ARTICLE OPEN

Check for updates

Genetic meta-analysis of levodopa induced dyskinesia in Parkinson's disease

Alejandro Martinez-Carrasco[1,2,3]✉, Raquel Real[1,2,3], Michael Lawton[4], Hirotaka Iwaki[5,6], Manuela M. X. Tan[7], Lesley Wu[1,2,3], Nigel M. Williams[8], Camille Carroll[9,10], Michele T. M. Hu[11,12], Donald G. Grosset[13], John Hardy[3,14,15,16,17,18], Mina Ryten[3,19,20], Tom Foltynie[1], Yoav Ben-Shlomo[4], Maryam Shoai[3,14,15] and Huw R. Morris[1,2,3]✉

The genetic basis of levodopa-induced-dyskinesia (LiD) is poorly understood, and there have been few well-powered genome-wide studies. We performed a genome-wide survival meta-analyses to study the effect of genetic variation on the development of LiD in five separate longitudinal cohorts, and meta-analysed the results. We included 2784 PD patients, of whom 14.6% developed LiD. We found female sex (HR = 1.35, SE = 0.11, $P = 0.007$) and younger age at onset (HR = 1.8, SE = 0.14, $P = 2 \times 10^{-5}$) increased the probability of developing LiD. We identified three genetic loci significantly associated with time-to-LiD onset. **rs72673189** on chromosome 1 (HR = 2.77, SE = 0.18, $P = 1.53 \times 10^{-8}$) located at the LRP8 locus, **rs189093213 on** chromosome 4 (HR = 3.06, SE = 0.19, $P = 2.81 \times 10^{-9}$) in the non-coding RNA *LINC02353* locus, and **rs180924818** on chromosome 16 (HR = 3.13, SE = 0.20, $P = 6.27 \times 10^{-9}$) in the *XYLT1* locus. Based on a functional annotation analysis on chromosome 1, we determined that changes in DNAJB4 gene expression, close to LRP8, are an additional potential cause of increased susceptibility to LiD. Baseline anxiety status was significantly associated with LiD (OR = 1.14, SE = 0.03, $P = 7.4 \times 10^{-5}$). Finally, we performed a candidate variant analysis of previously reported loci, and found that genetic variability in *ANKK1* (*rs1800497*, HR = 1.27, SE = 0.09, $P = 8.89 \times 10^{-3}$) and *BDNF* (*rs6265*, HR = 1.21, SE = 0.10, $P = 4.95 \times 10^{-2}$) loci were significantly associated with time to LiD in our large meta-analysis.

npj Parkinson's Disease (2023)9:128 ; https://doi.org/10.1038/s41531-023-00573-2

INTRODUCTION

Parkinson's disease (PD) is a common neurodegenerative disorder, characterised by the loss of dopaminergic neurons in the substantia nigra pars compacta. The development of levodopa induced dyskinesia (LiD) is a major clinical problem for PD patients and multiple pharmacological and neurosurgical approaches have been developed to try to prevent, attenuate or treat LiD. Dopamine is lost from the nigrostriatal pathway, which manifests as bradykinesia, muscular rigidity, rest tremor and postural instability[1,2]. There are several symptomatic treatments for PD motor symptoms, with the metabolic precursor of dopamine, levodopa, being the 'gold standard' drug. Levodopa improves motor function as measured by the Unified Parkinson's Disease Rating Scale (UPDRS) or the more recent MDS-UPDRS, widely used standard clinical assessments to evaluate the motor state in PD patients[3]. A comparison of an early levodopa treated group against a delayed treated group showed no difference in the rate of motor progression, suggesting that levodopa itself is not disease modifying or disease accelerating[4]. One of the major drawbacks of long-term levodopa treatment is that many PD patients experience levodopa-related motor complications, such as wearing off, dystonia and dyskinesia[5].

The prevalence of LiD varies across academic- and industry-led studies, averaging at around 20–40% after 4 years of levodopa treatment. There are two major LiD subtypes: peak-dose dyskinesia, which occur during the therapeutic window of levodopa treatment, and diphasic dyskinesia, which present at the start and end of a dose cycle[6].

Levodopa treatment is necessary for LiD development, but there are likely to be several other mediating factors[6]. Based on research in animal models, it is hypothesised that pulsatile delivery of oral levodopa, presynaptic nigrostriatal degeneration and intact striatal neurons are needed for the development of LiD[6]. Major risk factors for the development of LiD include young age at onset (AAO), female sex, low body weight, disease severity, disease duration and treatment duration (from the initiation of levodopa) as well as the total dose of levodopa[7,8]. Disease duration and treatment duration are closely related and delayed start study designs have evaluated the effect of delaying the initiation of levodopa, showing an association between longer delay and a decreased risk of LiD[9]. There is increasing evidence to

[1]Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, University College London, London, UK. [2]UCL Movement Disorders Centre, University College London, London, UK. [3]Aligning Science Across Parkinson's (ASAP) Collaborative Research Network, Chevy Chase, MD 20815, USA. [4]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK. [5]Center for Alzheimer's and Related Dementias (CARD), National Institute on Aging and National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA. [6]Data Tecnica International, Glen Echo, MD, USA. [7]Department of Neurology, Oslo University Hospital, Oslo, Norway. [8]Institute of Psychological Medicine and Clinical Neurosciences, MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff, UK. [9]Faculty of Health, University of Plymouth, Plymouth, UK. [10]Translational and Clinical Research Institute, Newcastle University, Newcastle, UK. [11]Nuffield Department of Clinical Neurosciences, Division of Clinical Neurology, University of Oxford, Oxford, UK. [12]Oxford Parkinson's Disease Centre, University of Oxford, Oxford, UK. [13]School of Neuroscience and Psychology, University of Glasgow, Glasgow, UK. [14]Department of Neurodegenerative Diseases, UCL Queen Square Institute of Neurology, University College London, London, UK. [15]UK Dementia Research Institute, University College London, London, UK. [16]Reta Lila Weston Institute, UCL Queen Square Institute of Neurology, London, UK. [17]National Institute for Health Research (NIHR) University College London Hospitals Biomedical Research Centre, London, UK. [18]Institute for Advanced Study, The Hong Kong University of Science and Technology, Hong Kong SAR, China. [19]Genetics and Genomic Medicine, UCL Great Ormond Street Institute of Child Health, University College London, London, UK. [20]NIHR Great Ormond Street Hospital Biomedical Research Centre, University College London, London, UK. ✉email: alejandro.carrasco.20@ucl.ac.uk; h.morris@ucl.ac.uk

Table 1. Cohort summary statistics.

Cohort	PD patients Post-QC (n)	Follow up, years	No.(%) LiD	No.(%) left-censored	No.(%) male	Time to midpoint event (mean ± sd)	AAO, years (mean ± sd)	AAB, years (mean ± sd)	Disease duration at baseline from onset, years (mean ± sd)	MDS-UPDRS part III at baseline (mean ± sd)	Levodopa dose at baseline (mean ± sd)
Tracking Parkinson's	1478	7.5	177 (12)	16 (1)	945 (64.23)	7.47 (2.18)	64.43 (9.16)	67.29 (9)	2.86 (1.58)	22.36 (11.69)	217 (197)
OPDC	705	9.0	92 (13)	8 (0.8)	451 (64)	7.87 (2.87)	64.35 (9.47)	67.21 (9.26)	2.85 (1.70)	26.27 (10.82)	280 (205)
PPMI	283	9.0	82 (21)	0 (0)	259 (66)	8.28 (2.27)	60.16 (9.93)	62.08 (9.78)	1.92 (1.30)	21.38 (9.10)	0 (0)
PD STAT	77	2.0	10 (13)	4 (4.9)	48 (62)	8.77 (2.83)	57.23 (8.7)	64.84 (9.24)	7.61 (1.73)	28.86 (11.61)	NA
PDBP	241	5.0	33 (14)	16 (6)	149 (62)	5.93 (2.66)	NA	64.58 (9.3)	2.85 (2.51)	20.9 (11.11)	414 (207)

No. (%) of LiD. This is the percentage with respect to (*n*).
No. (%) of left-censored. This is the percentage of left-censored patients with respect to (*n*).
No.(%) male. This is the percentage of males with respect to (*n*).
MDS-UPDRS part III (mean ± sd). MDS-UPDRS part III total at baseline.

suggest that genetics plays a role in the susceptibility to LiD. Rare variants in genes such as *PRKN*, *PINK1*, and *DJ-1* have been reported to be associated with higher rates of dyskinesia[10–12], although patients with autosomal recessive PD usually have early onset disease, which is in itself a risk factor for LiD. A study which corrected for age and disease duration variability did not replicate the findings of a higher LiD susceptibility among *PARK2* mutation carriers[13].

Common variation may also influence the risk of developing LiD. Variation at the DRD2, COMT, MAOA, BDNF, SLC6A3 and ADORA2A loci have all been reported to influence the risk of developing LiD[14–23]. Recently, an exome-wide association study of LiD in PD found that variants in *MAD2L2* and *MAP7* loci were associated with LiD, and replicated the association of the opioid receptor gene *OPRM1*[24]. Due to the high heterogeneity in the genetic determinants that regulate LiD, validation in large cohorts is needed.

Here, we investigated the genetic determinants of LiD by performing a meta-analysis of genome-wide survival to LiD in five different cohorts, and assessed previously reported loci. We also performed functional genetic annotation to better understand the nominated loci. Lastly, we have investigated the predictive power of a polygenic risk score (PRS), and explored baseline clinical features that were significantly associated with the development of LiD in PD using a stepwise regression approach.

RESULTS

Cohort clinical features and prevalence

Across all cohorts ($n = 2784$ PD patients), the incidence of LiD was 14% (Table 1), except in the PPMI cohort where it was 21%. This is consistent with the effect of age at onset on LiD[25–27], given that PPMI is a de novo study that recruited younger patients. We did not exclude any patient from the PPMI cohort due to left-censoring. We explored the effect of demographic and clinical factors previously reported to be associated with LiD. We merged baseline clinical data from all the cohorts. We found that patients with younger PD AAO (grouped as people with age at onset higher than 50 years and lower or equal than 50 years), had a higher probability of developing LiD than older patients along the time interval from disease onset to study end (HR = 1.8, SE = 0.14, $P = 2 \times 10^{-5}$) (data excluding PDBP as AAO was not available). Female PD patients showed a consistent increase in the probability of developing LiD during a 12.5 years time interval (eFig. 2a, b). Body mass index (BMI) was available in PPMI and Tracking Parkinson's, and smoking status data was available in the Tracking Parkinson's cohort only. We did not find a significant increase in the probability of developing dyskinesia either for PD patients with low baseline BMI nor for PD smokers at baseline (eFig. 2c, d).

Power analysis

We performed a power analysis to estimate the power to find a genetic association between time-to-LiD and genome-wide SNPs with the current sample size and LiD event rate, and to evaluate how this varied with a range of genotype hazard ratios and AFs. We were well-powered (80% power) to detect genetic variants associated with the development of LiD with a HR equal or higher than 2 and a minor allele frequency(MAF) as low as 0.01 (eFig. 3a). In addition, we performed a simulation to show as the sample size increases, the power to detect rarer associations improves. As we increased the simulated sample size to 18,000, we achieved 80% power for genetic variants with a MAF lower than 0.01, and with a HR lower than 2 (eFig. 3b).

Time-to-LiD GWAS

We ran time-to-LiD GWAS independently for each cohort, using the first appearance of LiD as the outcome. We confirmed that there was no genomic inflation in any cohort-specific GWAS (eTable 3). We identified three loci significantly associated with time-to-LiD onset in the meta-analysis of the adjusted model on chromosome 1, chromosome 16 and chromosome 4 (Fig. 1). The most significant SNPs at each loci were rs72673189, rs189093213, rs180924818. **rs72673189** (HR = 2.77, SE = 0.18, $P = 1.53 \times 10^{-8}$) in chromosome 1, is a variant in the third intron of the *LRP8* gene. **rs189093213** (HR = 3.06, SE = 0.19, $P = 2.81 \times 10^{-9}$) in chromosome 4 was found in the non-coding RNA *LINC02353 (PCDH7 1.2 Mb downstream)*. **rs180924818** (HR = 3,13, SE = 0.20, $P = 6.27 \times 10^{-9}$) in chromosome 16 was found very close (0.15 Mb upstream) to the 3'-UTR of the *XYLT1* protein coding gene in a non-coding region of the genome (Table 2). The direction of the effects was consistent and replicated across the meta-analysed cohorts in which the SNPs were present (Fig. 2). To visually represent the survival probability of patients carrying the lead SNP on each locus we found in our meta-analysis, we extracted per cohort patients' genotypes and showed the difference in the probability of LiD between carriers and non carriers through Kaplan-Meier curves (Fig. 3).

Sensitivity analysis

The three variants found to significantly increase LiD susceptibility in the adjusted model approach remained associated in the basic model including only known confounders (eTable 4). We found the correlation of the SNP metrics between the basic and the adjusted model to be high (eFig. 4). This indicated that adding additional predictors based on baseline variation increased the

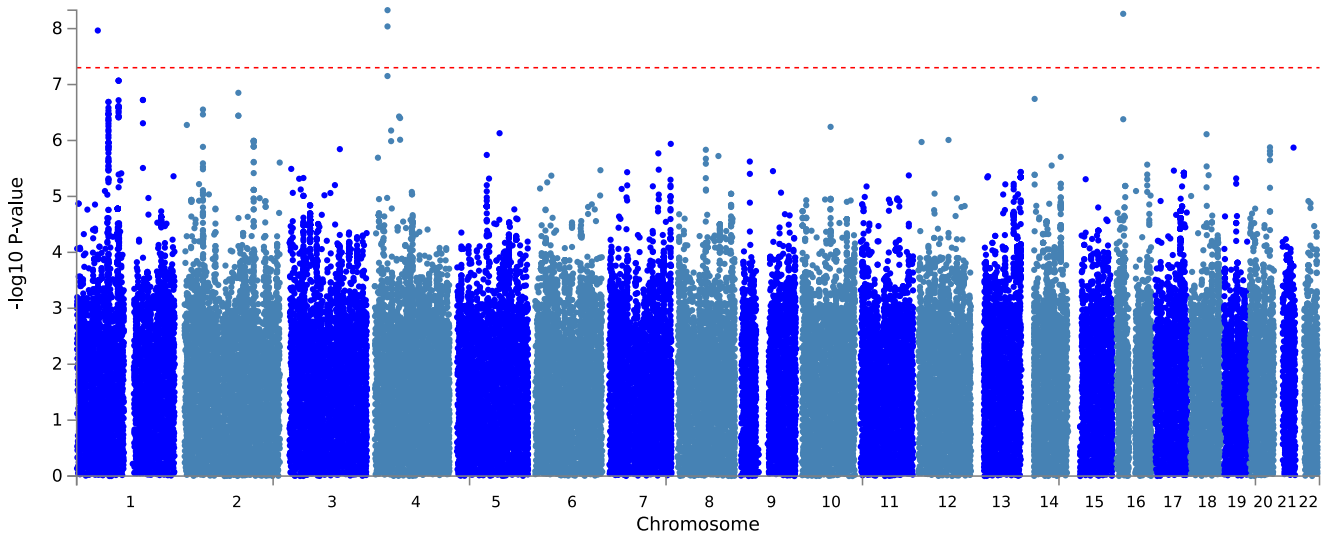

Fig. 1 LiD GWSS meta-analysis Manhattan plot. The GWSS was conducted using a Cox proportional hazards model in each cohort separately, and results were meta-analysed. Red dots indicate the variant with the lowest *P* value at each genome-wide significant genetic *locus*. Genome-wide significance was set at 5×10^{-8} and is indicated by the red dashed line.

Table 2. Independent significant SNPs with a *P* value lower than 1e–7.

CHR	BP	SNP	MAF	BETA	HR	SE	SNP P-value in the Adjusted model	SNP P-value in the Basic model	Number of SNPs	Nearest gene	Function
4	32435284	rs189093213	0.02	1.12	3.06	0.19	1.673e–09	6.15e–08	3	LINC02353	ncRNA intergenic
16	17044975	rs180924818	0.03	1.14	3.13	0.2	6.265e–09	8.20e–08	3	XYLT1	intergenic
1	53778300	rs72673189	0.03	1.02	2.77	0.18	1.527e–08	2.65e–08	2	LRP8	intronic
1	168645690	rs79432789	0.05	0.77	2.16	0.14	7.037e–08	2.47e–06	4	DPT	intergenic
1	39646765	rs71642678	0.01	1.61	5	0.3	8.555e–08	1.89e–07	12	MACF1	intronic
1	80950480	rs12133858	0.04	0.76	2.14	0.14	8.692e–08	1.01e–06	48	RP11-115A15	intergenic
9	22664277	rs77115593	0.02	1.26	3.52	0.24	9.192e–08	4.37e–07	1	LINC02551	ncRNA intronic
14	22020490	rs139943801	0.03	1	2.72	0.19	9.522e–08	2.63e–07	1	RBBP4P5	intergenic

power to detect SNP-outcome associations, presumably by explaining other sources of variance in the model, and that there was no source of confounding given by disease duration and severity measures (suggested by the high correlation in the SNP metrics).

Using data from Tracking Parkinson's only, we investigated whether these associations could be confounded by levodopa dose or the disease stage at the LiD event time point. For each of the genome-wide significant SNPs, we repeated the CPH analysis adjusting for levodopa dose or disease stage as measured by MDS-UPDRS part III at the first visit when the LiD threshold was reached or at the last available visit for patients who did not develop LiD during the study length. We did not find a change either in the hazard ratio or the test-statistics that could suggest an unaccounted source of confounding (eTable 4). Finally, excluding PDBP from the meta-analysis did significantly change the lead SNP's hazard ratio and significance levels eTable 5).

Functional annotation

We performed fine-mapping using ABF, SuSiE, FINEMAP, and Polyfun-SuSiE as described in Methods and found Consensus SNPs on each CPH GWAS nominated loci (eTable 6). We found the lead SNP at each locus to be Consensus SNPs, which are those selected by at least two different fine-mapping tools. We plotted each locus found to have at least one variant significantly associated with time to reach LiD against brain cell type-specific epigenomic data. We found that the lead (and fine-mapped SNP) at the *LRP8* locus belonged to a neuronal specific chromatin accessible region, which is a target region for DNA-associated proteins, as measured with the ATAC-seq and CHIP-seq (H3K27ac and H3K4me3) assays (Fig. 4). We also found this SNP to be part of a neuronal specific enhancer-promoter interaction within *LRP8*, as defined by PLAC-seq (Fig. 4). This implies that this specific *LRP8* intronic signal is an active neuronal enhancer of the *LRP8* expression, forming an anchored chromatin loop recruiting the transcription machinery to the *LRP8* transcription start site. In addition, we found suggestive evidence that the lead SNP lies in a transcription factor binding site (TFBS), as defined by the ENCODE project (eFig. 5). Similarly, we found that some of the fine-mapped SNPs (including the lead SNP) in the *XYLT1* locus were forming chromatin loops towards the *XYLT1* promoter, as measured by the PLAC-seq assay, suggesting that regulation of this gene associated with susceptibility to LiD (eFig. 6). We found this region to also overlap with TFBS marks (eFig. 7). We did not find any functional regulatory marks at the *LINC02353* locus.

Next, we performed colocalization analysis in all genes within 1 Mb from lead SNPs with $P < 1 \times 10^{-7}$. We found suggestive support for

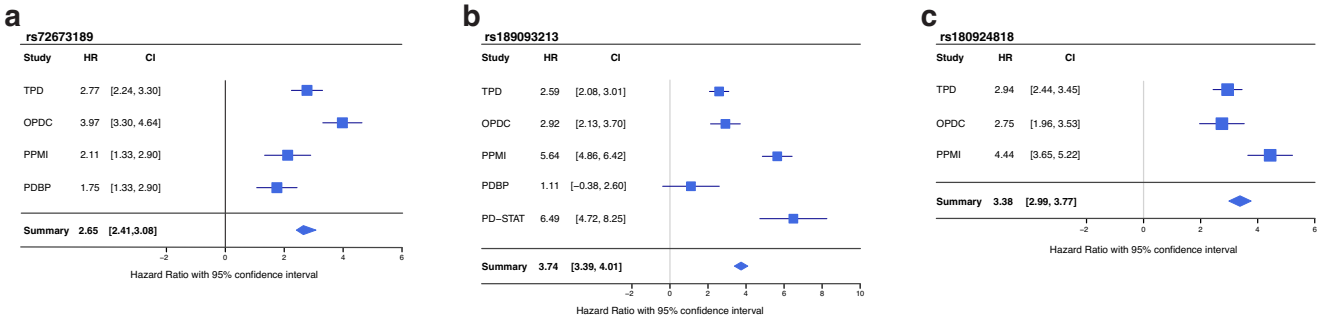

Fig. 2 Forest plots of lead genetic associated variants. a *LRP8* rs72673189 variant ($I^2 = 0$; Q: $\chi^2 = 0.24$, df = 3, $P = 1.53e\text{-}08$). **b** *LINC02353* rs189093213 variant ($I^2 = 21.4$;Q: $\chi^2 = 5.09$, df = 4, $P = 1.67e\text{-}09$). **c** XYLT1 rs180924818 variant ($I^2 = 0$; $\chi^2 = 0.77$, df = 2, $P = 6.27e\text{–}09$). $I^2 = I^2$ Index of heterogeneity HR Hazard ratio, P *P* value, Q Cochran's Cochran's *Q* test of heterogeneity, df degrees of freedom.

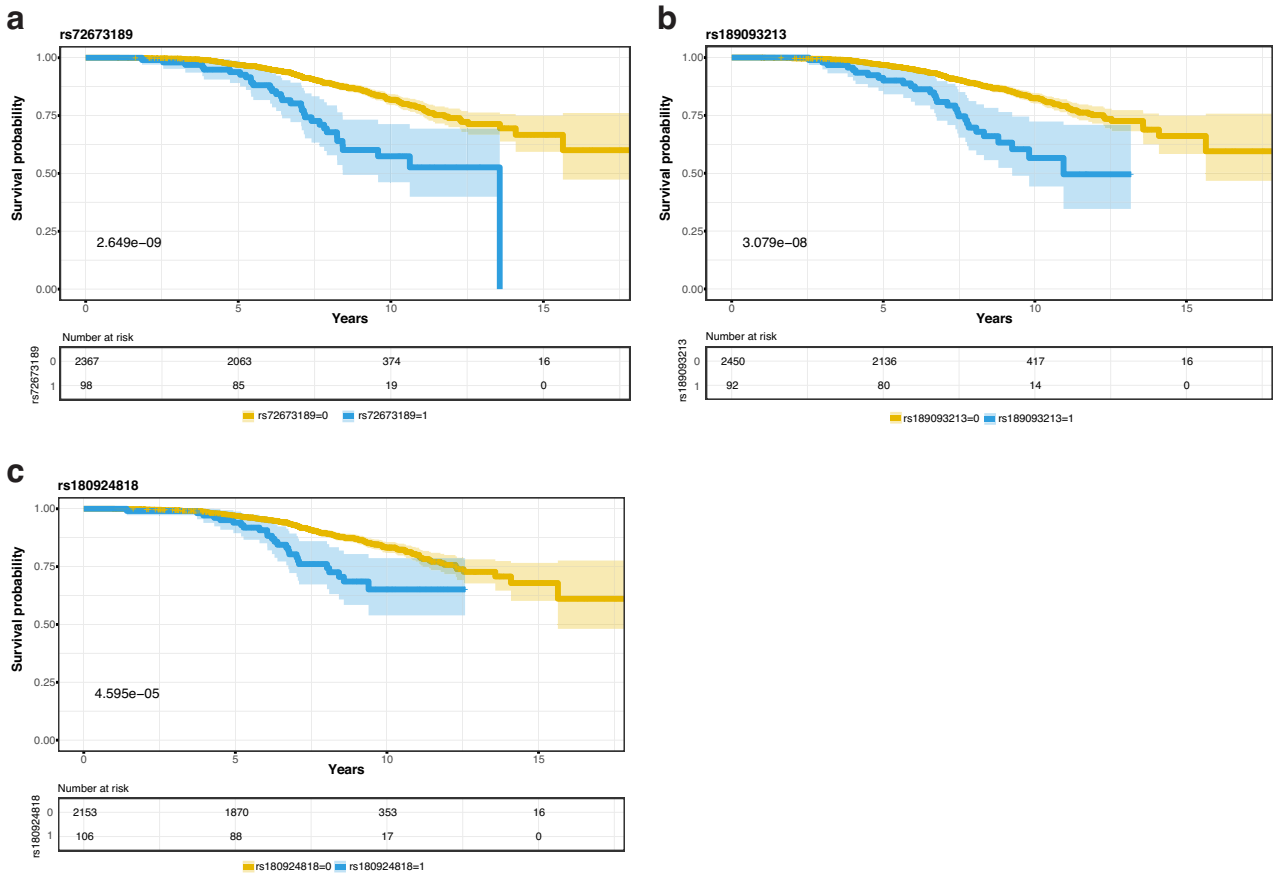

Fig. 3 Survival curves of candidate SNPs. a Kaplan-Meier curve for Survival probability (LiD free probability) based on rs72673189 carrier status in PD patients. **b** Kaplan-Meier curve for Survival probability (LiD free probability) based on rs189093213 carrier status in PD patients. **c** Kaplan-Meier curve for Survival probability (LiD free probability) based on rs180924818 carrier status in PD patients. The blue curve represents genetic variant carriers, whereas the yellow curve represents non-carriers. $p = p$ value. Number at risk represents the number of PD patients remaining on the study at the different time points (0, 5, 10, 15 years). The colour expansion on each curve represents the confidence interval (CI).

colocalization between the LiD GWAS meta-analysis signals and ci-eQTL data from Metabrain Cortex (PP H4 > 0.7 on the unadjusted colocalization analysis; PP H4 > 0.5 on the colocalization analysis after adjusting the priors based on the number of overlapping SNPs in the locus of interest) for the *DNAJB4* gene on chromosome 1 (eTable 7). We did not find evidence of colocalization in the *XYLT1, LRP8 n*or the non-coding RNA loci.

A few loci approaching genome-wide significance GWS in chromosome 1, were in proximity with *DNAJB4*. Therefore, we decided to investigate if the single causal variant assumption holds in the *DNAJB4* locus, necessary to validate the colocalization signal in *DNAJB4*. We ran GCTA-COJO under stepwise and conditional model selection procedures. We filtered all SNPs within the *DNAJB4* locus that were used to perform the colocalization analysis and that matched the AMP-PD reference panel (4590 out of 4840 SNPs included in the colocalization analysis). After performing the stepwise selection procedure assuming complete linkage equilibrium between SNPs that are more than 10 Mb from each other, and setting a collinearity cutoff of 0.9, only the lead SNP in the locus retained nominal significance

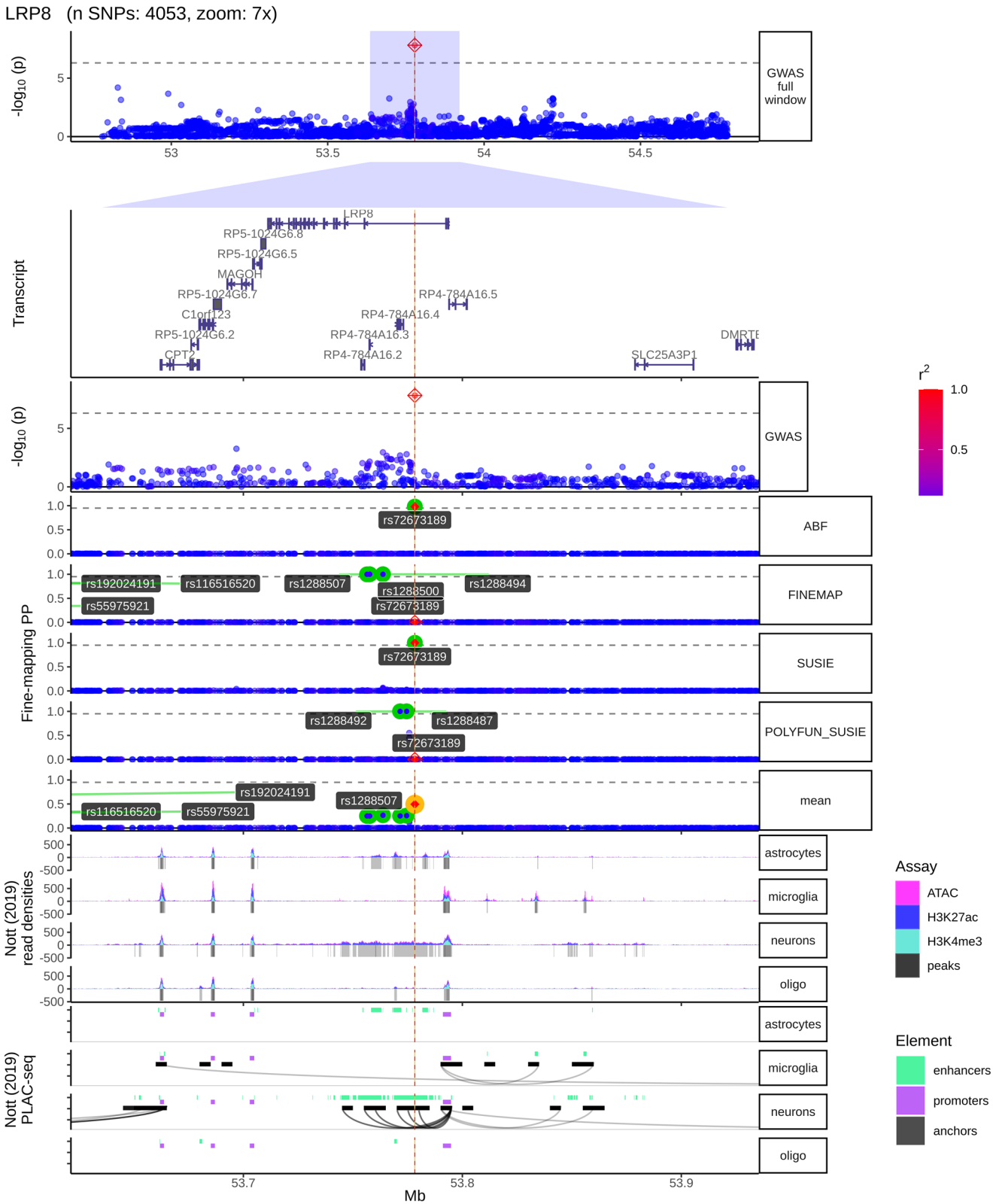

Fig. 4 LRP8 locus fine-mapping and brain cell type specific regulatory marks. From top to bottom, locus plot, transcript plot, the fine-mapping nominated variants across fine-mapping tools, brain cell type specific regulatory element marks. In the locus plot, the SNPs are coloured in red as LD (given by R2) increases, and blue as the LD decreases. In the fine-mapping track, we highlight the SNPs with the highest posterior probabilities for each fine-mapping tool (ABF, FINEMAP, SUSIE, POLYFUN_SUSIE). In addition, we highlight in yellow the Consensus SNP with the highest mean Posterior Probability (mean). In the cell type specific regulatory element marks, the first four rows are the density marks (y-axis) from ATAC-seq assay (in pink), and CHIP-seq assays (H3K27ac in blue, and H3K4me3 in cyan), in astrocytes, microglia, neurons, and oligodendrocytes. The next four rows are the distal anchored chromatin loops (black curves). We see how, only in neurons, there is a chromatin loop forming from the *LRP8* GWS and the fine-mapped consensus variant towards the LRP8 promoter (purple).

(rs278853, MAF = 0.26, beta = 0.40, se = 0.08, $P = 4.07 \times 10^{-6}$). Similarly, running an association analysis on each of the 4590 SNPs conditioning on the lead variant (rs278853) did not show any of these SNPs to be nominally significantly associated, confirming the single causal variant assumption and that the results obtained with coloc on the *DNAJB4* locus were unbiased. Lastly, to understand whether the *DNAJB4* signal was independent of the GWS *LRP8* locus signal, we ran an analysis conditioning on the genome-wide significant *LRP8* SNP (rs72673189). We found that rs278853 remained nominally associated ($P = 4.40 \times 10^{-6}$), indicating these two signals were independently associated with the risk of developing LiD.

Candidate variant analysis

We determined whether previously reported variants in the LiD literature (from LiDPD) had an impact on the time to LiD (eTable 8). We found *ANNK1* and *BDNF* variants to be nominally significantly associated ($P < 0.05$) with the time to dyskinesia. Nonetheless, ANNK1 or BDNF variants did not reach the significance threshold after applying Bonferroni correction according to the number of SNPs tested ($P < 2 \times 10^{-3}$).

LRP8, also known as Apolipoprotein E Receptor 2 (ApoER2), is part of the low-density lipoprotein receptor family[28]. In addition, using western blot analysis based LRP8 knockout mice models, have shown that LRP8 knockout increases the phosphorylation level of the microtubule-stabilising protein tau (MAPT)[29].A previous retrospective study including 855 caucasian PD patients found a suggestive association between the H1b *MAPT* haplotype and a higher likelihood of dyskinesia at an initial visit[30]. In the case of *XYLT1*, a previous study has found a regulatory effect of a *XYLT1* variant on the mRNA levels of *GBA* in the substantia nigra and cortex[31]. We investigated whether MAPT variants (rs1800547; rs242562; rs3785883; rs2435207) were associated with the time to LiD. In addition, we explored whether *APOE* and *GBA* variants increased the risk to develop LiD[32]. We did not find an association between time to LiD and *APOE* variants rs429358 and rs7412, or *GBA* rs2230288 variant (E326K), or MAPT rs1800547, rs242562, rs3785883, rs2435207 variants.

In addition, we explored genetic associations from *PINK1*, *DJ-1*, and *PRKN* intergenic variants. Whereas we did not find any genetic variant associated with time to LiD on the PINK1 locus, we found 26 *DJ-1* intergenic variants on the with a P value < 0.05 (rs1641433611 lead SNP; HR = 1.84, SE = 0.2, $P = 4 \times 10^{-4}$). Similarly, we found 162 intergenic variants with a P value < 0.05 in the PRKN locus (rs113276175 lead SNP; HR = 1.84, SE = 0.2, $P = 4 \times 10^{-4}$) (eTable 9).

PRS is capable of distinguishing patients that develop LiD

We nominated a total of 67 independent SNPs to compute the PRS in the Tracking Parkinson's cohort (eTable 10). We then validated the proposed SNP set on the OPDC cohort by measuring the ability to distinguish LiD PD patients. We found that genetic data as summarised by PRS, without any other clinical or demographic data, could accurately distinguish PD patients that developed LiD at 10 years from disease onset in two separate cohorts: Tracking Parkinson's (AUC 83.9) and OPDC (AUC 87.8) (eFig. 8). At 10 years from PD onset, we found that 16% of patients had LiD in the Tracking Parkinson' cohort, and 18% of patients had LiD in the OPDC cohort. Class imbalance can lead to inaccurate evaluation of classifiers. Therefore, we also computed precision recall curves (PROC) as large class imbalance can lead to biassed ROC curves when assessing the performance of a classifier. We found the PROC AUC to be lower in both TPD (AUC = 54.49) and OPDC (AUC = 33.24) (eFig. 9).

Stepwise regression approach to determine baseline predictors of LiD development

We used Tracking Parkinson's data at baseline in a stepwise regression approach using a logistic model. We then filtered out from the final model predictors that were not significantly associated after applying Bonferroni correction ($P < 0.05/\ 502 = 1 \times 10^{-4}$).

In addition to the PRS, which was significantly associated with a increase of the odds of LiD (OR = 962.94, SE = 0.57, $P = 1.07 \times 10^{-30}$), we found that anxiety at baseline (as measured by the Leeds Anxiety and Depression Scale[33]) was significantly associated with a increase of the odds of LiD (OR = 1.14, SE = 0.03, $P = 7.4 \times 10^{-5}$). We also explored clinical features previously reported as being associated with an increased or decreased LiD risk. Sex, AAO, and 5 principal components (PCs) were added in the base model of the stepwise regression approach. Consistent with previous studies as well as with our CPH model highlighted above, younger AAO increased the LiD odds (OR = 2.41, SE = 0.04, $P = 4 \times 10^{-3}$). However, sex was not found to be significantly associated in our final model including PRS and Leeds anxiety status.

Neither smoking status nor BMI were selected on the stepwise regression approach, consistent with what we found when we individually explored known LiD risk factors (eFig. 2). Interestingly, we also found that PD family history was selected on the stepwise regression analysis, and was nominally significantly associated with an increase in the odds of LiD (OR = 1.62, SE = 0.14, $P = 6.9 \times 10^{-4}$).

Finally, we attempted to replicate the association between dyskinesia state and anxiety using State-Trait Anxiety Inventory[34] available in PPMI. We did not find the Trait Anxiety Score to be significantly associated with LiD patients in PPMI (OR = −0.03, SE = 0.04, $P = 0.44$).

Patients with LiD have an average higher cognitive scoring

We assessed the cognitive status of LiD patients because of the association between the *LRP8* nominated locus and *APOE*. We explored whether the cognitive state differed between patients developing LiD and patients who did not develop LiD during the study length using the Wilcoxon rank sum non-parametric test with continuity correction, as we observed the data was not normally distributed. In addition, we also looked into differences in the MDS-UPDRS part III scores between the two groups, using the unpaired two samples t-test to compare the mean of two independent groups. We compared the LiD group ($N = 172$) against the non-LiD PD group ($N = 1318$) using data from Tracking Parkinson's alone as it is the largest deeply phenotyped cohort we had available. We did not find differences in the average MDS-UPDRS part III total score, either at baseline nor at the visit when patients first developed LiD (or the last available visit in cases who did not develop LiD) (eTable 11). However, PD patients who did not develop LiD through the study had a significantly lower MoCA score on average at baseline, as well as at the final visit (eTable 11).

DISCUSSION

We have performed an untargeted genome-wide study to define genetic variants associated with the time-to-LiD in PD, using a CPH model under a genetic additive effect and analysed the effect of genetic and baseline clinical variation on the development of LiD. We found genome-wide significant associations with the time-to-develop LiD at the *LRP8*, *LINC02353* and *XYLT1* loci. These associations were replicated across all the cohorts included in the meta-analysis. We also performed a candidate gene analysis, exploring genetic variants reported to be associated with LiD risk in our large GWAS meta-analysis. We found that genetic variability

in *BDNF* and *ANKK2*, were nominally associated with LiD. We did not replicate any other variant associated with LiD risk (eTable 8).

LRP8, also known as ApoER2, is a cell surface receptor part of the low-density lipoprotein receptor-family. Its expression is enriched in brain tissues such as the neocortex, cerebellum, hippocampus and olfactory bulb[28]. *LRP8*, together with *VLDLR*, is a mediator of the Reelin pathway, which contributes to development of the central nervous system as well as to facilitate neuronal migration[35,36]. LID develop in the context of ongoing neuronal loss, and synaptic/signalling changes related to dopamine therapy. Our finding suggests the changes in the Reelin pathway and neural development / plasticity may be important in the development of LID.

In addition, the *LRP8* protein stabilises *MAPT* and it has been shown that knocking out *LRP8* in mice increases tau phosphorylation[29]. Post-hoc functional annotation analysis revealed a chromatin loop between an enhancer within *LRP8* third intron (where the lead variant was found) and the *LRP8* promoter, thus providing functional support for *LRP8* as the causal gene at this locus. In addition, a colocalization analysis, looking at all genes within ±1 Mb from all GWAS variants with P value $< 1 \times 10^{-7}$ revealed a second association in chromosome 1 with the *DNAJB4* gene. Conditional analysis further confirmed that both regions were in linkage equilibrium, hence both *LRP8* and *DNAJB4* were independently associated with the time-to-LiD. We also found a similar event of distal regulation in the *XYLT1* locus, although the chromatin loop did not perfectly match with the GWAS signals, making the functional annotation analysis inconclusive. Moreover, we found that the two GWAS nominated signals overlapped with Transcription Factor Binding Sites marks from the ENCODE project, adding further support for the transcription machinery being recruited in the GWAS loci and regulating both genes expression after forming the enhancer-promoter distal chromatin loops. Nevertheless, whereas we found a chromatin loop suggesting regulation of *XYLT1* and *LRP8* gene expression, we did not find statistical support for gene regulation based on the colocalization Bayesian framework.

The three nominated protein coding genes have been previously reported to be functionally associated with putative PD genes, which may provide an insight into the development of LiD. *LRP8* encodes the low-density lipoprotein receptor-related protein 8, and it has been found to be associated with *APOE*. In addition, the *LRP8* protein stabilises *MAPT* and it has been shown that knocking out *LRP8* in mice increases tau phosphorylatlon[29]. DNAJB4 gene encodes a molecular chaperone tumour suppressor, and member of the heat shock protein-40 family. Mutations in the DNAJ family protein have been reported to cause or increase the risk of several neurological disorders, including Parkinson's disease[37]. *XYLT1* encodes a xylosyltransferase enzyme which takes part in the biosynthesis of glycosaminoglycan chains. A previous study has found a regulatory effect of a *XYLT1* variant on the mRNA levels of *GBA* in the substantia nigra and cortex[31]. We did not find support for colocalization with eQTLs nor evidence suggestive of epigenetic regulation of genes in the *LINC02353* locus. *PCDH7*, the nearest gene coding protein gene, encodes a protein with an extracellular domain containing 7 cadherin repeats. This gene has been described as a potential PD biomarker[38].

At an individual patient level, treatment strategies including levodopa and non-levodopa therapies, and the use of deep brain stimulation are determined by the emergence of motor complications including LID. The ability to develop a predictive algorithm to enhance clinical care would improve the outlook for PD treatment. Here, we have shown that both clinical and genetic variables have the potential to have a high predictive value for the development of LID. This will need to be validated in further cohorts and we hypothesise that the integration of further 'omics data (e.g. RNA and proteomics), using machine learning may lead to the definition of an accurate predictive model for defining PD patients at risk of developing dyskinesia.

We have analysed a large dataset with detailed clinical, drug exposure and genetic data. We have carefully tested for confounding by PD age at onset, sex, population structure and shown that our results are free of confounding effects as well as demonstrating they are consistent across cohorts. Because the dose of levodopa may be a major confounder in our study, we tested the effects of adjusting for levodopa dose on a sensitivity analysis, and found that the lead SNPs on *LRP8*, *LINC02353* and *XYLT1* loci remained significantly associated with the outcome, concluding that levodopa treatment was not a confounder in our study design. Likewise, adjusting for the MDS-UPDRS part III total score at the time of LiD development did not change the significance levels of the lead SNPs, suggesting that our findings were not confounded by motor severity or progression.

Although this is a large study there are limitations based on sample size. According to our sample calculation, we would be 80% powered to detect associations with the LiD phenotype from variants with a MAF of 0.01 when we reached a sample size of 18,000 patients. In addition, our results are limited to individuals of European ancestry and we have not explored whether there is a shared common genetic variability associated with changes in LiD survival across different ancestries. Expanding this analysis to PD genetic datasets with deeply phenotypic data available from initiatives such as the Global Parkinson's Genetic Programme (GP2) will give us new insight into the genetics of PD LiD patients as well as serve as a valuable resource for validation of findings[39].

MDS-UPDRS 4.1 is a simple but widely used measured which documents the appearance of LiD. Potentially, more detailed scales such as the Unified Dyskinesia Rating Scale[40] would provide a more accurate measure of the extent and impact of LiD, which would improve future GWAS.

Overall, we have found new evidence of common genetic variability associated with the time-to-LiD. We have been able to map genes at nearby risk loci, as well as provide fine mapping support of potential causal variants for LiD trait. Likewise, we hope to help design personalised medicine strategies that prevent PD patients developing dyskinesia according to their genetic burden which could be tested with the proposed PRS in this study. Similarly, we hope to help understand the molecular pathways that lead to LiD. Targeting nominated genes might allow the development of LiD treatment strategies. Further investigation regarding the overlap between anxiety GWAS and our GWAS might help understanding common causal pathways between the two conditions. Understanding shared mechanisms will help us prevent medication adverse events affecting non-targeted pathways and to fine-tune current treatments.

METHODS

The source code with all materials and methods are available on GitHub (https://github.com/AMCalejandro/LID-CPH.git; https://doi.org/10.5281/zenodo.8139563). The README explains each step of the workflow to conduct the analysis and a link to each relevant pipeline or protocol.

Patients data and LiD definition

We accessed clinical and genetic data from the Tracking Parkinson's (Tracking Parkinson's)[41], Oxford Parkinson's Disease Centre Discovery Cohort (OPDC)[42], Parkinson's Progression Markers Initiative (PPMI)[43], Parkinson's Disease Biomarkers Programme (PDBP)[44], and simvastatin as a neuroprotective treatment for PD trial (PD-STAT)[45] studies (eTable 1). Each subject provided written informed consent for participation according to the Declaration of Helsinki and all cohort studies were approved by the relevant ethics committee.

We carried out clinical data QC on each cohort independently (eFig. 1). Levodopa is necessary for PD patients to develop LiD[6], therefore we excluded those who were not exposed to levodopa. In addition, we removed patients who had a disease duration at study entry of more than 10 years from disease onset, patients without longitudinal data (patients with less than two clinical records available), and those with missing genotype data.

We defined PD patients as having dyskinesia if they reached an MDS-UPDRS item 4.1 score equal to or higher than two which is equivalent to a range of 26–50% of the waking time with dyskinesia, and the first appearance of LiD was defined as the event time. Patients were excluded if they had dyskinesia at study entry, as time to the development of dyskinesia could not be established.

Genotype data quality control and imputation

To perform quality control (QC) at both the sample and genotype levels before and after imputation of genotyping data, we used PLINK v1.9 (RRID:SCR_001757; https://www.cog-genomics.org/plink/1.9/)[46].

Sample level QC. We used X chromosome genotype data to check for sex discordance between the genotypic and phenotypic sex. We excluded individuals who were missing more than 5% of genotypes. Samples with excess or reduced heterozygosity in autosomes (defined as ± 4 standard deviations (SD) away from the mean heterozygosity rate within each cohort) were also excluded, as it can indicate contamination or increased homozygosity, respectively. We removed related individuals. Using GCTA software (version 1.93.0 beta for Linux; https://yanglab.westlake.edu.cn/software/gcta/#Overview)[47], we created a relationship matrix from pruned genotypes, and we filtered out individuals which had a similarity score higher than 0.125, equivalent to 1st degree relatives. Population stratification is a major confounder in genetic association studies due to differences in allele frequencies between ethnicity groups. We therefore excluded individuals of non-European ancestry by performing principal component analysis using the HapMap reference panel (Release number 3; ftp://ftp.ncbi.nlm.nih.gov/hapmap/)[48]. Individuals who were 6 SD away from the Northern and Western European ancestry (CEU) sample mean for any of the first 10 PCs were considered ancestry outliers and removed.

Variant QC. We removed variants that had a missing rate higher than 0.05, variants with a MAF of less than 0.01, and variants in which missing calls were not randomly distributed, by testing whether missingness status could be predicted from genotype calls at the two adjacent variants. We excluded variants that deviated from Hardy-Weinberg equilibrium (HWE), as extreme HWE deviations can be indicative of genotyping errors ($P < 1 \times 10^{-10}$)[49].

Imputation and post-imputation QC. We ran the Will Rayner tool (Version 4.2.10; https://www.well.ox.ac.uk/~wrayner/tools/) for further quality checks against the Haplotype Reference Consortium (HRC) (GRCh37/hg19) panel (version r1.1 2016; http://www.haplotype-reference-consortium.org/site). Likewise, we updated strand, position, and reference / alternate allele assignment, as well as removed A/T and G/C SNPs if MAF > 0.4, SNPs with allele frequency difference >0.2 compared to the reference panel, and SNPs not present in the HRC Panel[50]. We then imputed each QCed cohort in the Michigan Imputation Server (RRID:SCR_017579; https://imputationserver.sph.umich.edu)[51] using Minimac4 (version 1.0.0; RRID:SCR_009292 https://genome.sph.umich.edu/wiki/Minimac4 version 1.0.0) and Eagle2 (v2.4; RRID:SCR_017262) with 20-Mb chunk sizes used to estimate haplotype phasing. We used the HRC panel as the reference panel of individuals of predominantly European ancestry for imputation. Once the data was imputed, we filtered out variants with a Rsq score <0.7, to preserve only the variants imputed with high confidence. Finally, we removed variants with missingness rate >5% and MAF < 1%.

Whole-genome sequencing data

The PDBP and PPMI cohorts included in this study were whole-genome sequenced using Illumina HiSeq X Ten Sequencer. More information can be found in https://ida.loni.usc.edu/login.jsp. WGS data was QCed using the same pipeline as the array-based data.

Statistical analyses

We used the R programming language to perform all the statistical analysis (R Project for Statistical Computing, RRID:SCR_001905; version 4.1.3; https://www.R-project.org/).

We studied the association between genome-wide genetic variants and time to develop dyskinesia from self-reported age at PD motor onset with Cox proportional hazard (CPH) regression models under a genetic additive model, using the 'survival' R package (version 3.3-1; RRID:SCR_021137; https://cran.r-project.org/web/packages/survival/survival.pdf). All tests were two-tailed. To investigate the power to detect an association under a Cox regression model with the current sample size, as well as to perform a simulation on the relationship between power and allele frequency (AF), SNP hazard ratios (HR), and sample size, we used the R package survSNP (version 0.25; https://cran.r-project.org/web/packages/survSNP/index.html).

We ran time-to-LiD GWAS in each cohort separately, adjusting by AAO (or age at diagnosis in the cohorts where AAO was not available), sex, and first 5 PCs, using as our outcome the midpoint between the visit the threshold was met and the previous time point. Because the precise time patients first developed dyskinesia happens between the visit in which dyskinesia were first recorded and the previous visit, we set the time to develop dyskinesia as the midpoint between the last visit with no LiD and the first visit with LiD. We set age at motor onset as the start point to measure time to develop LiD. Patients who did not develop LiD at the end of the study or at the time of withdrawal were right-censored. For patients that withdrew from the study and that did not have the withdrawal time available, we set the censoring time as the midpoint between the last visit patients attended clinic and the next scheduled visit.

Multiple studies indicate that the risk of dyskinesia relates to disease severity. To improve the power to detect a genetic association, we explored the goodness-of-fit of the model in each cohort independently after adding the following baseline covariates, which provide surrogate measures of disease severity and dopaminergic denervation at baseline: levodopa or LEDD dose, disease duration from onset to baseline assessment and baseline motor score as measured by MDS-UPDRS part III. For each cohort, we selected the model which provided the most accurate prediction of LiD based on the Akaike Information Criteria (AIC). We used the resulting model as the main model in our analysis. We summarised the nominated set of covariates in each cohort (eTable 2). We verified that the proportional hazards assumption held true by assessing the independence between scaled Schoenfeld residuals and time through the cox.zph function from the 'survival' package. Schoenfeld residuals are obtained by subtracting the individuals' covariate values at the time '*t*' and the corresponding risk-weighted average of covariates among all those that are at risk at the time '*t*'. Then, they are scaled by performing a variance-weighted transformation. A non-significant relationship between the scaled residuals and time reveals proportionality of the hazards in the model.

We used METAL software (version released on the 2011-03-25; RRID:SCR_002013; https://genome.sph.umich.edu/wiki/

METAL_Documentation) for meta-analysis of genome wide association summary statistics, with a fixed effects model weighted by β coefficients and the inverse of the standard errors[52,53]. We applied a genomic control correction to the cohort-specific summary statistics by computing the inflation of the test statistic, and then applying the genomic control correction to the standard errors. We chose a meta-analysis over a merged analysis because of the heterogeneity in the inclusion and exclusion criteria across the clinical cohorts, as well as differences in the genotyping approaches (eTable 1). We applied a post meta-analysis QC step to remove genetic variants that were present in less than 3 out of 5 cohorts, with less than 1000 variants, as well as variants with high MAF heterogeneity across the cohorts (MAF > 0.15). In addition, we accounted for high heterogeneous variants by removing those with a signifiant Cochran's Q test as well as those with an I2 index higher than 80%.

Statistical significance was assessed at the conservative threshold of $P = 5 \times 10^{-8}$, derived from a Bonferroni correction accounting for the number of independent tests and the linkage disequilibrium (LD) structure of the genome[54].

We proved that the model met the proportional hazard assumption after including significant SNPs using the cox.zph function from the 'survival' package. We evaluated whether signals were replicated across different cohorts with the R package 'forestplot' (version 2.0.1; https://CRAN.R-project.org/package=forestplot).

Sensitivity analyses

To validate the genome wide significance findings, we performed four sensitivity analyses to discard the associations we found in our analysis were confounded. The first sensitivity analysis was designed to compare the basic and adjusted models. We tested whether high deviations in the SNP estimates and P-values arose after accounting for disease severity and dopaminergic denervation at baseline by measuring the correlation between the basic and adjusted GWAS meta-analyses. Next, we performed two separate sensitivity analyses to test whether either levodopa dose or the PD motor severity (as measured by MDS-UPDRS part III) at the time point where LiD were first documented, were confounding our findings. We performed this sensitivity analysis in Tracking Parkinson's, the largest dataset. We performed a CPH GWAS on the Tracking Parkinson's cohort adjusting by: a) known confounders, b) known confounders + motor severity (as measured by MDS-UPDRS part III) c) known confounders + levodopa dose. We compared the SNP metrics from the three models for the lead SNPs on the loci that reached genome-wide significance on the time-to-LiD GWAS meta-analysis. Lastly, because the PDBP cohort did not have age at onset available and we used AAD in the CPH model, we reran the time-to-LiD GWAS meta-analysis excluding PDBP to confirm that this cohort was not inflating the SNP test-statistics.

Post-GWAS analyses

We used the 'echolocatoR' R package (v 0.2.2; https://github.com/RajLabMSSM/echolocatoR) as a wrapper to perform fine-mapping which allows us to nominate causal variants for further study. In particular, we used the ABF approach through the 'coloc' R package, FINEMAP software in Unix (v v1.3; http://www.christianbenner.com/), the 'susieR' R package (v 0.11.92; https://cran.r-project.org/web/packages/susieR/index.html), and Polyfun-SuSiE(V1.0; https://github.com/omerwe/polyfun)[55–59]. We produced the 95% Probability Credible Set ($CS_{95\%}$), which is the minimum set of SNPs that contains all causal SNPs with 95% probability. We reported the consensus SNPs at each locus, i.e. those that were included in the 95 $CS_{95\%}$ of at least two fine-mapping tools, therefore increasing the confidence in the nominated causal SNPs. We reported the Posterior Probability (PP) as the mean PP across all fine-mapping tools. To account for SNP LD at each region, we used the precomputed LD matrix from the UK Biobank (https://alkesgroup.broadinstitute.org/UKBB_LD/)[60].

To evaluate the potential effect of SNPs on candidate loci on the control of gene expression we also used echolocatoR as a wrapper to access brain cell type-specific epigenetic marks from Nott and colleagues[61,62] (Data accessed using echolocatoR v 0.2.2). We mapped each locus to cell type-specific chromatin immunoprecipitation sequencing (ChIP-seq) results generated by quantifying H3K4me3 and H3K27ac epigenetic modifications, Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) results, and Proximity Ligation-Assisted ChIP-Seq (PLAC-Seq) results, to detect and quantify chromatin contacts anchored at genomic regions. In addition, we also mapped such loci to cell-type specific TFBS marks on Chip-seq experiments from the ENCODE project (RRID:SCR_006793; data accessed from echolocatoR R package v 0.2.2)[61,62]. This dataset contains 690 Chip-seq datasets representing 161 unique regulatory factors and spanning 91 human cell types. We used echolocatoR to query the ENCODE Uniform TFBS and retrieve the top four cell types with the highest probability density function for the top five regulatory elements.

To investigate whether there were several independently associated SNPs at each GWAS nominated locus, we performed a conditional and stepwise selection procedure with GCTA-COJO (version 1.93.0 beta for Linux; https://yanglab.westlake.edu.cn/software/gcta/#Overview)[47]. We used the Accelerating Medicines Partnership: Parkinson's Disease (AMP-PD, v.2.5)[63] data ($n = 10{,}418$) as the reference panel to estimate the correlation between SNPs. The reference sample was subjected to the same QC steps as described above, needed to get unbiased LD estimates[64].

We used the 'coloc' R package (version 5.1.0; https://cran.r-project.org/web/packages/colocr/index.html) to perform colocalization analysis between the SNPs associated with progression to LiD and SNPs defining gene expression in the region. We used cis-eQTL data from MetaBrain cortex tissue[65] ($N = 6{,}601$ individuals) and blood cis-eQTLs from eQTLGen ($N = 31{,}684$)[66]. We evaluated all genes within ±1 Mb from the lead variants with a $P < 1 \times 10^{-7}$ at each GWAS locus[6]. Coloc makes use of Bayesian inference to compute the posterior probability (PP) of five different hypothesis: No association with either trait (H0); Association with the LiD trait but not the eQTL trait (H1); Association with the eQTL trait but not the LiD trait (H2); Association with both traits, but the causal variant is distinct (H3); and the that there is a shared causal variant associated with both traits (H4). Each PP hypothesis lies between 0 and 1 and a high PPH4 (PPH4 > 0.8) is considered as evidence of colocalization between two traits tested, meaning the GWAS variant causes changes in specific gene expression. We ran coloc using default $p_1 = 1 \times 10^{-4}$, $p_2 = 1 \times 10^{-4}$, and $p_{12} = 1 \times 10^{-5}$ priors (p_1 and p_2 are the prior probability that any random SNP in the region is associated with trait 1 and 2, respectively, while p_{12} is the prior probability that any random SNP in the region is associated with both traits). However, it is worth noting that the prior for H3 hypothesis (association with both phenotypic and expression traits, but distinct causal variants) is $\approx n(n-1)p1.p2$, which scales with the square of n, resulting in H3 becoming more likely than H4 as the number of overlapping SNPs in the region increases[7]. Therefore, we adjusted the priors to account for the high number of overlapping SNPs ($p_1 = 3 \times 10^{-5}$, $p_2 = 3 \times 10^{-5}$, and $p_{12} = 5 \times 10^{-7}$)[8].

We used Functional Mapping and Annotation of Genome Wide Association Studies (FUMA) (RRID:SCR_017521; version 1.3.8; https://fuma.ctglab.nl/) to further characterise the nominated loci by querying GWAS Catalogue to retrieve uncharacterised GWAS loci SNPs in our meta-analysis and to get positional mapping information based on MAGMA[67]. We used a threshold of $P < 1 \times 10^{-6}$ to nominate tag SNPs. Additional SNPs that were in high LD with tag

SNPs were inferred using European samples 1kg Phase3 reference panel (with $r2 > 0.6$ and independent from each other with $r2 < 0.6$).

Candidate gene analysis

In order to validate variants that have been reported in previous studies to be associated with time-to-LiD or LiD risk, we accessed the LiDPD website (Date accessed: 12/01/2023; http://LiDpd.eurac.edu/) and downloaded a list of curated variants from the literature. We explored these in our time-to-LiD GWAS meta-analysis[68].

LiD prediction modelling

We used PRSice software (version 2; RRID:SCR_017057) to compute a polygenic risk score (PRS) using the summary statistics of our time-to-LiD meta-analysis as base data and the Tracking Parkinson's cohort as target data. We chose the Tracking Parkinson's cohort as it is the single largest cohort, which reduces the standard error (SE) of the PRS estimates, leading to more confident estimates. We then replicated the association of the nominated SNPs composing the PRS in the second largest cohort we had access to, OPDC, resembling a discovery/replication study design, although in this case the OPDC data had contributed to the LID PRS.

We set a threshold of $P < 1 \times 10^{-6}$ to nominate GWAS variants that make up the PRS. We selected independent SNPs by clumping within ±250 Kb from the index SNPs (the most significant SNP on a Kb window). We used the SNP betas as the estimated to compute the PRS from. Sex, standardised AAO, and the first 5 PCs were added as covariates to the PRS estimation process. To compute the LD estimates, we used the imputed cohorts from which we calculated the PRS, as they were large enough to provide accurate LD estimates ($N > 500$). To validate the PRS as an instrument to distinguish between PD patients with and without LiD, we derived time-dependent ROC curves, under the assumption that different PRS loads might cause changes to time-to-LiD onset. We used the Inverse Probability of Censoring Weighting estimation of Cumulative/Dynamic time-dependent ROC curve from the 'timeROC' R package (version 0.4; https://cran.r-project.org/web/packages/timeROC/index.html). To compute the weights, we used the Kaplan-Meier estimator of the censoring distribution.

Next, we used a stepwise logistic regression model with an in-house script using the 'stats' R base package (version 4.2.2; https://search.r-project.org/R/refmans/stats/html/00Index.html) to find whether any baseline clinical variable was significantly associated with LiD status. We used data from the Tracking Parkinson's cohort, as it is deeply phenotypically characterised (number of baseline covariates = 702). After removing variables with high missingness rate (missing rate >10%) or categorical variables with only one level, we defined a total of 502 baseline features (including the PRS) (eData 1). Then, we created a base logistic regression model (adjusted for sex the first 5 PCs and standardised AAO). At each step of the stepwise regression approach, we refitted the base model with each of the baseline predictors individually, and selected the model with the variable that decreased AIC the most. We ran the model until no variable further decreased the AIC, or until the AIC score was equal to 1. Once the model was fitted, we selected only those predictors that were significantly associated with the binary outcome, applying the conservative Bonferroni correction accounting for the number of predictors assessed. We set the significance threshold as $0.05 / 502 = 1 \times 10^{-4}$. To account for class imbalance in the evaluation of classifiers, we computed precision recall curves using the 'PRROC' R package (version 1.3.1; https://cran.r-project.org/web/packages/PRROC/index.html).

DATA AVAILABILITY

GWAS summary statistics are publicly available in the Zenodo ASAP data repository (https://doi.org/10.5281/zenodo.7795604). Supplementary Figures and Tables are available in the Zenodo ASAP data repository (https://zenodo.org/record/7802755#.ZC2RAnbMK38). TPD data are available upon access request from https://www.trackingparkinsons.org.uk/about-1/data/. The PDBP and PPMI data was accessed from Accelerating Medicines Partnership: Parkinson's Disease (AMP-PD) and data are available upon registration at https://www.amp-pd.org/. OPDC data are available upon request from the Dementias Platform UK (https://portal.dementiasplatform.uk/Apply). PD-STAT is available upon request to the principal investigator (C Carroll, Plymouth University, https://penctu.psmd.plymouth.ac.uk/pdstat/#:~:text=PD%20STAT%20%2D%20Simvastatin%20as%20a,brain%20from%20injury%20or%20loss.). HapMap phase 3 data (HapMap3) is available for download at ftp://ftp.ncbi.nlm.nih.gov/hapmap/. Cis-QTL eQTLGen data was downloaded from (https://www.eqtlgen.org/cis-eqtls.html). MetaBrain cis-eQTL data can be accessed upon access request form (https://www.metabrain.nl/cis-eqtls.html). eQTL data from eQTL catalogue can be ftp-accessed (https://www.ebi.ac.uk/eqtl/Data_access/). ENCODE TFBS marks and Nott brain cell type-specific enhancer-promoter interactome data were accessed through echolocatoR (https://github.com/RajLabMSSM/echolocatoR).

CODE AVAILABILITY

All the code has been made publicly available on GitHub (https://github.com/AMCalejandro/LID-CPH.git) https://doi.org/10.5281/zenodo.8139563) Analyses were performed using open-source tools as described in the Methods section.

Received: 25 May 2023; Accepted: 16 August 2023;
Published online: 31 August 2023

REFERENCES

1. Kalia, L. V. & Lang, A. E. Parkinson's disease. *Lancet* **386**, 896–912 (2015).
2. Gibb, W. R. & Lees, A. J. The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry* **51**, 745–752 (1988).
3. Fahn, S. et al. Levodopa and the progression of Parkinson's disease. *N. Engl. J. Med.* **351**, 2498–2508 (2004).
4. Verschuur, C. V. M. et al. Randomized delayed-start trial of Levodopa in Parkinson's Disease. *N. Engl. J. Med.* **380**, 315–324 (2019).
5. Jankovic, J. & Tan, E. K. Parkinson's disease: etiopathogenesis and treatment. *J. Neurol. Neurosurg. Psychiatry* **91**, 795–808 (2020).
6. Espay, A. J. et al. Levodopa-induced dyskinesia in Parkinson disease: current and evolving concepts. *Ann. Neurol.* **84**, 797–811 (2018).
7. Manson, A., Stirpe, P. & Schrag, A. Levodopa-induced-dyskinesias clinical features, incidence, risk factors, management and impact on quality of life. *J. Parkinsons Dis* **2**, 189–198 (2012).
8. Tran, T. N., Vo, T. N. N., Frei, K. & Truong, D. D. Levodopa-induced dyskinesia: clinical features, incidence, and risk factors. *J. Neural Transm.* **125**, 1109–1117 (2018).
9. Cilia, R. et al. The modern pre-levodopa era of Parkinson's disease: insights into motor complications from sub-Saharan Africa. *Brain* **137**, 2731–2742 (2014).
10. Khan, N. L. et al. Parkin disease: a phenotypic study of a large case series. *Brain* **126**, 1279–1292 (2003).
11. van Duijn, C. M. et al. Park7, a novel locus for autosomal recessive early-onset parkinsonism, on chromosome 1p36. *Am. J. Hum. Genet.* **69**, 629–634 (2001).
12. Lin, M. K. & Farrer, M. J. Genetics and genomics of Parkinson's disease. *Genome Med.* **6**, 48 (2014).
13. Lohmann, E. et al. A multidisciplinary study of patients with early-onset PD with and without parkin mutations. *Neurology* **72**, 110–116 (2009).
14. Oliveri, R. L. et al. Dopamine D2 receptor gene polymorphism and the risk of levodopa-induced dyskinesias in PD. *Neurology* **53**, 1425–1430 (1999).
15. Darmopil, S., Martín, A. B., De Diego, I. R., Ares, S. & Moratalla, R. Genetic inactivation of dopamine D1 but not D2 receptors inhibits L-DOPA-induced dyskinesia and histone activation. *Biol. Psychiatry* **66**, 603–613 (2009).
16. Falla, M., Di Fonzo, A., Hicks, A. A., Pramstaller, P. P. & Fabbrini, G. Genetic variants in levodopa-induced dyskinesia (LID): a systematic review and meta-analysis. *Parkinsonism Relat. Disord.* **84**, 52–60 (2021).
17. de Lau, L. M. L., Verbaan, D., Marinus, J., Heutink, P. & van Hilten, J. J. Catechol-O-methyltransferase Val158Met and the risk of dyskinesias in Parkinson's disease. *Mov. Disord.* **27**, 132–135 (2012).

18. Yin, Y., Liu, Y., Xu, M., Zhang, X. & Li, C. Association of COMT rs4680 and MAO-B rs1799836 polymorphisms with levodopa-induced dyskinesia in Parkinson's disease—a meta-analysis. *Neurol. Sci.* **42**, 4085–4094 (2021).
19. Solís, O. et al. Human COMT over-expression confers a heightened susceptibility to dyskinesia in mice. *Neurobiol. Dis.* **102**, 133–139 (2017).
20. Kusters, C. D. J. et al. Dopamine receptors and BDNF-haplotypes predict dyskinesia in Parkinson's disease. *Parkinsonism Relat. Disord.* **47**, 39–44 (2018).
21. Foltynie, T. et al. BDNF val66met influences time to onset of levodopa induced dyskinesia in Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry* **80**, 141–144 (2009).
22. Cheshire, P. et al. Influence of single nucleotide polymorphisms in COMT, MAO-A and BDNF genes on dyskinesias and levodopa use in Parkinson's disease. *Neurodegener. Dis.* **13**, 24–28 (2014).
23. Bialecka, M. et al. The association of functional catechol-O-methyltransferase haplotypes with risk of Parkinson's disease, levodopa treatment response, and complications. *Pharmacogenet. Genom.* **18**, 815–821 (2008).
24. König, E. et al. Exome-wide association study of levodopa-induced dyskinesia in Parkinson's disease. *Sci. Rep.* **11**, 19582 (2021).
25. Ku, S. & Glass, G. A. Age of Parkinson's disease onset as a predictor for the development of dyskinesia. *Mov. Disord.* **25**, 1177–1182 (2010).
26. Sharma, J. C., Bachmann, C. G. & Linazasoro, G. Classifying risk factors for dyskinesia in Parkinson's disease. *Parkinsonism Relat. Disord.* **16**, 490–497 (2010).
27. Warren Olanow, C. et al. Factors predictive of the development of Levodopa-induced dyskinesia and wearing-off in Parkinson's disease. *Mov. Disord.* **28**, 1064–1071 (2013).
28. Passarella, D. et al. Low-density lipoprotein receptor-related protein 8 at the crossroad between cancer and neurodegeneration. *Int. J. Mol. Sci.* **23**, 8921 (2022).
29. Hiesberger, T. et al. Direct binding of Reelin to VLDL receptor and ApoE receptor 2 induces tyrosine phosphorylation of disabled-1 and modulates tau phosphorylation. *Neuron* **24**, 481–489 (1999).
30. Deutschlander, A. B. et al. Association of MAPT subhaplotypes with clinical and demographic features in Parkinson's disease. *Ann. Clin. Transl. Neurol.* **7**, 1557–1563 (2020).
31. Schierding, W. et al. Common variants coregulate expression of GBA and modifier genes to delay Parkinson's disease onset. *Mov. Disord.* **35**, 1346–1356 (2020).
32. Szwedo, A. A. et al. GBA and APOE impact cognitive decline in Parkinson's Disease: a 10-year population-based study. *Mov. Disord.* **37**, 1016–1027 (2022).
33. Snaith, R. P., Bridge, G. W. & Hamilton, M. The Leeds scales for the self-assessment of anxiety and depression. *Br. J. Psychiatry* **128**, 156–165 (1976).
34. Spielberger, C. D., Gorsuch, R. L. & Lushene, R. E. *STAI Manual for the State-trait Anxiety Inventory ('Self-evaluation Questionnaire')*. (Consulting Psychologists Press, 1970).
35. Reddy, S. S., Connor, T. E., Weeber, E. J. & Rebeck, W. Similarities and differences in structure, expression, and functions of VLDLR and ApoER2. *Mol. Neurodegener.* **6**, 30 (2011).
36. Hirota, Y., Kubo, K.-I., Fujino, T., Yamamoto, T. T. & Nakajima, K. ApoER2 controls not only neuronal migration in the intermediate zone but also termination of migration in the developing cerebral cortex. *Cereb. Cortex* **28**, 223–235 (2016).
37. Zarouchlioti, C., Parfitt, D. A., Li, W., Gittings, L. M. & Cheetham, M. E. DNAJ Proteins in neurodegeneration: essential and protective factors. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **373**, 20160534 (2018).
38. Sun, A.-G. et al. Identifying distinct candidate genes for early Parkinson's disease by analysis of gene expression in whole blood. *Neuro Endocrinol. Lett.* **35**, 398–404 (2014).
39. Global Parkinson's Genetics Program. GP2: The Global Parkinson's Genetics Program. *Mov. Disord.* **36**, 842–851 (2021).
40. Goetz, C. G., Nutt, J. G. & Stebbins, G. T. The unified Dyskinesia rating scale: presentation and clinimetric profile. *Mov. Disord.* **23**, 2398–2403 (2008).
41. Malek, N. et al. Tracking Parkinson's: study design and baseline patient data. *J. Parkinsons. Dis.* **5**, 947 (2015).
42. Publications: Frances Mary Ashcroft. OPDC home. https://www.dpag.ox.ac.uk/opdc.
43. Parkinson Progression Marker Initiative. The Parkinson Progression Marker Initiative (PPMI). *Prog. Neurobiol.* **95**, 629–635 (2011).
44. Rosenthal, L. S. et al. The NINDS Parkinson's disease biomarkers program. *Mov. Disord.* **31**, 915–923 (2016).
45. Carroll, C. B. et al. Simvastatin as a neuroprotective treatment for Parkinson's disease (PD STAT): protocol for a double-blind, randomised, placebo-controlled futility study. *BMJ Open* **9**, e029740 (2019).
46. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
47. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
48. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
49. Input filtering. https://www.cog-genomics.org/plink/1.9/filter.
50. McCarthy Tools. https://www.well.ox.ac.uk/~wrayner/tools.
51. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
52. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190 (2010).
53. Tseng, G. C., Ghosh, D. & Feingold, E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.* **40**, 3785–3799 (2012).
54. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
55. Schilder, B. M., Humphrey, J. & Raj, T. echolocatoR: an automated end-to-end statistical and functional genomic fine-mapping pipeline. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btab658 (2021).
56. Wakefield, J. Bayes factors for genome-wide association studies: comparison with *P*-values. *Genet. Epidemiol.* **33**, 79–86 (2009).
57. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
58. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).
59. Weissbrod, O. et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).
60. Bycroft, C. et al. The UK biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
61. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
62. Nott, A. et al. Brain cell type-specific enhancer-promoter interactome maps and disease risk association. *Science* **366**, 1134 (2019).
63. Iwaki, H. et al. Accelerating medicines partnership: Parkinson's disease. genetic resource. *Mov. Disord.* **36**, 1795–1804 (2021).
64. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
65. de Klein, N. et al. Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases. *Nat. Genet.* **55**, 377–388 (2023).
66. Võsa, U. 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**, 1300 (2021).
67. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
68. Home. http://lidpd.eurac.edu/.

ACKNOWLEDGEMENTS

This research was funded in whole or in part by Aligning Science Across Parkinson's [Grant number: ASAP-000478] through the Michael J. Fox Foundation for Parkinson's Research (MJFF). For the purpose of open access, the author has applied a CC BY public copyright licence to all Author Accepted Manuscripts arising from this submission. This research was supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. The UCL Movement Disorders Centre is supported by the Edmond J. Safra Philanthropic Foundation. This work was supported in part by the Intramural Research Programme of the National Institute on Aging (NIA). Data used in the preparation of this article were obtained from the AMP-PD Knowledge Platform (https://www.amp-pd.org). AMP-PD is a public–private partnership managed by the FNIH and funded by Celgene, GSK, Michael J. Fox Foundation for Parkinson's Research, the National Institute of Neurological Disorders and Stroke, Pfizer, and Verily. Clinical data and biosamples used in preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI), and the Parkinson's Disease Biomarkers Programme (PDBP). PPMI is funded by the Michael J Fox Foundation for Parkinson's Research and funding partners, including: Abbvie, AcureX, Aligning Science Across Parkinson's, Amathus Therapeutics, Avid Radiopharmaceuticals, Bial Biotech, Biohaven, Biogen, BioLegend, Bristol-Myers Squibb, Calico Labs, Celgene, Cerevel, Coave, DaCapo Brainscience, 4D Pharma, Denali, Edmond J Safra Foundation, Eli Lilly, GE Healthcare, Genentech, GlaxoSmithKline, Golub Capital, Insitro, Janssen Neuroscience, Lundbeck, Merck, Meso Scale Discovery, Neurocrine Biosciences, Prevail Therapeutics, Roche, Sanofi Genzyme, Servier, Takeda, Teva, UCB, VanquaBio, Verily, Voyager Therapeutics, and Yumanity. PPMI was approved by the ethics committees at each participating site, and written informed consent was obtained from all participants prior to inclusion in the study. The Parkinson's Disease Biomarker Programme (PDBP) consortium is supported by the National Institute of Neurological Disorders and Stroke (NINDS) at

the National Institutes of Health. A full list of PDBP investigators can be found at https://pdbp.ninds.nih.gov/policy. The PDBP Investigators have not participated in reviewing the data analysis or content of the manuscript. PDBP was approved by the ethics committees at each participating site, and written informed consent was obtained from all participants prior to inclusion in the study. Both TPD and OPDC cohorts are primarily funded and supported by Parkinson's UK (https://www.parkinsons.org.uk/) and supported by the National Institute for Health and Care Research (NIHR) Clinical Research Network (CRN). The TPD study is also supported by NHS Greater Glasgow and Clyde. The OPDC cohort is also supported by the NIHR Oxford Biomedical Research Centre, based at the Oxford University Hospitals NHS Trust, and the University of Oxford. TPD has multi-centre research ethics approval from the West of Scotland Research Ethics Committee: IRAS 70980, MREC 11/AL/0163 (ClinicalTrials.gov, NCT02881099). OPDC has multi-centre research ethics approval from the South Central Oxford A Research Ethics Committee 16/SC/0108. PD-STAT is funded and supported by grants from the Cure Parkinson's Trust (https://cureparkinsons.org.uk/) and JP Moulton Charitable Foundation (https://www.perscitusllp.com/moulton-charity-trust/), co-ordinated by the Peninsula Clinical Trials Unit, University of Plymouth and sponsored by University Hospitals Plymouth NHS Trust. PD-STAT has Newcastle and North Tyneside 2 Research Ethics Committee approval, 12/10/2015, ref: 15/NE/0324.

AUTHOR CONTRIBUTIONS

H.R.M. and A.M.C. designed the study. H.R.M. supervised the study. D.G.G., M.T.M.H., Y.B.-S., M.A.L., J.H. and H.R.M. conceived and led the TPD and OPDC clinical cohorts, as well as performed data management and curation. A.M.C. performed all the quality control on each cohort and performed all analyses in the present manuscript. C.C. led the PD STAT study, as well as performed data management and curation. M.M.X.T. helped with the design of the quality control strategy. L.W. provided access to an harmonised version of the AMP-PD genetic data. M.A.L. helped with the statistical study design. M.S. helped to design the statistical and quality control methodology. H.I. provided the base code to run a stepwise regression approach for clinical features selection. T.F. reviewed the manuscript methodology and results. A.M.C. wrote the initial manuscript. All authors critically reviewed the manuscript.

COMPETING INTERESTS

H.R.M. reports paid consultancy from Roche. Research Grants from Parkinson's UK, Cure Parkinson's Trust, PSP Association, CBD Solutions, Drake Foundation, Medical Research Council (MRC), Michael J. Fox Foundation. Dr Morris is a co-applicant on a patent application related to C9ORF72 - Method for diagnosing a neurodegenerative disease (PCT/GB2012/052140). D.G.G. has received grants from Michael's Movers, the Neurosciences Foundation, and Parkinson's UK, and honoraria from AbbVie, BIAL Pharma, Britannia Pharmaceuticals, GE Healthcare, and consultancy fees from Acorda Therapeutics and the Glasgow Memory Clinic. M.T.M.H. 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. She also received payment for Advisory Board attendance/consultancy for Biogen, Roche, Sanofi, CuraSen Therapeutics, Evidera, Manus Neurodynamica, Lundbeck. Y.B.-S. has received grant funding from the MRC, NIHR, Parkinson's UK, NIH, and ESRC. C.C receives salary from University Hospitals Plymouth NHS Trust and National Institute of Health and Care Research. She has received advisory, consulting or lecture fees from AbbVie, Bial, Scient, Orkyn, Abidetex, UCB, Pfizer, EverPharma, Lundbeck, Global Kinetics, Kyowa Kirin, Britannia and Medscape, and research funding from Parkinson's UK, Edmond J Safra Foundation, National Institute of Health and Care Research and Cure Parkinson's. J.H. is supported by the UK Dementia Research Institute, which receives its funding from DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society, and Alzheimer's Research UK. He is also supported by the MRC, Wellcome Trust, Dolby Family Fund, National Institute for Health Research University, College London Hospitals Biomedical Research Centre. All other authors report no competing interests. M.A.L. received fees for advising on a secondary analysis of an RCT sponsored by North Bristol NHS trust that studied off-state Dyskinesia.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41531-023-00573-2.

Correspondence and requests for materials should be addressed to Alejandro Martinez-Carrasco or Huw R. Morris.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.