscle_Skeletal
Cervix_Endocervix
Brain_Amygdala
Nerve_Tibial
Brain_Anterior_cingulate_cortex_BA24
Fallopian_Tube
Artery_Aorta
Brain_Cortex
Brain_Frontal_Cortex_BA9
Brain_Caudate_basal_ganglia
Brain_Putamen_basal_ganglia
Artery_Coronary
in_Nucleus_accumbens_basal_ganglia
Testis
Cervix_Ectocervix
Uterus
Ovary
sophagus_Gastroesophageal_Junction
Esophagus_Muscularis
Artery_Tibial