bioRxiv preprint doi: https://doi.org/10.1101/388165; this version posted March 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

TITLE

Expanding Parkinson's disease genetics: novel risk loci, genomic context, causal insights and heritable risk.

AUTHORS

Mike A. Nalls[1,2,CA]*, Cornelis Blauwendraat[1]*, Costanza L. Vallerga[3,4]*, Karl Heilbron[5]*, Sara Bandres-Ciga[1]*, Diana Chang[6]*, Manuela Tan[7], Demis A. Kia[7], Alastair J. Noyce[7,8], Angli Xue[3,4], Jose Bras[9,10], Emily Young[11], Rainer von Coelln[12], Javier Simón-Sánchez[13,14], Claudia Schulte[13,14], Manu Sharma[15], Lynne Krohn[16,17], Lasse Pihlstrom[18], Ari Siitonen[19,20], Hirotaka Iwaki[1,2,21], Hampton Leonard[1,2], Faraz Faghri[1,22], J. Raphael Gibbs[1], Dena G. Hernandez[1], Sonja W. Scholz[23,24], Juan A. Botia[7,25], Maria Martinez[26], Jean-Christophe Corvol[27], Suzanne Lesage[27], Joseph Jankovic[11], Lisa M. Shulman[11], The 23andMe Research Team[5], System Genomics of Parkinson's Disease (SGPD) Consortium, Margaret Sutherland[28], Pentti Tienari[29,30], Kari Majamaa[19,20], Mathias Toft[18,31], Ole A. Andreassen[32,33], Tushar Bangale[6], Alexis Brice[27], Jian Yang[3,4], Ziv Gan-Or[16,17,34], Thomas Gasser[13,14], Peter Heutink[13,14], Joshua M Shulman[11,35,36], Nicolas Wood[7], David A. Hinds[5], John A. Hardy[7], Huw R Morris[37,38], Jacob Gratten[3,4], Peter M. Visscher[3,4], Robert R. Graham[6], Andrew B. Singleton[1] for the International Parkinson's Disease Genomics Consortium.

AFFILIATIONS

[1] Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, 20892 USA
[2] Data Tecnica International, Glen Echo, MD, 20812 USA
[3] Queensland Brain Institute, The University of Queensland, Brisbane, QLD 4072 Australia
[4] Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072 Australia
[5] 23andMe, Inc., Mountain View, California 94041 USA
[6] Department of Human Genetics, Genentech, South San Francisco, 94080, CA, USA
[7] Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK
[8] Preventive Neurology Unit, Wolfson Institute of Preventive Medicine, Queen Mary University of London, London, UK
[9] Center for Neurodegenerative Science, Van Andel Research Institute, Grand Rapids, Michigan, USA
[10] Department of Neurodegenerative Diseases, UCL Institute of Neurology, University College London, London, UK
[11] Department of Neurology, Baylor College of Medicine, Houston, USA
[12] Department of Neurology, University of Maryland School of Medicine, Baltimore, MD, USA
[13] Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany
[14] German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany.

bioRxiv preprint doi: https://doi.org/10.1101/388165; this version posted March 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

[15] Centre for Genetic Epidemiology, Institute for Clinical Epidemiology and Applied Biometry, University of Tubingen, Germany
[16] Department of Human Genetics, McGill University, Montreal, Quebec, Canada
[17] Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada
[18] Department of Neurology, Oslo University Hospital, Oslo, Norway
[19] Institute of Clinical Medicine, Department of Neurology, University of Oulu, Oulu, Finland
[20] Department of Neurology and Medical Research Center, Oulu University Hospital, Oulu, Finland
[21] The Michael J. Fox Foundation, New York, New York, 10036 USA
[22] Department of Computer Science, University of Illinois Urbana-Champaign, Champaign, IL, 61820, USA
[23] Neurodegenerative Diseases Research Unit, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892, USA
[24] Department of Neurology, Johns Hopkins University Medical Center, Baltimore, MD 21287, USA
[25] Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, Spain
[26] INSERM UMR 1220; and Paul Sabatier University, Toulouse, France
[27] INSERM U1127, CNRS UMR 7225, Sorbonne Université UMR S1127, APHP, Institut du Cerveau et de la Moelle épinière, ICM, Paris F-75013, France
[28] National Institute on Neurological Diseases and Stroke, National Institutes of Health, Bethesda, MD 20892 USA
[29] Clinical Neurosciences, Neurology, University of Helsinki, Helsinki, Finland
[30] Helsinki University Hospital, Helsinki, Finland
[31] Institute of Clinical Medicine, University of Oslo, Oslo, Norway
[32] Jebsen Centre for Psychosis Research, University of Oslo, Oslo, Norway
[33] Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway
[34] Department of Neurology & Neurosurgery, McGill University, Montreal, Quebec, Canada
[35] Departments of Molecular & Human Genetics and Neuroscience, Baylor College of Medicine, Houston, USA
[36] Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston
[37] Department of Clinical Neuroscience, UCL Institute of Neurology, London UK
[38] UCL Movement Disorders Centre, UCL Institute of Neurology, London, UK
*denotes shared first authorship.
[CA]denotes corresponding author, mike[at]datatecnica[dot]com

Full consortia membership (PubMed indexed) is available in the supplemental materials (Text S1).

ACKNOWLEDGEMENTS AND FUNDING

See supplemental materials (Text S2).

ABSTRACT

bioRxiv preprint doi: https://doi.org/10.1101/388165; this version posted March 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

We performed the largest genome-wide association study of PD to date, involving the analysis of 7.8M SNPs in 37.7K cases, 18.6K UK Biobank proxy-cases, and 1.4M controls. We identified 90 independent genome-wide significant signals across 78 loci, including 38 independent risk signals in 37 novel loci. These variants explained 26-36% of the heritable risk of PD. Tests of causality within a Mendelian randomization framework identified putatively causal genes for 70 risk signals. Tissue expression enrichment analysis suggested that signatures of PD loci were heavily brain-enriched, consistent with specific neuronal cell types being implicated from single cell expression data. We found significant genetic correlations with brain volumes, smoking status, and educational attainment. In sum, these data provide the most comprehensive understanding of the genetic architecture of PD to date by revealing many additional PD risk loci, providing a biological context for these risk factors, and demonstrating that a considerable genetic component of this disease remains unidentified.

INTRODUCTION

Parkinson's disease (PD) is a neurodegenerative disorder, affecting up to 2% of the population older than 60 years, an estimated 1 million individuals in the United States alone. PD patients suffer from a combination of progressive motor and non-motor symptoms that increasingly impair daily function and quality of life. There are no treatments that delay or alter PD[1]. As the global population continues to age, the prevalence of PD is projected to double in some age groups by 2030, creating a substantial burden on healthcare systems.[1,2,3]

Early investigations into the role of genetic factors in PD focused on the identification of rare mutations underlying familial forms of the disease,[4–6] but over the past decade there has been a growing appreciation for the important contribution of genetics in sporadic disease[7,8]. Genetic studies of sporadic PD have altered the foundational view of disease etiology as much of sporadic disease was formerly thought to be environmental.

With this in mind, we executed a series of experiments to further explore the genetics of PD (summarized in Figure 1). We performed the largest-to-date GWAS for PD, including 7.8M SNPs, 37.7K cases, 18.6K UK Biobank (UKB) "proxy-cases" and 1.4M controls. We identified putatively causal genes for PD, providing valuable targets for therapeutic research. We assessed the function of these putatively causal genes on a larger scale than in previous studies of PD via Mendelian randomization (MR), expression enrichment, and protein-protein interaction network analysis [9,10,11]. We estimated PD heritability, developed a polygenic risk score that predicted a substantial proportion of this heritability, and leveraged these results to inform future studies of PD genetics. Finally, we identified putative PD biomarkers and risk factors using genetic correlation and Mendelian randomization.

METHODS

See Supplementary Methods

bioRxiv preprint doi: https://doi.org/10.1101/388165; this version posted March 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

RESULTS

Novel loci and multiple signals in known loci identified

To maximize our power for locus discovery we used a single stage design, meta-analyzing all available GWAS summary statistics. In support of this design, we found strong genetic correlations between GWAS using PD cases ascertained by clinicians compared to 23andMe self-reported cases (rG = 0.85, SE = 0.06) and UKB proxy cases (rG = 0.84, SE = 0.134).

We identified a total of 90 independent genome-wide significant association signals through our meta-analysis and conditional analyses of 37,688 cases, 18,618 UKB proxy-cases and 1,417,791 controls at 7,784,415 SNPs (Figure 2, Table 1, Supplementary Appendices, Table S1, Table S2). Of these, 38 signals are new and more than 1MB from loci described in a previous report by Chang *et al*. 2017 (Table S3).

In an attempt to detect multiple independent signals within loci we implemented conditional and joint analysis (GCTA-COJO, http://cnsgenomics.com/software/gcta/) with a large study-specific reference genotype series, as well as a participant-level conditional analysis using 23andMe data [12]. We considered independent risk signals from conditional analyses to share the same locus if they were within 250kb of each other. We detected 10 loci containing more than one independent risk signal (22 risk SNPs in total across these loci), of which nine had been identified by previous GWAS, including multi-signal loci in the vicinity of *GBA, NUCKS1 / RAB29, GAK / TMEM175, SNCA* and *LRRK2*. The novel multi-signal locus comprised independent risk variants rs2269906 (*UBTF* / *GRN*) and rs850738 (*FAM171A2*). Detailed summary statistics on all nominated loci can be found in Table S2.

Refining heritability estimates and determining extant genetic risk

To quantify how much of the genetic liability we have explained and what direction to take with future PD GWAS we calculated updated heritability estimates and polygenic risk scores (PRS). Using LD score regression (LDSC) on a meta-analysis of all 11 clinically-ascertained datasets from our GWAS and estimated the liability-scale narrow-sense heritability of PD as 0.22 (95% CI 0.18 - 0.26), only slightly lower than a previous estimate derived using GCTA (0.27, 95% CI 0.17 - 0.38)[10,13,14]. This may be because LDSC is known to be more conservative than GCTA, however, our LDSC heritability estimate does fall within the 95% confidence interval of the GCTA estimate.

Next, we sought to determine the proportion of SNP-based heritability explained by our PD GWAS results using polygenic risk scores (PRSs). We utilized a two-stage design for our PRS analyses, with variant selection and training in the NeuroX-dbGaP dataset (5,851 cases and 5,866 controls) and then validation in the Harvard Biomarker Study (HBS, 527 cases and 472 controls). We focused on the NeuroX-dbGaP and HBS cohort as both of these clinically characterized cohorts were genotyped on the same PD-focused array (NeuroX) and have been used in previous studies of PRSs [8,15–18]. In addition, both of these studies directly genotyped

bioRxiv preprint doi: https://doi.org/10.1101/388165; this version posted March 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

larger effect, rare variants within *LRRK2* (rs34637584, G2019S) and *GBA* (rs76763715, N370S) of great interest in previous PRS analyses.

In order to prevent bias, we estimated the effect size of each SNP contributing to the PRS using a meta-analysis of all PD GWAS datasets except NeuroX-dbGAP and HBS. Using permutation testing in the NeuroX-dbGAP training cohort, we found that the optimal *P* threshold for variant inclusion was 1.35E-03, which included 1809 variants. Two PRSs were tested in HBS, one limited to 88 of the 90 genome-wide significant variants (two variants failed to pass quality control in the HBS study), and the other incorporating 1805 variants from the training phase (four variants failed to pass quality control in HBS due to low imputation quality). The 88 variant PRS had an area under the curve (AUC) of 0.651 (95% CI 0.617 - 0.684), while the 1805 variant PRS had an AUC of 0.692 (95% CI 0.660 - 0.725). The AUCs from our 88 variant PRS in both the NeuroX-dbGAP cohort and the HBS cohort were significantly larger than the AUCs in those same cohorts using a published PRS (Chang *et al.* 2017, AUC = 0.624, P < 0.002 from DeLong's test). Although the HBS cohort was used to discover the 90 PD GWAS risk variants, therefore potentially biasing our 88 variant PRS, all 90 variants remained genome-wide significant in a meta-analysis of all GWAS datasets excluding the HBS study. Extended results for all included studies at all P-value thresholds can be found in the Supplementary Appendix.

Using equations from Wray *et al*. 2010 and our current heritability estimates, the 88 variant PRS explained approximately 16% of the genetic liability of PD assuming a global prevalence of 0.5%[13,19]. The 1805 variant PRS explained roughly 26% of PD heritability. In a high-risk population with a prevalence of 2%, the 1805 variant PRS explained 36% of PD heritable risk [13,19] (Table S4).

We then attempted to quantify strata of risk in our more inclusive PRS. Compared to individuals with PRS values in the lowest quartile , the PD odds ratio for individuals with PRS values in the highest quartile was 3.74 (95% CI = 3.35 - 4.18) in the NeuroX-dbGaP cohort and 6.25 (95% CI = 4.26 - 9.28) in the HBS cohort (Table 2, Figure 3, Figure S1).

Variants in the range of 5E-08 < P < 1.35E-03 (used in the 1805 variant PRS) were rarer and had smaller effect estimates than variants reaching genome-wide significance. These sub-significant variants had a median minor allele frequency of 21.3% and a median effect estimate (absolute value of the log odds ratio of the SNP parameter from regresion) of 0.047. Genome-wide significant risk variants were more common with a median minor allele frequency of 25.1%, and had a median effect estimate of 0.081. We performed power calculations to forecast the number of additional PD cases needed to achieve genome-wide significance at 80% power for a variant with a minor allele frequency of 21.3% and an effect estimate of 0.047 [20]. Assuming that all incoming data is well harmonized with current data and that disease prevalence is 0.5%, we estimated that we would need a total of ~99K cases, ~2.3 times as many as our current analysis. Variant discovery at this point will help us work towards the maximum achievable AUC for a genetic predictor in PD (estimated 85%). Past this point it is possible that effect estimates get too marginal, variants get too rare and they are no longer useful in predictions or in estimating heritability [21].

bioRxiv preprint doi: https://doi.org/10.1101/388165; this version posted March 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Functional causal inferences via Quantitative Trait Loci (QTL)

There were 305 genes within the 78 GWAS loci. We sought to identify the causal gene in each locus in order to help direct future high-throughput functional studies. Specifically, we used large QTL datasets and summary-data-based Mendelian randomization (SMR) to test whether the expression or methylation of these genes led to a causal change in PD risk (Table 3, Table S5, Table S6)[11]. This method allows for functional inferences between two datasets to be made within a similar framework as a randomized controlled trial, treating the genotype as the randomizing factor.

We used four QTL datasets: a large meta-analysis of mRNA expression across brain tissues, mRNA expression in the substantia nigra, mRNA expression in blood, and methylation in blood [22–26]. Of the 305 genes under linkage disequilibrium (LD) peaks around our risk variants of interest, 237 were possibly associated with at least one QTL and were therefore testable via SMR (Supplementary Methods, Table S6). The expression or methylation of 151 of these 237 genes (63.7%) was significantly linked to a causal change in PD risk.

Of the 90 PD GWAS risk variants, 70 were in loci containing at least one of these putatively causal genes after multiple test correction (Table 3 summarizes top QTL per gene). For 53 out of these 70 PD GWAS hits (75.7%), the gene nearest to the sentinel SNP was a putatively causal gene (Table S2). Most loci tested contained multiple putatively causal genes. Interestingly, the nearest putatively causal gene to the rs850738 / *FAM171A2* GWAS risk signal is *GRN*, a gene known to be associated with frontotemporal dementia (FTD)[27]. Mutations in this gene (*GRN*) have also been shown to be connected with another lysosomal storage disorder, neuronal ceroid lipofuscinosis[28].

Rare coding variant burden analysis

As an orthogonal approach for nominating putatively causal genes, we also carried out rare coding variant burden analyses. While the main GWAS analysis was limited to MAF ≥ 1% (except for known coding risk variants), we carried out rare variant burden analyses in a subset of studies (Supplementary Methods). We performed kernel-based burden tests on the 113 genes in our PD GWAS loci that contained two or more rare coding variants (MAF< 5% or MAF < 1%). After Bonferroni correction for 113 genes, we identified 7 significant putatively causal genes: *LRRK2*, *GBA*, *CATSPER3* (rs11950533/*C5orf24* locus), *LAMB2* (rs12497850/*IP6K2* locus), *LOC442028* (rs2042477/*KCNIP3* locus), *NFKB2* (rs10748818/*GBF1* locus), and *SCARB2* (rs6825004 locus). These results suggest that some of the risk associated with these loci may be due to rare coding variants. The *LRRK2* and *NFKB2* associations at MAF < 1% remained significant after correcting for all ~20,000 genes in the human genome (P = 2.15E-10 and P = 4.02E-07, Table S7, Table S5).

Tissue and cell specific expression enrichment plus protein-protein interactions

bioRxiv preprint doi: https://doi.org/10.1101/388165; this version posted March 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

In order to better understand the function of the genes highlighted by this study, we tested whether these genes were enriched in 10,651 biological pathways. We tested for gene expression and pathway enrichment in PD loci using Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) and webgestaltR, respectively [9,29]. We found 10 significantly enriched pathways (false discovery rate [FDR]-adjusted P < 0.05, Table S8), including four related to vacuolar function and three related to known drug targets (calcium transporters: ikeda_mir1_targets_dn and ikeda_mir30_targets_up, kinase signaling: kim_pten_targets_dn). Known pathways of interest relating to lysosomal function, endocytosis, and dopamine metabolism were significantly enriched when using a more lenient *P* value (FDR-adjusted *P* < 0.1). At least three candidate genes within novel loci are involved in lysosomal storage disorder (*GUSB*, *GRN,* and *NEU1)*, a pathway of interest in recent PD research [30].

Next, we sought to determine the tissues and cell types most relevant to PD etiology using FUMA[9]. We tested whether the genes highlighted by our PD GWAS were enriched for expression in 53 tissues from across the body. We found 13 significant tissues, all of which were brain-derived (Figure S2A), in contrast to what has been seen in Alzheimer's disease which shows a strong bias towards blood, spleen, lungs and microglial enrichments [31]. To further disentangle the enrichment in brain tissues, we tested whether our PD GWAS genes were enriched for expression in 88 brain cell types using single cell RNA sequencing reference data from DropViz (http://dropviz.org)[32]. After false discovery rate correction we found seven significant brain cell types, all of which were neuronal (Figure S2B). The strongest enrichment was for neurons in the substantia nigra (SN) at P = 1.0E-06, with additional significant results at P < 5.0E-4 for the globus pallidus (GP), thalamus (TH), posterior cortex (PC), frontal cortex (FC), hippocampus (HC) and entopeduncular nucleus (ENT).

Finally, we analyzed protein-protein interaction networks using webgestaltR[29] and found that the genes highlighted by our PD GWAS were enriched in six functional ontological networks (FDR-adjusted *P* < 0.1). The majority of these networks were related to chemical signaling pathways or response to some type of stressor. The most significant protein-protein interaction was related to response to interferon-gamma (Table S9, Figure S3A, Figure S3B).

Genetic correlations and Mendelian randomization across phenotypes

Next, we used cross-trait genetic correlation and Mendelian randomization to identify putative PD biomarkers and risk factors. We estimated the cross-trait genetic correlation between our PD GWAS and 757 other GWAS datasets curated by LD hub [33]. We found four significant genetic correlations (FDR-adjusted *P* < 0.05, Table 4, Table S10) including positive correlations with intracranial volume (rG = 0.351, SE = 0.077, P = 4.64E-06) and putamen volume (rG = 0.248, SE = 0.064, P = 9.55E-05, respectively)[34], and negative correlations with current tobacco use (rG = -0.134, SE = 0.034, P = 7.92E-05) and "academic qualifications: National Vocational Qualifications (NVQ) or Higher National Diploma (HND) or Higher National Certificate (HNC) or equivalent" (rG = -0.169, SE 0.045, P = 2.00E-04)[35]. The negative association with one's academic qualifications suggests that individuals without a college education may be at less risk

bioRxiv preprint doi: https://doi.org/10.1101/388165; this version posted March 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

of PD than individuals with higher levels of education. The correlation between PD and smoking status may not be independent from the correlation between PD and education as smoking status and years of education were significantly correlated (rG = -0.361, SE = 0.064, P = 1.64E-08) [36].

We used Mendelian randomization to assess whether there was evidence of a causal relationship between PD and five phenotypes related to academic qualification, smoking, and brain volumes described above (Figure S4). Cognitive performance had a large, significant causal effect on PD risk (MR effect = 0.213, SE = 0.041, Bonferroni-adjusted P = 8.00E-07), while PD risk did not have a significant causal effect on cognitive performance (Bonferroni-adjusted P = 0.125). Educational attainment also had a significant causal effect on PD risk (MR effect = 0.162, SE = 0.040, Bonferroni-adjusted P = 2.06E-04), but PD risk also had a weak but significant causal effect on educational attainment (MR effect = 0.007, SE = 0.002, Bonferroni-adjusted P = 7.45E-3). There was no significant causal relationship between PD and current smoking status (forward analysis: MR effect = -0.069, SE = 0.031, Bonferroni-adjusted P = 0.125; reverse analysis: MR effect = 0.004, SE = 0.010, Bonferroni-adjusted P = 1). Smoking initiation (the act of ever starting smoking) did not have a causal effect on PD risk (MR effect = -0.063, SE = 0.034, Bonferroni-adjusted P = 0.315), whereas PD had a small, but significantly positive causal effect on smoking initiation (MR effect = 0.027, SE = 0.006, Bonferroni-adjusted P = 1.62E-05). Intracranial volume could not be tested because its GWAS did not contain any genome-wide significant risk variants. There was no significant causal relationship between PD and putamen volume (P > 0.05 in both the forward and reverse directions).

DISCUSSION

Our work marks a significant step forward in our understanding of the genetic architecture of PD and provides a genetic reference set for the broader research community. We identified 90 independent common genetic risk factors for PD, nearly doubling the number of known PD risk variants. We re-evaluated the cumulative contribution of genetic risk variants, both genome-wide significant and not-yet discovered, in order to refine our estimates of heritable Parkinson's disease risk. We also nominated likely causal genes at each locus for further follow-up using QTL analyses and rare variant burden analyses. Our work has highlighted the pathways, tissues, and cell types involved in PD etiology. Finally, we identified intracranial and putaminal volume as potential PD biomarkers, and cognitive performance as a PD risk factor. Altogether, the data presented here has significantly expanded the resources available for future investigations into potential PD interventions.

Using a PRS constructed from our GWAS results, we were able to explain up to 36% of PD heritability. Power estimates suggest that expansions of case numbers to 99K cases will continue to reveal additional insights into PD genetics. While these yet-to-be defined risk variants will have relatively small effects, cumulatively they will improve our ability to predict PD and will help to further expand our knowledge of the genes and pathways that drive PD risk.

bioRxiv preprint doi: https://doi.org/10.1101/388165; this version posted March 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Population-wide screening for individuals who are likely to develop PD is currently not feasible using our 1805 variant PRS alone. There would be roughly 14 false positives per true positive assuming a prevalence of 0.5%. While large-scale genome sequencing and non-linear machine learning methods will likely improve these predictive models, we have previously shown that we will need to incorporate other data sources (*e.g.* smell tests, family history, age, sex) in order to generate algorithms that are useful for population-wide screening [18].

Evaluating these results in the larger context of pathway, tissue, and cellular functionality revealed that genes near PD risk variants showed enrichment for expression in the brain, contrasting with previous work in Alzheimer's disease. Notably, we showed that the expression enrichment of genes at PD loci occured exclusively in neuronal cell types. We also found that PD genes were enriched in chemical signaling pathways and pathways involving the response to a stressor. These observations may be informative for disease modeling efforts, highlighting the importance of disease modeling in neurons and of incorporating a cellular stress component. This will help inform and focus stem cell derived therapeutic development efforts that are currently underway.

Using cross-trait LD score regression, we found four phenotypes that were genetically correlated with PD. Putamen and intracranial volumes may prove to be valuable PD biomarkers. Our bi-directional GSMR results suggest a complex etiological connection between smoking initiation and PD that will require further follow-up and should be viewed with some caution. One of the implications of this work is that PD trials of nicotine or other smoking-related compound(s) may be less likely to succeed. The strong causal effect of cognitive performance on PD is supported by observational studies [37].

While this study marks major progress in assessing genetic risk factors for PD, there remains a great deal to be done. No defined external validation dataset was used, which may be seen as a limitation. Simulations have suggested that without replication variants with *P* values between 5E-08 and 5E-9 should be interpreted with greater caution [38,39]. We found 16 risk variants in this *P* value range, including two known variants near *WNT3* (proximal to the *MAPT* locus) and *BIN3*. To a degree, the fact that we filtered our variants with a secondary random-effects meta-analysis may make our 90 PD GWAS hits somewhat more robust due to the conservative nature of random-effects. Secondly, this study focused on PD risk in individuals of European ancestry. Adding datasets from non-European populations would be helpful to further improve our granularity in association testing and ability to fine-map loci through integration of more variable LD signatures while also evaluating population specific associations. Additionally, large ancestry-specific PD LD reference panels, such as those for Ashkenazi Jewish patients, will help us further unravel the genetic architecture of loci such as GBA and LRRK2. This may be particularly crucial at these loci where LD patterns may be quite variable within European populations, accentuating the possible influence of LD reference series on conditional analyses in some cases [40]. Finally, our work utilized state-of-the-art QTL datasets to nominate candidate genes, but many QTL associations are hampered by both small sample size and low *cis*-SNP density. Larger QTL studies and PD-specific network data from large scale cellular screens would allow us to build a more robust functional inference framework.

bioRxiv preprint doi: https://doi.org/10.1101/388165; this version posted March 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

As the field moves forward there are some critical next steps that should be prioritized. First, allowing researchers to share participant-level data in a secure environment would facilitate inclusiveness and uniformity in analyses while maintaining the confidentiality of study participants. Our work suggests that GWASes including up to 99,000 cases will continue to provide useful biological insights into PD. In addition to studies of the genetics of PD risk, studies of disease onset, progression, and subtype will be important and will require large series of well-characterized patients. We also believe that work across diverse populations is important, not only to be able to best serve these populations but also to aid in fine mapping of loci. Notably, the use of genome sequencing technologies could further improve discovery by capturing rare variants and structural variants, but with the caveat that very large samples sizes will be required. While there is still much left to do, we believe that our current work represents a significant step forward and that the results and data will serve as a foundational resource for the community to pursue this next phase of PD research.

WORKS CITED

1. Gasser, T. Genetics of parkinson's disease. *Ann. Neurol.* **44**, S53–S57 (1998).
2. Dorsey, E. R. *et al.* Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. *Neurology* **68**, 384–386 (2007).
3. Dorsey, E. R., Ray Dorsey, E. & Bloem, B. R. The Parkinson Pandemic—A Call to Action. *JAMA Neurol.* **75**, 9 (2018).
4. Polymeropoulos, M. H. *et al.* Mapping of a gene for Parkinson's disease to chromosome 4q21-q23. *Science* **274**, 1197–1199 (1996).
5. Singleton, A. B. *et al.* alpha-Synuclein locus triplication causes Parkinson's disease. *Science* **302**, 841 (2003).
6. Parkinson, J. An Essay on the Shaking Palsy. *J. Neuropsychiatr.* **14**, 223–236 (2002).
7. Fung, H.-C. *et al.* Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol.* **5**, 911–916 (2006).
8. Chang, D. *et al.* A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat. Genet.* **49**, 1511–1516 (2017).
9. Wang, J., Vasaikar, S., Shi, Z., Greer, M. & Zhang, B. WebGestalt 2017: a more

bioRxiv preprint doi: https://doi.org/10.1101/388165; this version posted March 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res.* **45**, W130–W137 (2017).

10. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
11. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
12. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–75, S1–3 (2012).
13. Keller, M. F. *et al.* Using genome-wide complex trait analysis to quantify 'missing heritability' in Parkinson's disease. *Hum. Mol. Genet.* **21**, 4996–5009 (2012).
14. 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).
15. Nalls, M. A. *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.* **46**, 989–993 (2014).
16. Nalls, M. A. *et al.* NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases. *Neurobiol. Aging* **36**, 1605.e7–12 (2015).
17. Nalls, M. A. *et al.* Genetic risk and age in Parkinson's disease: Continuum not stratum. *Mov. Disord.* **30**, 850–854 (2015).
18. Nalls, M. A. *et al.* Diagnosis of Parkinson's disease on the basis of clinical and genetic classification: a population-based modelling study. *Lancet Neurol.* **14**, 1002–1009 (2015).
19. Wray, N. R., Yang, J., Goddard, M. E. & Visscher, P. M. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet.* **6**, e1000864 (2010).
20. Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38**, 209–213 (2006).

bioRxiv preprint doi: https://doi.org/10.1101/388165; this version posted March 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

21. Kim, H., Grueneberg, A., Vazquez, A. I., Hsu, S. & de los Campos, G. Will Big Data Close the Missing Heritability Gap? *Genetics* **207**, 1135 (2017).
22. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
23. Qi, T. *et al.* Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat. Commun.* **9**, 2282 (2018).
24. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
25. Wu, Y. *et al.* Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat. Commun.* **9**, 918 (2018).
26. Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv* 447367 (2018). doi:10.1101/447367
27. Cruts, M. *et al.* Null mutations in progranulin cause ubiquitin-positive frontotemporal dementia linked to chromosome 17q21. *Nature* **442**, 920–924 (2006).
28. Smith KR, E. al. Strikingly different clinicopathological phenotypes determined by progranulin-mutation dosage. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/22608501. (Accessed: 10th January 2019)
29. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
30. Robak, L. A. *et al.* Excessive burden of lysosomal storage disorder gene variants in Parkinson's disease. *Brain* **140**, 3191–3203 (2017).
31. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* (2019). doi:10.1038/s41588-018-0311-9
32. Saunders A, E. al. Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/30096299. (Accessed: 31st October 2018)

bioRxiv preprint doi: https://doi.org/10.1101/388165; this version posted March 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

33. LD Hub. Available at: http://ldsc.broadinstitute.org. (Accessed: 26th June 2018)

34. Hibar, D. P. *et al.* Common genetic variants influence human subcortical brain structures. *Nature* **520**, 224–229 (2015).

35. Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK Biobank. *Neale lab* Available at: http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank. (Accessed: 24th June 2018)

36. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

37. Valdés, E. G. *et al.* Occupational complexity and risk of Parkinson's disease. *PLoS One* **9**, e106676 (2014).

38. Wu, Y., Zheng, Z., Visscher, P. M. & Yang, J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. *Genome Biol.* **18**, 86 (2017).

39. Pulit, S. L., de With, S. A. J. & de Bakker, P. I. W. Resetting the bar: Statistical significance in whole-genome sequencing-based association studies of global populations. *Genet. Epidemiol.* **41**, 145–151 (2017).

40. Rivas, M. A. *et al.* Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population. *PLoS Genet.* **14**, e1007329 (2018).

Table 1: Novel loci associated with Parkinson's disease.
Summary statistics for all 90 genome-wide significant PD variants. Columns include single nucleotide polymorphism ID (SNP), chromosome (CHR), base pair position (BP), nearest gene annotation for the variant, effect allele designation and frequency, as well as metrics for the odds ratio (OR), regression coefficient (beta), and standard error of the beta for the SNP from fixed-effects meta-analysis as well as the index of heterogeneity (I2). We also include four p-values from: fixed-effects meta-analyses, random-effects meta-analyses, standard conditional analyses in 23andMe, and a conditional joint analysis approach (COJO).

Table 2: Summary of genetic predictive model performance.
These are estimates of performance for predictive models including single study estimates, estimates from meta-analyses across studies, as well as a two stage design. Here the best P

bioRxiv preprint doi: https://doi.org/10.1101/388165; this version posted March 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

threshold column denotes the filtering value for SNP inclusion to achieve the maximal pseudo (Nagelkerke's) R2. The odds ration (OR) colum is the exponent of the regression coefficient (beta) from logistic regression of the polygenic risk score (PRS) on case status, with the standard error (SE) representing the precision of these estimates. These same metrics are derived across array types and datasets using random-effects meta-analyses. The area under the curve (AUC) is included as the most common metric for predictive model performance. In the table, * denotes R2 approximation adjusted for an estimated prevalence of 0.5%, equivalent to roughly half of the unadjusted R2 estimates for the PRS. All calculations and reported statistics include only the PRS and no other parameters after adjusting for principal components 1-5, age and sex at variant selection in the NeuroX-dbGaP dataset.

Table 3: Summary of significant functional inferences from QTL associations via Mendelian randomization for nominated genes of interest.
Multi-SNP eQTL Mendelian randomization results focusing only on the most significant association per nearest genes to PD risk loci after Bonferroni correction. If a locus was significantly associated with both brain and blood QTLs after multiple test correction, we opted to show the most signficant brain tissue derived association here after filtering for possible polygenicity (HEIDI P > 0.01). All tested QTL summary statistics can be found in Supplementary Table S6. Effect estimates represent the change in PD odds ratio per one standard deviation increase in gene expression or methylation.

Table 4: Significant cross-trait genetic correlations.
The genetic correlations between PD and four significantly-associated traits from LD Hub. An extended version of this table included in the Supplementary Materials (Supplementary Table S10) showing data for all tested correlations. In this table, h2 represents the heritability estimate of the trait.

FIGURES

Figure 1: Workflow and rationale summary.
This figure describes study design and rationale behind the analyses included in this report.

Figure 2: Manhattan plot.
The nearest gene to each of the 90 significant variants are labeled in green for previously-identified loci and in blue for novel loci. –log10P values were capped at 40. Variant points are color coded red and orange, with orange representing significant variants at P 5E-08 and 5E-9 and red representing significant variants at P < 5E-9. The X axis represents the base pair position of variants from smallest to largest per chromosome (1-22).

Figure 3: Predictive model details.
A. The odds ratio of developing PD for each quartile of polygenic risk score (PRS) compared to the lowest quartile of genetic risk. B. PRS receiver-operator curves for each array type and sampling design.

bioRxiv preprint doi: https://doi.org/10.1101/388165; this version posted March 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

SUPPLEMENTARY MATERIALS

Supplementary Methods: Detailed methods section.

Supplemental Appendix: This appendix is split into four sections detailing: first comparisons of effect estimates across GWAS cohorts (beta~beta plots), second forest plots for each significant variant, thirdly locus plots showing regional GWAS results, and QTL and burden associations for each variant, finally the fourth section including extended PRS results. Beta~beta plots compare the regression coefficients for up to 90 of the significant variants in one study to a meta-analysis of all others via linear regression. Forest plots communicate similar sensitivity analyses, for each of the 92 variants of interest. In the forest plots, box size indicates relative sample size for that study, and the width of the diamond representing the meta-analysis effect estimates indicate the 95% confidence interval. The locus plots are a zoomed-in version of Figure 2 for each of the 90 significant variants. These plots are truncated at a -log10 P value of 50 for display purposes and include the most significant burden test and QTL analysis results per gene denoted by label color-coding in each figure. In each locus plot, R2 is measured in our in-house LD reference dataset and shows the correlation between the most significant local SNP and all other proximal SNPs. Additional detailed PRS results for a subset of cohorts are available in the appendix summarizing PRS estimates at varied P thresholds. Each cohort specific PRS in the appendix is based on meta-analyses excluding that cohort when calculating SNP weights. A smaller table summarizing PRS associations at the P threshold with the highest r^2 is also included. Column headers in the PRS section of the appendix mirror that of Table 2.

Figure S1: The odds ratio of developing PD for each decile of PRS, comparing each decile to all others for all samples in this analysis.

Figure S2: Results of FUMA analysis for tissue and cell type specific expression enrichment. A. Tissue enrichment. B. Cell type-specific enrichment. Red bars indicate levels of significance surpassing multiple test correction.

Figure S3: Panel A: Gene ontology term connectivity within protein-protein networks. This panel shows network of gene ontology (GO) terms from pathway analyses. Most significant GO terms are shown in green. Panel B: Gene level connectivity within protein-protein networks. This panel shows connectivity between genes across enriched pathways.

Figure S4: Comparison of regression coefficients in Mendelian randomization analyses across traits. Each cross represents a SNP, with the dashed lines representing the trend across all variants. Axes position are regression coefficients from GWAS for significant SNPs from either GWAS. Panel A includes results for cognitive performance, panel B includes results for educational attainment, panel C includes results for putamen volume, panel D includes results for smoking initiation and panel E includes results for current smoking status.

Table S1: Descriptive statistics and quality control summaries for meta-analyzed genome-wide association studies. ! denotes age at exam for both cases and controls. $ denotes age at death,

bioRxiv preprint doi: https://doi.org/10.1101/388165; this version posted March 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

onset not available. * based on 599 PD cases and 715 controls. ^ denotes samples checked for overlap across datasets as per Nalls et al. 2014 and Chang et al. 2017, ^^ denotes checked for overlap within IPDGC sample series, ^^^denotes a combination of both workflows for identifying sample overlap.

Table S2: Summary statistics for all nominated risk variants, known and novel. For binary variables, 0 = negative and 1 = positive. Some specific notes include: delineations of all studies, new studies and previous studies as discussed in the methods section. Betas and standard errors (StdErr) refer to effect estimates per SNP from logistic regression or fixed-effects meta-analyses.I2 is the index of heterogeneity. QTL Nominated Gene = genes which represent the nearest cis-QTL for that locus significant in MR.

Table S3: Comparison with novel results from Chang et al., 2017. This table summarizes linkage disequilibrium estimates between Chang et al., 2017 novel loci and variants passing quality control in this report.

Table S4: Estimates of genetic liability explained in different scenarios.
Here we compare how different AUC estimates and prevalence rates change the amount of genetic liability explained by GWAS.

Table S5: SNPs of interest tagging genes for functional inferences and networks analysis. Nominated genes and SNPs for follow-up analyses based on minimum $r^2 > 0.5$ within +/- 1MB of one of our 90 risk loci.

Table S6: Complete summary statistics for QTL Mendelian randomization. Output from the SMR package for all QTLs of interest. Additional columns include QTL reference dataset, dataset-level Bonferroni corrected P values and a binary indicator if a candidate association passed multiple test correction. All columns prefixed by SMR indicate multi-SNP SMR results.

Table S7: Rare coding variant burden analyses for genes under GWAS peaks. Detailed results of burden tests for genes proximal to risk loci. This includes variant counts, test statistics (rho, q, P, adjusted P) for each gene of interest.

Table S8: FUMA expression pathway enrichment analysis results.
Pathway enrichment from collapsed GWAS summary statistics.

Table S9: Protein network analysis for linked genes under association peaks. Gene ontology terms passing false discovery rate adjustment.

Table S10: Bivariate LDscores. Default output from LD Hub. Abbreviations defined in main text and methods section.

Text S1: Authors and affiliations.

bioRxiv preprint doi: https://doi.org/10.1101/388165; this version posted March 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Text S2: Acknowledgements and Funding.

AUTHOR CONTRIBUTIONS

Study level analysis

MAN, CB, CLV, KH, SB-C, DC, MT, DK, LR, JS-S, LK, LP, ABS

Additional analysis and data management

MAN, CB, SB-C, AJN, AX, JY, JG, PMV, ABS

Design and funding

MAN, CB, CLV, KH, SB-C, LP, MS, KM, MT, AB, JY, ZG-O, TG, PH, JMS, NW, DAH, JH, HRM, JG, PMV, RRG, ABS

Critical review and writing the manuscript

MAN, CB, CLV, KH, SB-C, DC, MT, DAK, AJN, AX, JB, EY, RvC, JS-S, CS, MS, LK, LP, AS, HI, HL, FF, JRG, DGH, SWS, JAB, MM, J-CC, SL, JJ, LMS, MS, PT, KM, MT, AB, JY, ZG-O, TG, PH, JMS, NW, DAH, JH, HRM, JG, PMV, RRG, ABS

DATA ACCESS

GWAS summary statistics for 23andMe datasets (post-Chang and data included in Chang et al. 2017 and Nalls et al. 2014) will be made available through 23andMe to qualified researchers under an agreement with 23andMe that protects the privacy of the 23andMe participants. Please visit http://research.23andme.com/collaborate/#publication for more information and to apply to access the data. An immediately accessible version of the summary statistics is available here https://drive.google.com/file/d/1FZ9UL99LAqyWnyNBxxlx6qOUlfAnublN/view?usp=sharing excluding Nalls et al. 2014, 23andMe post-Chang et al. 2017 and Web-Based Study of Parkinson's Disease (PDWBS) but including all analyzed SNPs. After applying with 23andMe, the full summary statistics including all analyzed SNPs and samples in this GWAS meta-analysis will be accessible to the approved researcher(s). Underlying participant level IPDGC data is available to potential collaborators, please contact ipdgc.contact@gmail.com.

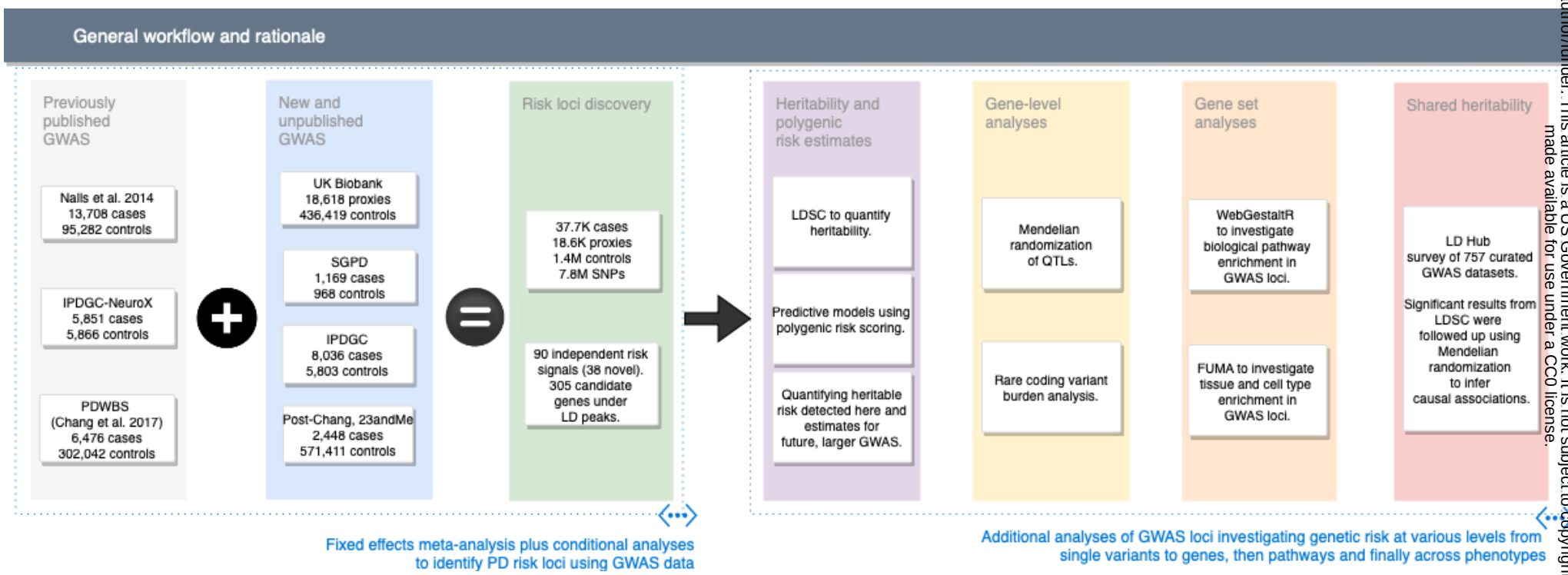
SNP	CHR	BP	Nearest Gene	Effect allele	Other allele	Effect allele frequency	OR	Beta	SE	P, fixed-effects	P, COJO	P, conditional	P, random-effects	I2, %
rs6658353	1	161469054	FCGR2A	c	g	0.501	1.07	0.065	0.009	6.10E-12	4.69E-12	1.38E-05	3.71E-05	40.2
rs11578699	1	171719769	VAMP4	t	c	0.195	0.93	-0.070	0.012	4.47E-09	4.45E-09	2.63E-03	1.09E-07	5.1
rs76116224	2	18147848	KCNS3	a	t	0.904	1.12	0.110	0.019	1.27E-08	1.27E-08	3.75E-07	1.27E-08	0
rs2042477	2	96000943	KCNIP3	a	t	0.242	0.94	-0.066	0.012	1.38E-08	1.48E-08	3.49E-05	1.38E-08	0
rs6808178	3	28705690	LINC00693	t	c	0.379	1.07	0.066	0.010	8.09E-12	7.18E-12	8.84E-05	8.09E-12	0
rs55961674	3	122196892	KPNA1	t	c	0.172	1.09	0.086	0.013	9.98E-12	8.30E-12	2.80E-06	9.98E-12	0
rs11707416	3	151108965	MED12L	a	t	0.367	0.94	-0.063	0.010	1.13E-10	1.02E-10	2.66E-04	1.77E-07	10.9
rs1450522	3	161077630	SPTSSB	a	g	0.674	0.94	-0.062	0.010	5.01E-10	4.90E-10	3.51E-04	2.27E-05	24.6
rs34025766	4	17968811	LCORL	a	t	0.159	0.92	-0.084	0.013	2.87E-10	2.82E-10	7.43E-06	2.87E-10	0
rs62333164	4	170583157	CLCN3	a	g	0.326	0.94	-0.064	0.010	2.00E-10	1.77E-10	5.10E-05	2.17E-05	21.3
rs26431	5	102365794	PAM	c	g	0.703	1.06	0.062	0.010	1.57E-09	1.65E-09	6.00E-03	2.36E-07	7.9
rs11950533	5	134199105	C5orf24	a	c	0.102	0.91	-0.092	0.016	7.16E-09	6.73E-09	5.08E-04	2.68E-08	1.9
rs9261484	6	30108683	TRIM40	t	c	0.245	0.94	-0.064	0.011	1.62E-08	1.43E-08	1.26E-06	1.62E-08	0
rs12528068	6	72487762	RIMS1	t	c	0.284	1.07	0.066	0.010	1.63E-10	1.79E-10	9.80E-06	1.63E-10	0
rs997368	6	112243291	FYN	a	g	0.805	1.07	0.071	0.012	1.84E-09	1.97E-09	2.61E-05	1.84E-09	0
rs75859381	6	133210361	RPS12	t	c	0.967	0.80	-0.221	0.034	1.04E-10	9.67E-11	1.09E-06	1.04E-10	0
rs76949143	7	66009851	GS1-124K5.11	a	t	0.051	0.87	-0.143	0.025	1.43E-08	1.51E-08	5.47E-09	2.04E-06	12.3
rs2086641	8	130901909	FAM49B	t	c	0.723	0.94	-0.061	0.011	1.81E-08	1.57E-08	6.07E-06	1.81E-08	0
rs6476434	9	34046391	UBAP2	t	c	0.734	0.94	-0.062	0.011	6.58E-09	6.56E-09	2.74E-04	6.58E-09	0
rs10748818	10	104015279	GBF1	a	g	0.851	0.92	-0.079	0.013	1.05E-09	1.23E-09	7.47E-06	1.05E-09	0
rs7938782	11	10558777	RNF141	a	g	0.878	1.09	0.087	0.015	2.12E-09	1.97E-09	2.17E-07	2.12E-09	0
rs7134559	12	46419086	SCAF11	t	c	0.404	0.95	-0.054	0.010	3.96E-08	3.80E-08	1.69E-02	1.84E-05	25.2
rs11610045	12	133063768	FBRSL1	a	g	0.490	1.06	0.060	0.009	1.77E-10	1.62E-10	3.57E-05	8.79E-07	19.5
rs9568188	13	49927732	CAB39L	t	c	0.740	1.06	0.062	0.011	1.15E-08	1.11E-08	4.29E-06	2.46E-04	21.4
rs4771268	13	97865021	MBNL2	t	c	0.230	1.07	0.068	0.011	1.45E-09	1.67E-09	1.41E-04	1.45E-09	0
rs12147950	14	37989270	MIPOL1	t	c	0.438	0.95	-0.053	0.010	3.54E-08	3.58E-08	1.06E-03	3.54E-08	0
rs3742785	14	75373034	RPS6KL1	a	c	0.787	1.07	0.071	0.012	1.92E-09	2.08E-09	2.22E-06	8.18E-06	24.8
rs2904880	16	28944396	CD19	c	g	0.309	0.94	-0.065	0.011	7.87E-10	8.68E-10	1.39E-05	7.87E-10	0
rs6500328	16	50736656	NOD2	a	g	0.599	1.06	0.059	0.010	1.82E-09	1.53E-09	1.43E-03	1.82E-09	0
rs12600861	17	7355621	CHRNB1	a	c	0.648	0.95	-0.057	0.010	1.01E-08	1.15E-08	5.10E-03	1.01E-08	0
rs2269906	17	42294337	UBTF	a	c	0.653	1.07	0.063	0.010	6.24E-10	8.63E-09	1.17E-05	6.24E-10	0
rs850738	17	42434630	FAM171A2	a	g	0.606	0.93	-0.071	0.011	1.29E-11	3.55E-10	4.18E-04	2.17E-07	17
rs61169879	17	59917366	BRIP1	t	c	0.164	1.09	0.082	0.013	9.28E-10	9.40E-10	9.07E-07	6.21E-06	16.4
rs666463	17	76425480	DNAH17	a	t	0.833	1.08	0.076	0.013	3.20E-09	2.90E-09	1.62E-05	4.17E-04	41
rs1941685	18	31304318	ASXL3	t	g	0.498	1.05	0.053	0.009	1.69E-08	1.61E-08	1.64E-08	1.69E-08	0
rs8087969	18	48683589	MEX3C	t	g	0.550	0.94	-0.058	0.010	1.41E-08	1.46E-08	1.09E-04	1.41E-08	0
rs77351827	20	6006041	CRLS1	t	c	0.128	1.08	0.080	0.014	8.87E-09	7.94E-09	1.84E-05	4.38E-07	11.2
rs2248244	21	38852361	DYRK1A	a	g	0.283	1.07	0.071	0.011	2.74E-11	2.51E-11	6.31E-05	8.78E-06	34.3

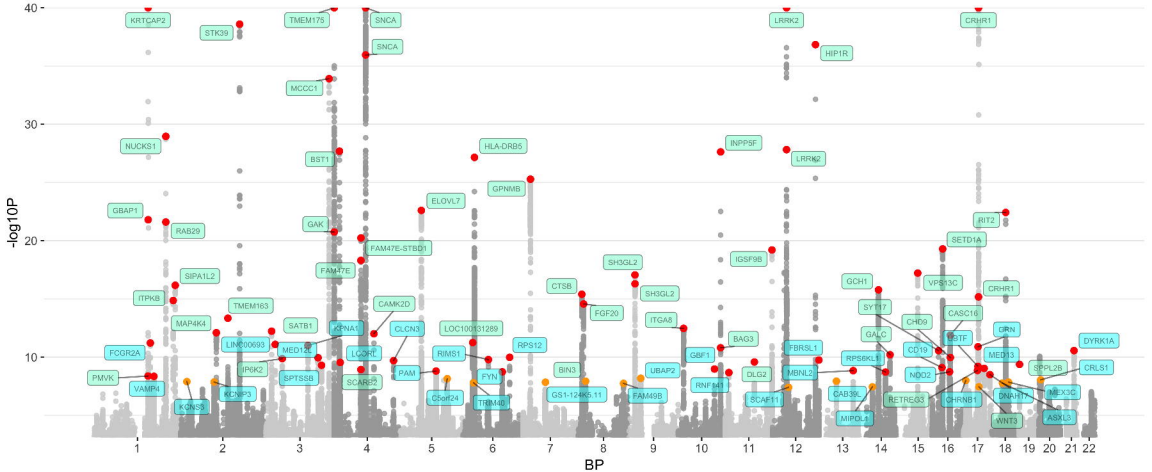
Study	Max P threshold	pseudo R2 from PRS*	OR	Beta	SE	P	OR, highest quartile PRS	95% CI, highest quartile PRS	N SNPs	N samples	AUC	95% CI (DeLong)	Sensitivity	Specificity	Positive predictive value (PPV)	Negative predictive value (NPV)	Balanced accuracy
Training dataset: IPDGC - Neurox	1.35E-03	0.029	1.74	0.553	0.022	8.99E-135	3.74	3.35 - 4.18	1809	11,243	0.640	0.630 - 0.650	0.569	0.632	0.591	0.611	0.601
Test dataset: HBS	4.00E-02	0.054	2.03	0.709	0.072	8.28E-23	6.25	4.26 - 9.28	1805	999	0.692	0.660 - 0.725	0.628	0.686	0.691	0.623	0.657

Gene	Probe	CHR	Probe, BP	Top SNP, BP	Top SNP	N SNPs	QTL reference	Effect	SE	Odds ratio	P	Bonferroni adjusted P
VAMP4	ENSG00000117533	1	171,690,343	171,717,417	rs10913587	98	Võsa et al. 2018 - blood expression	-0.272	0.05	0.762	5.67E-07	1.19E-04
KCNIP3	ENSG00000115041	2	96,007,438	95,989,766	rs3772034	14	Qi et al. 2018 - brain expression	-0.161	0.04	0.851	1.12E-05	1.15E-03
MAP4K4	ENSG00000071054	2	102,410,880	102,338,377	rs6733355	3	Võsa et al. 2018 - blood expression	1.119	0.24	3.063	2.32E-06	4.87E-04
TMEM163	ENSG00000152128	2	135,344,950	135,248,544	rs598668	28	Qi et al. 2018 - brain expression	0.074	0.02	1.077	3.55E-07	3.65E-05
KPNA1	ENSG00000114030	3	122,187,294	122,201,610	rs73190142	110	Võsa et al. 2018 - blood expression	0.310	0.05	1.363	1.56E-06	3.28E-04
GAK	ENSG00000178950	4	884,612	906,131	rs11248057	1	Qi et al. 2018 - brain expression	0.508	0.10	1.663	7.47E-07	7.69E-05
CAMK2D	ENSG00000145349	4	114,527,635	114,730,260	rs115671064	146	Võsa et al. 2018 - blood expression	-0.006	0.05	0.994	5.74E-06	1.21E-03
PAM	ENSG00000145730	5	102,228,247	102,118,633	rs2432162	679	Võsa et al. 2018 - blood expression	-0.031	0.01	0.970	2.08E-06	4.36E-04
LOC100131289	cg21339923	6	27,636,378	27,636,378	rs78149975	2	Qi et al. 2018 - brain methylation	-0.094	0.02	0.911	1.53E-06	3.06E-04
TRIM40	cg01641092	6	30,094,300	30,094,315	rs9261443	8	Qi et al. 2018 - brain methylation	0.072	0.01	1.075	6.15E-06	1.23E-03
HLA-DRB5	cg26036029	6	32,552,443	32,570,311	rs34039593	8	Qi et al. 2018 - brain methylation	-0.153	0.02	0.858	7.53E-10	1.51E-07
GPNMB	ENSG00000136235	7	23,295,156	23,294,668	rs858274	74	Qi et al. 2018 - brain expression	0.090	0.01	1.094	2.73E-21	2.81E-19
CTSB	ENSG00000164733	8	11,713,495	11,699,279	rs4631423	33	Qi et al. 2018 - brain expression	-0.150	0.04	0.861	4.37E-09	4.50E-07
BIN3	ENSG00000147439	8	22,502,296	22,456,517	rs71513892	32	Qi et al. 2018 - brain expression	0.046	0.01	1.047	1.43E-06	1.48E-04
SH3GL2	ENSG00000107295	9	17,688,103	17,684,784	rs10756899	15	Qi et al. 2018 - brain expression	0.252	0.05	1.287	5.83E-08	6.00E-06
ITGA8	ENSG00000077943	10	15,659,036	15,548,925	rs7910668	6	Qi et al. 2018 - brain expression	-0.201	0.05	0.818	6.13E-05	6.32E-03
RNF141	ENSG00000110315	11	10,548,001	10,553,355	rs4910153	120	Võsa et al. 2018 - blood expression	-0.054	0.05	0.947	6.25E-07	1.31E-04
IGSF9B	cg25790212	11	133,800,774	133,800,477	rs11223626	1	Qi et al. 2018 - brain methylation	-0.172	0.04	0.842	3.24E-06	6.48E-04
FBRSL1	cg03621470	12	133,137,479	133,138,334	rs10781619	16	Qi et al. 2018 - brain methylation	-0.057	0.01	0.944	6.35E-05	1.27E-02
CAB39L	ENSG00000102547	13	49,950,524	49,918,175	rs35214871	30	Qi et al. 2018 - brain expression	0.097	0.02	1.102	3.51E-08	3.62E-06
GCH1	ENSG00000131979	14	55,339,148	55,348,837	rs3825611	6	Qi et al. 2018 - brain expression	0.113	0.03	1.120	2.76E-04	2.85E-02
SYT17	ENSG00000103528	16	19,229,472	19,273,554	rs727747	4	Qi et al. 2018 - brain expression	0.177	0.05	1.193	1.54E-04	1.58E-02
SETD1A	ENSG00000099381	16	30,982,526	30,950,352	rs7206511	34	Võsa et al. 2018 - blood expression	-0.710	0.09	0.492	2.75E-13	5.77E-11
CHRNB1	ENSG00000170175	17	7,354,703	7,373,595	rs60488855	18	Qi et al. 2018 - brain expression	0.115	0.03	1.122	1.67E-05	1.72E-03
UBTF	ENSG00000108312	17	42,290,697	42,297,631	rs113844752	34	Võsa et al. 2018 - blood expression	-0.466	0.09	0.628	5.68E-06	1.19E-03
MAPT	ENSG00000186868	17	44,038,724	44,862,347	rs199502	6	Qi et al. 2018 - brain expression	0.265	0.03	1.304	7.13E-24	7.35E-22
WNT3	ENSG00000108379.5	17	44,875,148	44,908,263	rs9904865	2	GTEx v7 - substantia nigra brain expression	-0.082	0.02	0.921	4.01E-06	4.81E-05
DNAH17	cg09006072	17	76,425,972	76,427,732	rs589582	3	Qi et al. 2018 - brain methylation	0.100	0.02	1.106	2.44E-05	4.88E-03
MEX3C	ENSG00000176624	18	48,722,797	48,731,131	rs12458916	40	Võsa et al. 2018 - blood expression	-0.291	0.05	0.748	5.28E-05	1.11E-02

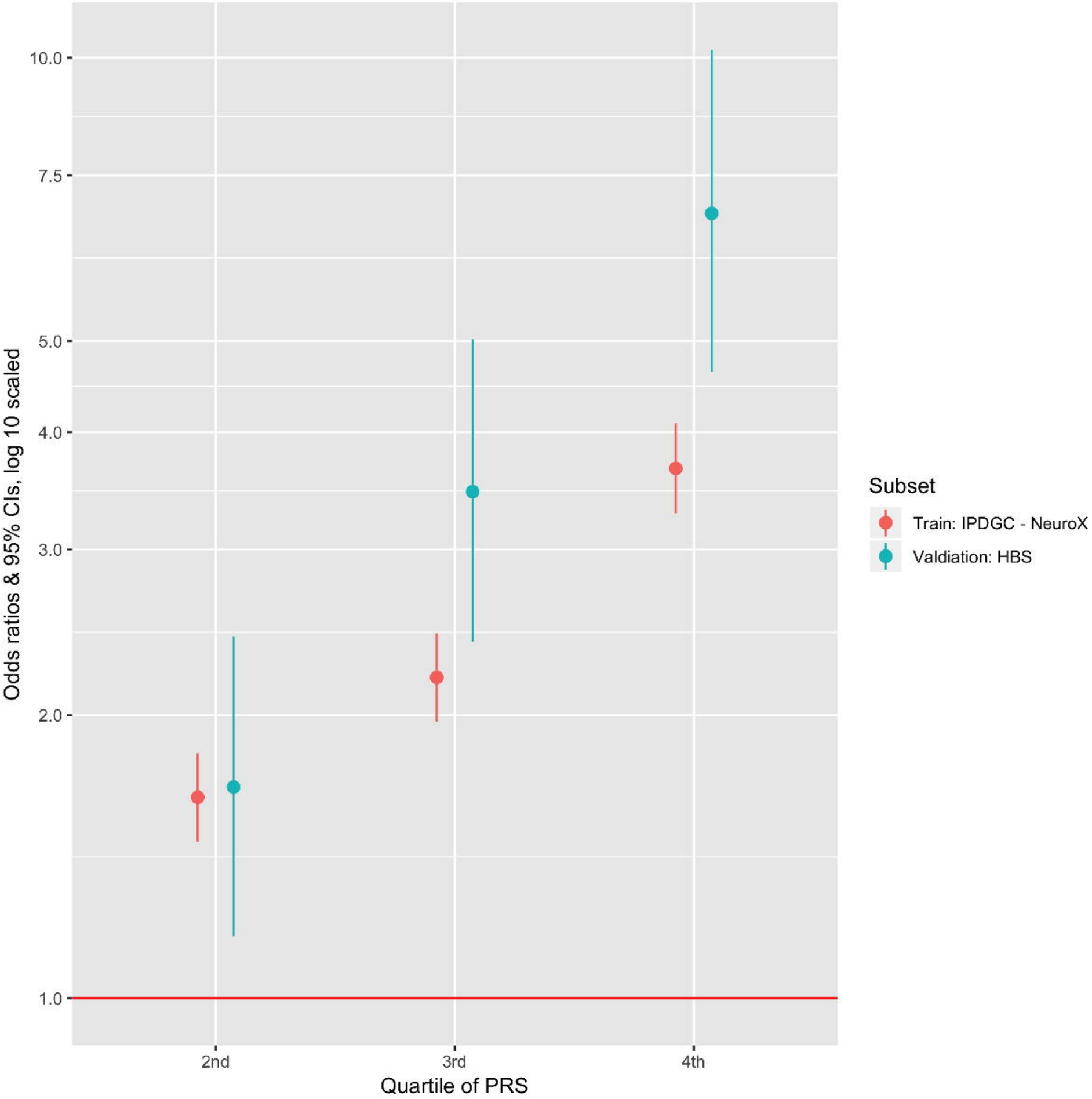
Trait of interest	PMID	correlation,	SE, RG	Z, RG	P, RG	P, FDR adjusted	Observed H2	SE, H2	H2 intercept	SE, H2 intercept	Cross trait intercept	SE, Cross trait intercept
Intracranial volume	25607358	0.351	0.077	4.580	4.64E-06	3.51E-03	0.166	0.045	1.003	0.007	-0.013	0.005
Current tobacco smoking	Not available (UKB)	-0.134	0.034	-3.947	7.92E-05	2.41E-02	0.055	0.003	1.014	0.010	0.004	0.007
Mean Putamen	25607358	0.248	0.064	3.902	9.55E-05	2.41E-02	0.282	0.047	0.952	0.007	-0.007	0.006
Qualifications: NVQ or HND or HNC or equivalent	Not available (UKB)	-0.169	0.045	-3.726	2.00E-04	3.79E-02	0.015	0.002	1.011	0.007	0.005	0.005

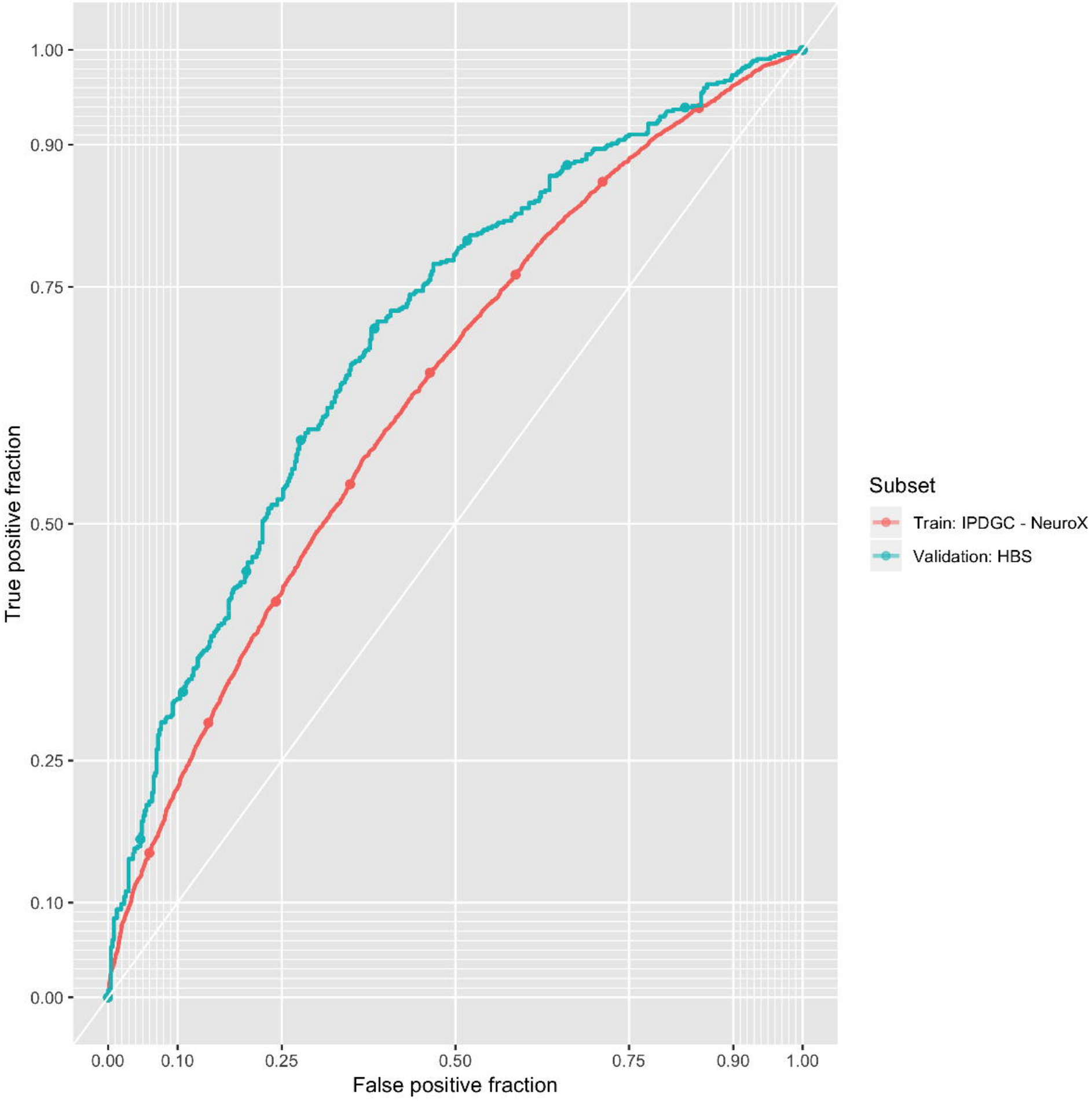
bioRxiv preprint doi: https://doi.org/10.1101/388165; this version posted March 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.





-log10P
BP
KRTCAP2
STK39
TMEM175
SNCA
SNCA
LRRK2
CRHR1
HIP1R
MCCC1
NUCKS1
BST1
HLA-DRB5
INPP5F
LRRK2
GPNMB
ELOVL7
GBAP1
RAB29
GAK
RIT2
SETD1A
FAM47E-STBD1
FAM47E
IGSF9B
SH3GL2
SH3GL2
SIPA1L2
CTSB
GCH1
VPS13C
CRHR1
ITPKB
TMEM163
CAMK2D
FGF20
SYT17
CHD9
CASC16
MAP4K4
SATB1
KPNA1
CLCN3
LOC100131289
ITGA8
GALC
GRN
LINC00693
MED12L
LCORL
RPS12
BAG3
GBF1
FBRSL1
CD19
BSTF
DYRK1A
FCGR2A
IP6K2
RIMS1
RPS6KL1
MED13
PMVK
SPTSSB
SCARB2
PAM
FYN
BIN3
UBAP2
DLG2
MBNL2
NDO2
SPPL2B
CRLS1
VAMP4
KCNIP3
C5orf24
TRIM40
GS1-124K5.11
FAM49B
RNF141
SCAF11
CAB39L
RETREG3
CHRNB1
DNAH17
MEX3C
KCNS3
MIPOL1
WNT3
ASXL3


10.0
7.5
5.0
4.0
3.0
2.0
1.0
Odds ratios & 95% CIs, log 10 scaled
2nd
3rd
4th
Quartile of PRS
Subset
Train: IPDGC - NeuroX
Valdiation: HBS


1.00
0.90
0.75
0.50
0.25
0.10
0.00
True positive fraction
0.00
0.10
0.25
0.50
0.75
0.90
1.00
False positive fraction
Subset
Train: IPDGC - NeuroX
Validation: HBS