npj | parkinson's disease

www.nature.com/npjparkd

ARTICLE OPEN

Check for updates

The IPDGC/GP2 Hackathon - an open science event for training in data science, genomics, and collaboration using Parkinson's disease data

Hampton L. Leonard[1,2,3,4✉], Ruqaya Murtadha[2], Alejandro Martinez-Carrasco[5], Alina Jama[6], Amica Corda Müller-Nedebock[7,8], Ana-Luisa Gil-Martinez[9,10], Anastasia Illarionova[4], Anni Moore[1], Bernabe I. Bustos[11], Bharati Jadhav[12], Brook Huxford[13], Catherine Storm[5], Clodagh Towns[5], Dan Vitale[1,2,3], Devina Chetty[6], Eric Yu[14,15], Francis P. Grenn[1], Gabriela Salazar[16], Geoffrey Rateau[17], Hirotaka Iwaki[1,2,3], Inas Elsayed[18,19], Isabelle Francesca Foote[13,20], Zuné Jansen van Rensburg[6], Jonggeol Jeff Kim[1,13], Jie Yuan[21], Julie Lake[1], Kajsa Brolin[22], Konstantin Senkevich[14,23], Lesley Wu[5], Manuela M. X. Tan[5,24], María Teresa Periñán[25,26], Mary B. Makarious[1,5], Michael Ta[1], Nikita Simone Pillay[27], Oswaldo Lorenzo Betancor[28,29], Paula R. Reyes-Pérez[30], Pilar Alvarez Jerez[1,2], Prabhjyot Saini[14,15], Rami al-Ouran[31], Ramiya Sivakumar[32], Raquel Real[5], Regina H. Reynolds[8,9], Ruifneg Hu[21], Shameemah Abrahams[6], Shilpa C. Rao[33,34], Tarek Antar[1], Thiago Peixoto Leal[33], Vassilena Iankova[35], William J. Scotton[8], Yeajin Song[1,2], Andrew Singleton[1,2], Mike A. Nalls[1,2,3], Sumit Dey[13], Sara Bandres-Ciga[1,2], Cornelis Blauwendraat[1,2], Alastair J. Noyce[13] on behalf of The International Parkinson Disease Genomics Consortium (IPDGC) and The Global Parkinson's Genetics Program (GP2)

Open science and collaboration are necessary to facilitate the advancement of Parkinson's disease (PD) research. Hackathons are collaborative events that bring together people with different skill sets and backgrounds to generate resources and creative solutions to problems. These events can be used as training and networking opportunities, thus we coordinated a virtual 3-day hackathon event, during which 49 early-career scientists from 12 countries built tools and pipelines with a focus on PD. Resources were created with the goal of helping scientists accelerate their own research by having access to the necessary code and tools. Each team was allocated one of nine different projects, each with a different goal. These included developing post-genome-wide association studies (GWAS) analysis pipelines, downstream analysis of genetic variation pipelines, and various visualization tools. Hackathons are a valuable approach to inspire creative thinking, supplement training in data science, and foster collaborative scientific relationships, which are foundational practices for early-career researchers. The resources generated can be used to accelerate research on the genetics of PD.

npj Parkinson's Disease (2023)9:33 ; https://doi.org/10.1038/s41531-023-00472-6

INTRODUCTION

An abundance of Parkinson's disease (PD) data spanning many different modalities (genetic, transcriptomic, proteomic, epigenomic, clinical, and more) has been generated over the past few years. However, many challenges remain in effectively using and integrating this data to produce meaningful and impactful

[1]Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA. [2]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. [3]Data Tecnica International LLC, Washington, DC, USA. [4]German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany. [5]Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, University College London, London, UK. [6]Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, University College London, London, UK. [7]Division of Molecular Biology and Human Genetics, Department of Biomedical Sciences, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa. [8]South African Medical Research Council/Stellenbosch University Genomics of Brain Disorders Research Unit, Stellenbosch University, Cape Town, South Africa. [9]Department of Neurodegenerative Disease, University College London, London, UK. [10]Great Ormond Street Institute of Child Health, Genetics and Genomic Medicine, University College London, London, UK. [11]The Ken & Ruth Davee Department of Neurology and Simpson Querrey Center of Neurogenetics, Feinberg Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA. [12]Department of Genetics and Genomic Sciences and Mindich Child Health and Development Institute, Icahn School of Medicine at Mount, Hess Center for Science and Medicine, New York, NY 10029, USA. [13]Preventive Neurology Unit, Wolfson Institute of Population Health, Queen Mary University of London, London, UK. [14]Department of Human Genetics, McGill University, Montreal, QC, Canada. [15]The Neuro (Montreal Neurological Institute-Hospital), McGill University, Montreal, QC, Canada. [16]INNCOSYS, Col. Morelos Second Section, 50120 Toluca de Lerdo, México. [17]Institut du Cerveau - Institute of Brain and Spine (ICM), Hôpital Pitié, 47 Bd de l'Hôpital, 75013 Paris, France. [18]Faculty of pharmacy, University of Gezira, Wad MedaniP.O. Box 20Sudan. [19]International Parkinson Disease Genomics Consortium (IPDGC)-Africa, University of Gezira, Wad MedaniP.O. Box 20Sudan. [20]Unit for Psychological Medicine, Wolfson Institute of Population Health, Queen Mary University of London, London, UK. [21]Center for Advanced Parkinson Research, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA. [22]Translational Neurogenetics Unit, Wallenberg Neuroscience Center, Department of Experimental Medical Science, Lund University, Lund, Sweden. [23]Department of Neurology and Neurosurgery, McGill University, Montréal, QC, Canada. [24]Department of Neurology, Oslo University Hospital, Oslo, Norway. [25]Unidad de Trastornos del Movimiento, Servicio de Neurología y Neurofisiología Clínica, Instituto de Biomedicina de Sevilla, Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Seville, Spain. [26]CIBERNED, Madrid, Spain. [27]South African National Bioinformatics Institute (SANBI), South African Medical Research Council Bioinformatics Unit, University of the Western Cape, Bellville, South Africa. [28]Veterans Affairs Puget Sound Health Care System, Seattle, WA, USA. [29]Department of Neurology, University of Washington School of Medicine, Seattle, WA, USA. [30]Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de, México, Juriquilla, México. [31]Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA. [32]University of Southern California, Los Angeles, CA 90007, USA. [33]Department of Genomic Medicine, Lerner Research Institute, Cleveland Clinic Foundation, Cleveland, OH 44195, USA. [34]Department of Molecular Medicine, Case Western Reserve University, Cleveland, OH 44106, USA. [35]Department of Neurology With Friedrich Baur Institut, University Hospital of Ludwig-Maximilians-Universität München, Munich, Germany. A full list of members and their affiliations appears in the Supplementary Information. ✉email: leonardhl@nih.gov

Published in partnership with the Parkinson's Foundation

results. A lack of access to good training resources and limited connections to other researchers can delay progress. These problems are even more evident in underrepresented populations, where a lack of resources may hinder researchers and clinicians from performing necessary research on diverse ancestry populations and building local research capacity. To accelerate research, the wide adoption of open science practices is critical. Bringing researchers together to share data, code, tools, and pipelines will help reproducibility and create lower entry points to generate the results needed to make the necessary progress in PD research.

With the intention of creating pipelines and tools to facilitate PD genetics research, 49 early-career scientists from 12 countries collaborated in a virtual 3-day "Hackathon" event in May 2021. The event combined scientists from two initiatives, the International Parkinson's Disease Genomics Consortium (IPDGC)[1] and the first resource project of the Aligning Science Across Parkinson's (ASAP) initiative, the Global Parkinson's Genetics Program (GP2)[2]. IPDGC and GP2 exist to drive forward research into the genetic basis of PD, and for both initiatives, training and collaboration are vital aspects of accelerating and diversifying research efforts. Hackathons are a helpful tool that can help promote this mutual effort, providing a creative and engaging outlet to facilitate networking and team building. Hackathons have recently become popular in the health and biology domain, where the outcomes have focused on a variety of research disciplines and challenges. Some hackathons focus on the development of useful tools, such as the event hosted by DNAnexus and the Baylor College of Medicine in 2020, where participants created novel tools for structural variation and SARS-CoV-2 research[3]. Others focus on reaching a solution or creative idea to a research question through competition, such as the events using data from The Alzheimer's Disease Neuroimaging Initiative (ADNI), where teams competed to predict mild cognitive impairment (MCI) conversion to Alzheimer's Disease[4]. The first combined 'GP2/IPDGC Hackathon 2021' event provided teams with a choice of nine PD genetics and genomics analysis topics. The focus of the GP2/IPDGC Hackathon was not competition, but training, networking, collaboration, and the development of tools that would be useful to the broader research community.

As part of an open science initiative, the Accelerating Medicines Partnership Parkinson's Disease Program (AMP PD, https://amp-pd.org/)[5] aims to identify and validate biomarkers by providing researchers access to a large, harmonized dataset that includes clinical, genomic, transcriptomic, and proteomic data. GP2 has recently partnered with AMP PD to make one space where researchers can access multiple PD cohorts and a range of data with a single sign-on. Both AMP PD and GP2 use Terra[6], a platform for researchers to access data and run analysis tools (Fig. 1). The Terra platform allows analysis to be done directly in the cloud, navigating many data sharing and research governance issues that arise from combining data from different geographical regions. Additionally, cloud analysis allows for ease in collaboration, replication of results, and addresses privacy and data use concerns by preventing the download of individual-level data. Using Terra is necessary for accessing the wealth of AMP PD and GP2 data, so many of the project topics for this hackathon involved gaining experience and building pipelines on Terra. This allowed hackathon participants to get comfortable using Terra and cloud computing and make available tools to help future researchers use Terra easily and quickly. Whether the teams employed Terra or another platform to create their tools, all were created with the idea to openly share whenever possible to any interested party to accelerate progress.

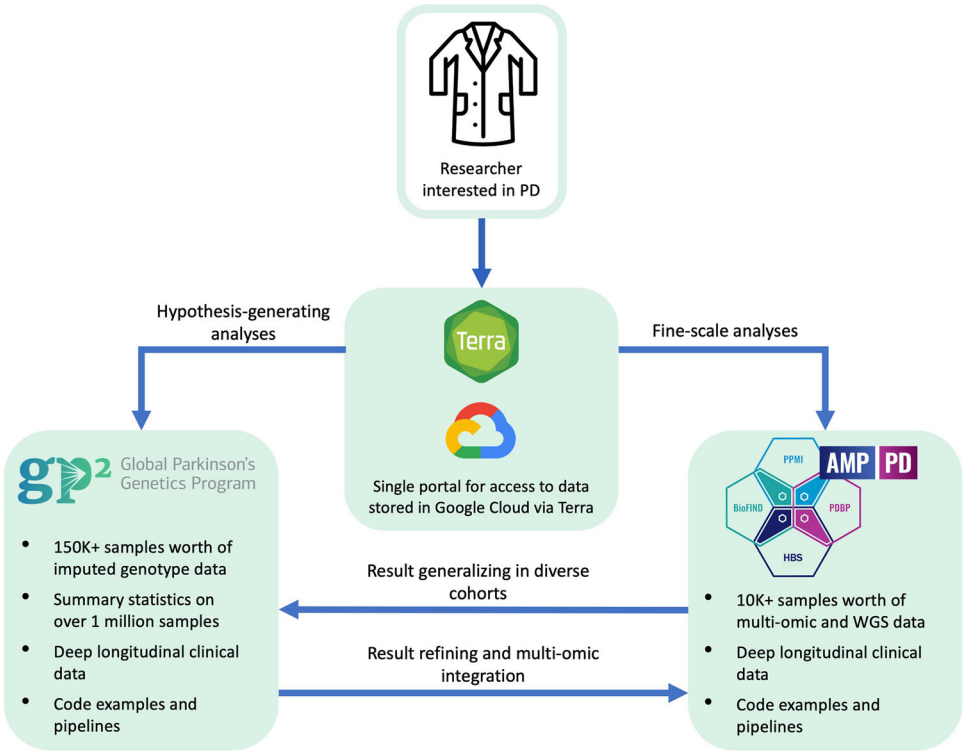

Fig. 1 The AMP PD and GP2 collaboration framework. GP2 sample numbers represented in this figure refer to planned numbers, not current. A single data use agreement (DUA) and the Terra cloud platform handle access to both data resources. Each resource addresses different but complementary research questions. GP2 and its wealth of diverse population genotype data were suited for large population-based and hypothesis-generating analyses. In contrast, AMP PD and its multi-omic and whole-genome sequencing data were suited for fine-scale analyses.

RESULTS

Hackathon description

The "GP2/IPDGC Hackathon 2021" event included 49 participants from 12 countries: Brazil, Canada, France, Germany, India, Mexico, Norway, Russia, South Africa, Spain, Sudan, Sweden, the UK, and the United States. This event was open to all levels of experience and career stages, so participants had a range of coding and genetic analysis skills. Each of the nine teams were designed in order to spread participant experience evenly so that the skill levels were well-balanced. Participant demographics are further outlined in Table 1.

Each of the 3 days of the event had scheduled teamwork sessions, but teams were allowed to create their own schedules outside of those sessions to accommodate the wide array of time-zone differences inherent to an international event. The event also provided scheduled training and seminar sessions. On the first day, participants were given a seminar on the use of machine learning in healthcare as well as a hands-on demonstration of Terra to prepare them for some of the hackathon project topics. On the second day, participants were given a seminar on code "health". The first two days also included time for icebreakers and games, both within each team and across all hackathon participants, in order to build better connections and facilitate continued academic contact outside of the event. The third and final day was reserved for working to finish the projects and final presentations. Github was used to share code and applications were deployed publicly.

Hackathon outcomes

To simplify the description of the outcomes of the hackathon, the nine projects from this event have been grouped into one of three categories; genome-wide association study (GWAS)-level and post-GWAS analyses, downstream analyses of genetic variation, and data visualization. GWAS-level and post-GWAS projects were designed to be helpful to researchers looking for ways to follow up their GWAS analyses easily. The projects for this topic included efforts to create general post-GWAS and pathway and cell-type enrichment workflows and examples. Downstream analyses of genetic variation projects aimed at providing examples and pipelines for additional investigation of genetic variation, including colocalization, variant interaction, and network generation and visualization. Finally, the data visualization projects aimed to provide resources that other researchers may use to help further their research or generate hypotheses). These included a GP2 cohort browser, expansion of a PD locus browser, visualization of longitudinal PD clinical outcomes, and visualizing longitudinal and cross-variant effects. A summary of the projects and the topics they belong to, and links to available code and applications are found in Table 2. A glossary for any potentially unfamiliar terms can be found in Box 1.

Table 1. Summary of hackathon participant demographics.

	Raw counts	% Participants
Institutional affiliation		
Academic	36	73.5
Industry	1	2.0
Government	12	24.5
Other	0	0
Python experience		
Beginner - Intermediate	38	77.6
Advanced - Expert	11	22.4
Notebook experience (Jupyter, Google Colab)		
Beginner - Intermediate	37	75.5
Advanced - Expert	12	24.5
R experience		
Beginner - Intermediate	28	57.1
Advanced - Expert	21	
Command line interface experience		
Beginner - Intermediate	36	42.9
Advanced - Expert	13	26.5
Terra experience		
Beginner - Intermediate	45	91.8
Advanced - Expert	4	8.2

DISCUSSION

In the first IPDGC and GP2 joint hackathon, 49 early-career researchers from across the globe came together virtually to create the tools described here. The event lasted three days and was broken up into virtual team meetings, icebreakers and networking breaks, work sessions, and training opportunities. Each of the nine projects had teammates assigned according to their interests and skill level. At the end of the 3-day event, each team presented their progress and results. Within the following week, teams refined their Github content or deployed their apps and the final tools are now available to the public to aid future research.

As the amount of data available for disease research grows and becomes increasingly cloud-centric, public tools like these will help reduce the difficulty and time it takes to visualize, analyze, and understand this data. In order to do effective and prompt research, sharing tools and code for analyses must become the standard.

In addition to creating several helpful pipelines and visualization applications for the PD research community, this hackathon revealed the need for further documentation and training on cloud computing in the disease research field. Many of the tools created during this event are designed to be used in a cloud setting to assist new researchers in analyzing cloud-based data. However, more resources like these will be needed to ensure cloud resources can be used efficiently.

Hackathons are a valuable tool for prototyping new ideas and tools, but they are also helpful for creating new collaborative networks and working relationships among researchers. This event's virtual setting allowed many people of different backgrounds and skill sets to work together on creative solutions. Encouraging this kind of creative thinking and creating opportunities for trainees to invest in new and necessary skills is integral to facilitating productive research.

The "GP2/IPDGC Hackathon 2021" event resulted in the creation of novel pipelines and applications designed to assist future genetics research. Many of these novel pipelines were designed for Terra, to make genetic analysis in the cloud more accessible to researchers inexperienced with cloud computing platforms. An added benefit of this event was that participants were given the opportunity to increase their analysis skills and build connections by working together on a new project. By providing training opportunities to scientists and producing original and novel applications, this event further demonstrated collaborative and international hackathons as an important tool for the scientific community.

METHODS

GWAS-level and post-GWAS analyses

GWAS of PD have nominated 90 independent risk signals in individuals of European ancestry, explaining ~16–36% of the heritable risk[7], as well as two additional risk signals in Asian

Table 2. Summary of the goals and outcomes of each of the hackathon projects.

Topic	Project title	Goal	Outcome	Links to code or applications
GWAS-level and post-GWAS analyses	Project 1: Post-GWAS analysis	Develop a pipeline for some common post-GWAS follow-up analyses	GREML-LDMS and PRS analyses coded and tested on AMP PD data	Github: https://github.com/ipdgc/GP2-post-GWAS-analysis Zenodo: https://doi.org/10.5281/zenodo.6477900
	Project 2: Pathway and cell-type enrichment pipeline	Develop a pipeline for investigating pathway and cell-type enrichment from GWAS summary statistics	Code for gene set enrichment with WebGestaltR on Terra as well as formatting summary statistics for FUMA	Github: https://github.com/ipdgc/GP2-pathway-enrichment-pipeline Zenodo: https://doi.org/10.5281/zenodo.6477914
Downstream analyses of genetic variation	Project 3: Colocalization pipeline	Develop a pipeline for colocalization analysis	Code for colocalization and visualization with the coloc and eQTpLot R packages	Github: https://github.com/ipdgc/Colocalization-Pipeline Zenodo: https://doi.org/10.5281/zenodo.6477921
	Project 4: Network generation and visualization pipeline	Develop a pipeline that generates and visualizes gene regulatory networks	Code for generating Leiden networks with eQTL and genetic data with the leidenalg python package	Github: https://github.com/ipdgc/GP2-network-generation Zenodo: https://doi.org/10.5281/zenodo.6477923
	Project 5: Variant interaction pipeline	Develop a pipeline that investigates variant interaction	Code for data prep with Plink1.9 and ANNOVAR, as well as interaction tests with glm in R	Github: https://github.com/ipdgc/GP2-Variant-Interaction-Pipeline Zenodo: https://doi.org/10.5281/zenodo.6477931
Data visualization	Project 6: GP2 cohort tracker visualization	Develop a dashboard for tracking and investigating the progress of cohort integration for GP2	Both an internal and external dashboard can be used to investigate information about participating GP2 cohorts	Application: https://gp2.org/cohort-dashboard/
	Project 7: PD GWAS Loci Browser expansions	Expand the functionality of the PD GWAS Loci Browser	Locus zoom plots for conditional analyses, power calculations, violin plots for expression data, and user statistics were added to the Loci Browser	Application: https://pdgenetics.shinyapps.io/GWASBrowser/ RRID: SCR_022188
	Project 8: Visualization of longitudinal UPDRS/HY scores	Visualize longitudinal clinical measures of PD	Python Streamlit application for visualizing UPDRS and HY score progression	Application: https://share.streamlit.io/tantar/hack/main/GP2_data_visualization.py RRID: SCR_022187
	Project 9: Visualize longitudinal and cross-sectional variant effects	Visualize longitudinal and cross-sectional variant effects from GWAS	Code for running an R shiny application that visualizes variant effects across cohorts	Github: https://github.com/ipdgc/GP2-visualize-longitudinal-variant-effects Zenodo: https://doi.org/10.5281/zenodo.6477935

Access to the code or applications developed during the hackathon is included in the "Links to Code or Applications" column.

Box 1 Glossary for unfamiliar terms

Accelerating medicines partnership Parkinson's disease (AMP PD)	A program with the aim of deep molecular characterization and longitudinal clinical profiling of PD patient data and biosamples with the goal of identifying and validating diagnostic, prognostic, and/or disease progression biomarkers for PD.
ANNOVAR	A software tool to utilize update-to-date information to functionally annotate genetic variants.
BioFIND	An observational clinical study designed to discover and verify biomarkers of PD.
Broad-sense heritability	The proportion of phenotypic variance that can be attributed to genetic causes.
Colocalization	A process that determines whether a single variant is responsible for both the GWAS and eQTL signals in a locus.
Expressive quantitative trait locus (eQTL)	A locus that explains a fraction of the genetic variance of a gene expression phenotype.
Functional mapping and annotation (FUMA)	A platform that can be used to annotate, prioritize, visualize, and interpret GWAS results.
Genome-wide association study (GWAS)	An analysis used to identify inherited genetic variants associated with the risk of disease or a particular trait.
The Global Parkinson's Genetics Program (GP2)	An ambitious program to genotype >150,000 volunteers around the world to further understand the genetic architecture of PD.
GREML-LDMS	Method to estimate heritability for human complex traits in unrelated individuals using whole-genome sequencing (WGS) data
Hoehn and Yahr (HY) scale	A scoring system that allows for the quantification of the different stages of PD.
The International Parkinson Disease Genomics Consortium (IPDGC)	A worldwide collaboration dedicated to the identification of both Mendelian and risk genes is important for PD.
Leiden algorithm	Network building algorithm with potential benefits over the Louvain algorithm.
Locus (plural: Loci)	The specific physical location of a gene or other DNA sequence on a chromosome.
Louvain	Network building algorithm.
Linkage disequilibrium (LD)	The nonrandom association of alleles at different loci.
Narrow-sense heritability	The fraction of phenotypic variance that can be attributed to additive genetic variation.
Parkinson's progression markers initiative (PPMI)	A study is collaborating with partners around the world to create a robust open-access dataset and biosample library for PD.
The Parkinson's Disease Biomarkers Program (PDBP)	A program dedicated to accelerating the pace of PD biomarkers research.
Polygenic risk score (PRS)	A measure of disease risk calculated by the cumulative effect of multiple risk variants.
Posterior probability (PP)	The statistical probability that a hypothesis is true when calculated in the light of relevant observations.
Single nucleotide polymorphism (SNP)	A variation in a single base pair in a DNA sequence.
Unified Parkinson's disease rating scale (UPDRS)	A rating tool used to gauge the severity and progression of PD in patients.

populations[8]. Typically, published GWAS are accompanied by various follow-up analyses, but performing these analyses is not always straightforward. Common post-GWAS analyses include heritability estimation and polygenic risk score (PRS) calculation. Heritability analyses estimate the percentage of disease risk accounted for by common genetic variants, and PRS can be used to predict disease risk by aggregating the effects at multiple common risk loci[9]. Another follow-up approach to help interpret GWAS results is to combine Single Nucleotide Polymorphisms (SNPs) into a group of functionally related genes, such as genes belonging to a single biological pathway or cell type. This is called gene set enrichment analysis and is a widely used approach to examine the cumulative effect of SNPs in a particular biological process and determine whether there are particular pathways, processes, or cell types affected in disease.

Project 1 (post-GWAS analysis). We used AMP PD version 1 release data to develop a Terra-based pipeline for assessing SNP-based heritability, as well as polygenic risk score calculation. Using the GREML-LDMS method[10] applied to data from the AMP PD version 1 release, we estimated narrow-sense heritability (h^2) to be roughly 52%, which is much higher than the typical estimate of 22%[7], and is likely biased due to the recruitment of specific variant carriers present in the AMP PD cohort. AMP PD provides information regarding the recruitment arm. Researchers should investigate this information to determine if specific samples need to be removed from certain analyses in future work. We then used PLINK v1.9[11] and estimated risk effect sizes from the summary statistics of the latest PD GWAS[7] to calculate the genetic risk scores of PD *LRRK2* mutation carriers ($n = 382$), control *LRRK2* mutation carriers ($n = 275$), and control individuals without PD causing mutations ($n = 3435$) from the AMP PD version 2.5 dataset. We tested normalized *z*-scores for association with *LRRK2* carrier disease status. Mean ± standard deviation of unadjusted PRS scores were higher in PD *LRRK2* mutation carriers (-0.0166 ± 0.004), compared to control *LRRK2* mutation carriers (-0.0182 ± 0.004) and controls (-0.0180 ± 0.004), suggesting that PD *LRRK2* mutation carriers share a common polygenic risk profile with idiopathic PD (iPD), contrary to control *LRRK2* mutation carriers (Fig. 2a).

Project 2 (pathway and cell enrichment pipeline). We aimed to create a pipeline to annotate GWAS summary statistics to test the enrichment of biological pathways and cell types. Analyses were performed on the Terra platform using the most recent PD GWAS summary statistics[7]. We created a pipeline to correctly format the GWAS summary statistics for a common annotation tool Functional Mapping and Annotation of Genome-Wide Association

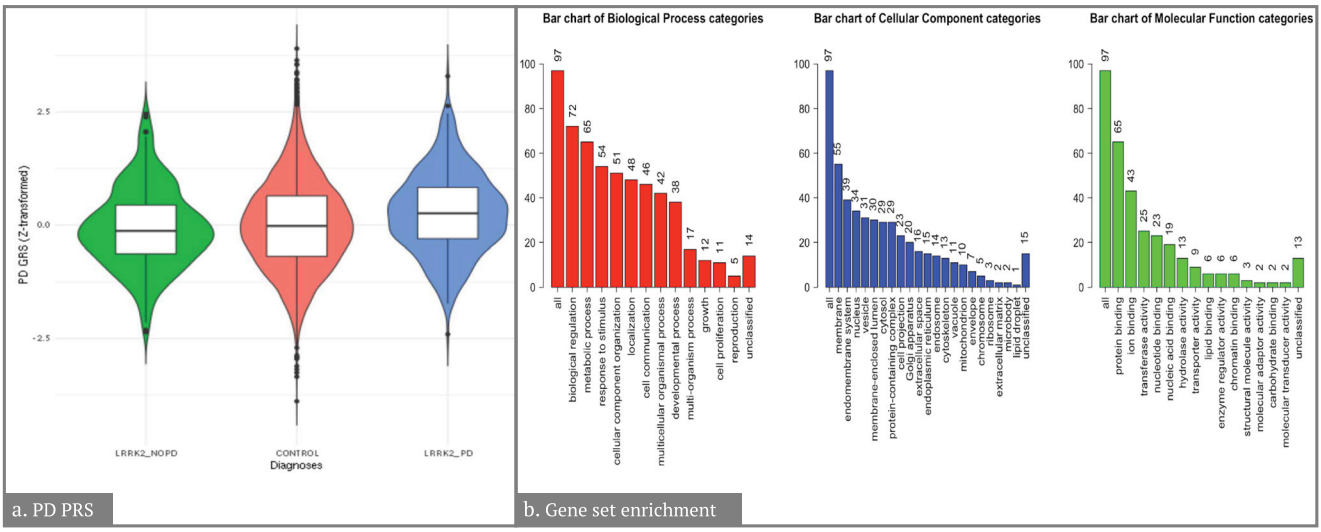

Fig. 2 Results from the GWAS-level and post-GWAS analyses projects. a Violin plots comparing z-transformed Parkinson's disease (PD) genetic risk score distributions in PD-LRRK2 cases, non-PD-LRRK2 carriers, and controls. Within the violins, box plots display the median and the bounds of the box correspond to the 25th and 75th percentiles. The upper and lower limits of the whiskers correspond to 1.5 times the limits of the 25th and 75th percentiles. PD-LRRK2 individuals had a higher risk of developing PD compared to control LRRK2 mutation carriers (OR = 1.60, 95% CI = 1.33–1.93, $P = 1.1 \times 10^{-6}$). The mean of the unadjusted GRS score was also significantly higher in PD-LRRK2 cases compared to non-PD-LRRK2 carriers ($P = 2.9 \times 10^{-6}$) and controls ($P = 5.1 \times 10^{-7}$) in the pairwise Wilcoxon rank-sum test. **b** Summary of input genes from WebGestalt showing the number of PD genes (from GWAS significant SNPs) which overlap with the annotated genes in the Gene Ontology Slim terms from the biological process, cellular component, and molecular function.

Studies (FUMA)[12], then downloaded the formatted data and uploaded it to FUMA. We also ran the WEB-based GEne SeT AnaLysis Toolkit (WebGestalt) directly on Terra using the WebGestaltR package[13]. We selected the nearest genes to the GWAS significant SNPs ($P < 5 \times 10^{-8}$) from Nalls et al. 2019 GWAS summary statistics. Using WebGestaltR, we conducted an over-representation analysis and gene set enrichment analysis. We identified 97 unique genes from the genome-wide significant hits in the PD GWAS summary statistics. We generated summary data for these PD genes annotated by biological processes, cellular components, and molecular functions (Fig. 2b). There was no significant enrichment of any Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway gene sets in the overrepresentation analysis (FDR $P < 0.05$).

Downstream analyses of genetic variation

While GWAS have identified many common variants associated with complex diseases like PD, it is follow-up analyses that have started to decode GWAS results, and more downstream analysis is needed to unravel the implications of observed genetic variation in PD. Three types of analysis were the focus for the downstream analyses of the genetic variation topics, including colocalization, variant interaction, and network generation and visualization. Colocalization analysis allows the calculation and estimation of the correlation between a GWAS locus and an expression quantitative trait locus (eQTL). Variant interaction, or epistasis, is an interaction of genetic variation at two or more loci to produce a phenotypic outcome that is not predicted by the additive combination of effects attributable to the individual loci[14]. Its importance in humans continues to be a matter of debate[15,16], but it may explain some of the "missing heritability" underlying complex diseases such as PD[16–18]. In addition to investigating individual variant effects with colocalization and epistasis, visualizing biological networks can help with understanding complex molecular relationships and interactions. In PD research, genetic and gene expression data has been used in community network analysis to nominate pathways and genes for drug target and functional prioritization[19,20].

Project 3 (colocalization pipeline). Colocalization analysis takes into account five hypotheses: H0 (no association between the locus and either trait), H1 (locus has an association with first trait only), H2 (locus has an association with second trait only), H3 (locus has an association with both traits but driven by different SNPs which are not in linkage disequilibrium (LD)), H4 (locus has an association with both traits driven by same SNPs). For Project 3: Colocalization pipeline, we considered colocalization analysis with a posterior probability of colocalization in H4 (PPH4) greater than 0.8 to be significant. We utilized the coloc R package[21,22] and summary statistics from ref. [7]. We used eQTL data from a cerebellar cortical meta-analysis of four cohorts[23], publicly available from the AMP-AD Knowledge Portal[24]. As an example for our pipeline, we extracted the region ± 500 kb around *DYRK1A*, nominated in Nalls et al. 2019, from the GWAS summary statistics and eQTL data. To visualize the results, we employed the eQTpLot R package[25], which can generate different plots for GWAS and eQTL signal colocalization, as well as the correlation between their *p* values and enrichment of eQTLs among variants and LD of loci of interest, allowing efficient and intuitive visualization of gene expression and trait interaction. We used our previously generated results for *DYRK1A* and whole brain eQTL as an example for creating visualizations using this package. (Fig. 3a).

Project 4 (network generations and visualization pipeline). We sought to develop a Leiden network and subsequent visualization pipeline for transcriptomic and genomic data to identify and visualize both a priori and complex phenotype gene regulatory networks. The Leiden algorithm is one option for community detection of networks and can be faster and return more reliable results than the more well-known Louvain algorithm[26]. We relied on the leidenalg[27] package in Python to produce weighted and unweighted networks on GWAS summary statistics and then visualized the resulting networks. (Fig. 3b). Data used consisted of AMP PD genomic, transcriptomic data, and public eQTL data from the eQTL catalog[28] and PD summary statistics from the most recent PD GWAS[7]. This project was designed as a proof-of-concept for a pipeline for detecting gene networks and relating them to PD phenotype information via GWAS summary stats.

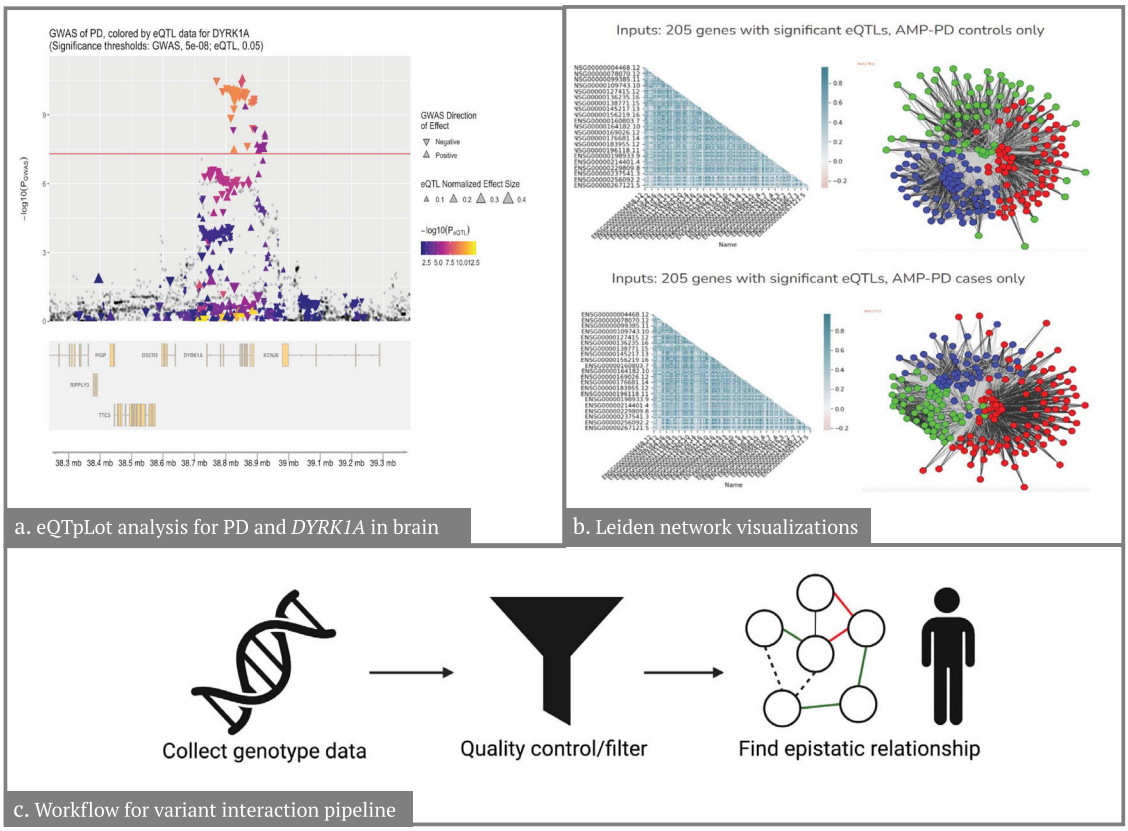

Fig. 3 Results from the downstream analysis of genetic variation projects. a Displays the locus of interest, in this case, ±500 kb from DYRK1A, and the horizontal line depicts the GWAS significance threshold of $P = 5 \times 10^{-8}$. Displays the genes in the locus of interest. **b** Depicts the Leiden gene networks and correlations for significant eQTLs for PD controls and PD cases. **c** Depicts the general workflow for the variant interaction pipeline.

Project 5 (variant interaction pipeline). We developed a workflow that can be summarized as follows: (1) We utilized individual-level test data in binary format to perform data harmonization with PLINK v1.9[11] to ensure that the risk allele was consistent for all the variants; (2) We established a minor allele frequency (MAF) threshold >0.05 to subset variants, keeping only common genetic variation; (3) We annotated variants of interest using ANNOVAR[29], differentiating between coding and non-coding as well as annotated predicted gene consequence; (4) We carried out interaction analyses in R 3.6 using the glm() function and adjusting for age, gender, and the first five components. (Fig. 3c).

Data visualization

Visualization of clinical and genetic data plays an essential role in research. It can be used to inform the progress of initiatives like GP2, help researchers to view data in a meaningful way, and generate and corroborate hypotheses. As GWAS and other analyses nominate more PD risk loci, efforts to decode the role of these variants and how they interact with both longitudinal and cross-sectional phenotypes will be needed. Four projects focused on data visualization, including a GP2 cohort tracker, updates to the IPDGC locus browser, visualization of longitudinal clinical phenotypes, and visualization of longitudinal and cross-sectional variant effects.

Project 6 (GP2 cohort tracker visualization). We designed the GP2 cohort tracker visualization to show essential information about cohorts recruited for GP2 and showcase their diversity, geographical location of enrollment, ancestry representation, and additional relevant metadata. We designed this visualization to inform progress and inspire others to contribute to this initiative. In the form of a one-page dashboard developed with the open-source Python software Streamlit, the visualization includes separate maps for complex and monogenic cohorts. It was critical to include easy-to-use search and discovery aspects built into the dashboard. If a user knows the name of a particular cohort, then they can pull up information for that cohort that populates the rest of the dashboard. The user can also filter by general methods such as cohort size or country. This design is used internally and externally on the GP2 website (https://gp2.org/cohort-dashboard/) to inform those interested in the progress of GP2's cohort integration (Fig. 4a).

Project 7 (IPDGC GWAS loci browser expansions). To facilitate investigations of nominated risk variants, members of IPDGC have created a PD GWAS locus browser (https://pdgenetics.shinyapps.io/GWASBrowser/) that makes relevant statistics and datasets available to the public[30]. Throughout the hackathon, our team continued the development of this browser through the addition of new datasets and features. To identify secondary association signals at each locus from the Nalls et al. 2019 study, we performed conditional analysis using the Genome-wide Complex Trait Analysis (GCTA) tool[31,32]. Locus zoom plots were added to display the results of this conditional analysis (Fig. 4b)[33]. Power calculations were done for each risk variant by Nalls et al. 2019 to determine if the findings were sufficiently powered. To do so, we followed methods used by the Genetic Association Study Power Calculator tool (https://csg.sph.umich.edu/abecasis/gas_power_calculator/), using summary statistics from Nalls et al. 2019, a disease prevalence of 0.01, and a significance level of 0.05 as input. We queried blood gene expression data included in the AMP PD version 2.5 release to measure

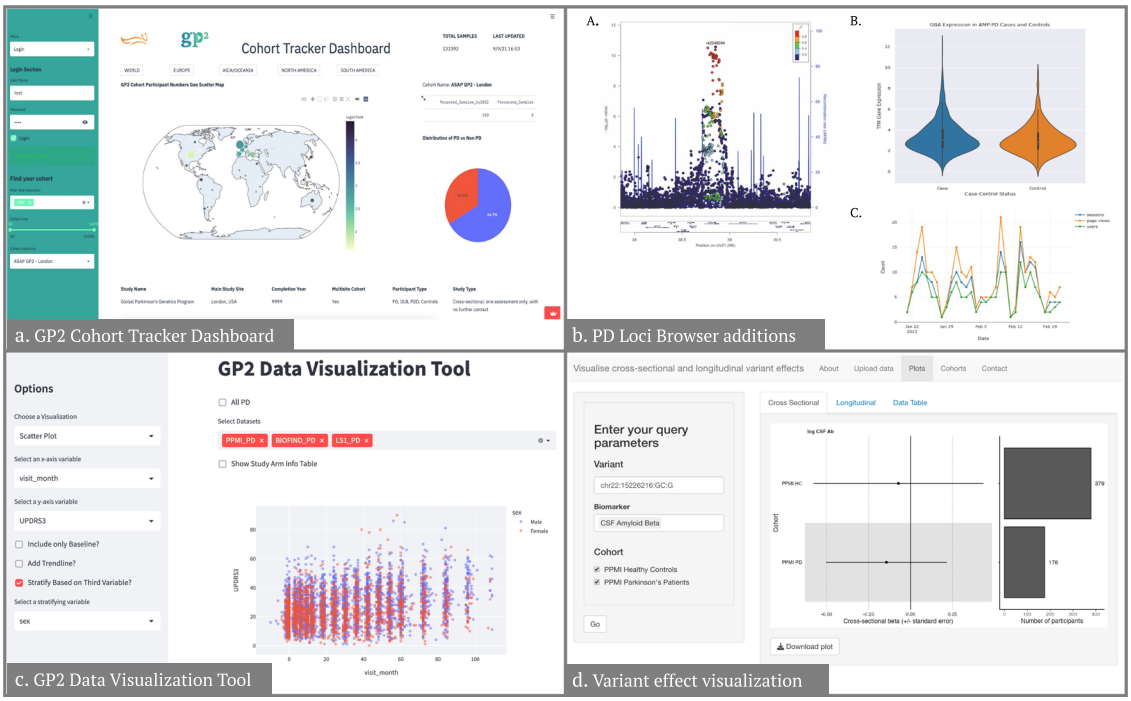

Fig. 4 Results from the data visualization projects. a The left banner allows for filtering and specific cohort selection, the map depicts cohort origin, and the right panel depicts PD vs. non-PD distribution. **b** (A) Locus zoom plot generated using conditional analysis statistics for locus 78 PD risk variant rs2248244. (B) Violin plot of GBA expression in AMP PD cases and controls. (C) Example plot of browser user visits over time. **c** Depicts an example image from the app, in this case, a scatter plot visualization of UPDRS2 scores across visits, color-coded for sex. **d** On the left, users can input their query parameters, including variant, biomarker(s), and cohort(s) of interest. On the right, a forest plot demonstrates the regression beta for the variant of interest in cross-sectional data. A bar plot demonstrates the number of participants in each cohort. The exact visualization is also available for longitudinal data, with all available data available in a tabular format in the "Data Table" tab.

expression levels in PD cases and controls. We obtained TPM expression at baseline for samples that had case or control status and no PD mutations in whole-genome sequencing data, leaving a total of 1710 samples. Expression data for each gene was displayed in a violin plot and added to the expression section of the browser (Fig. 4b). The literature section of the browser was updated to display a description, PubMed hit count, and word cloud plot for each gene within 1 MB of a PD risk variant. Our last addition to the browser was a display of user statistics. We used the googleAnalyticsR package[34] to record and visualize the number of visits for the browser and each risk variant within a period specified by the user (Fig. 4b).

Project 8 (visualization of longitudinal UPDRS/HY scores). The Unified Parkinson's Disease Rating Scale (UPDRS) and Hoehn-Yahr (HY) stage are two of the most common measures of the severity of PD. We set out to develop a user-friendly and adaptable app to display a diverse set of visualizations of longitudinal UPDRS/HY scores, based on data from GP2, utilizing the Streamlit library from Python. During the Hackathon, we successfully integrated data from three cohorts: Parkinson's Progression Markers Initiative (PPMI), Parkinson's Disease Biomarkers Program (PDBP), and BioFIND[35–37]. We were also able to produce four different visualizations (Fig. 4c). First, we created bar graphs to visualize changes in scores from the data over time, and we added the option to include baseline patients only. Second, we created line graphs showing confidence intervals for longitudinal changes in HY and UPDRS scores. Our final visualizations were experimental, but we produced proof-of-concept visualizations with limited options. We created a Sankey graph visualization that better visualized how participants moved between different subsets of the population over time. Lastly, the fourth visualization is a Kaplan–Meier curve showing the time to reach a certain threshold within our progression scores.

Project 9 (Visualize longitudinal and cross-sectional variant effects). We set out with the aim of creating an interactive and user-friendly web application that would allow users to (i) visualize the effect of a genetic variant across multiple cohorts using publicly available GWAS summary statistics and (ii) input their GWAS summary statistics for visualization and meta-analysis with existing data. As test data, we used a small subset of results from a study of amyloid-β levels in cerebrospinal fluid derived from healthy control individuals and individuals with PD from the PPMI dataset[35]. Amyloid-β levels were measured at baseline and in follow-up visits; thus, results were available from both a cross-sectional and longitudinal GWAS. Using the R shiny[38] framework, we produced a skeleton framework for our web application, with several tabs, including (i) an "Upload data" tab where users could upload and query their data and (ii) a "Plots" tab where users could query the available test data and visualize it. In the "Plots" tab, we allowed users to query by a variant of interest, with the option to choose which biomarker(s) and cohort(s) they wished to visualize. The variant beta was visualized across cohorts and biomarkers using a forest plot, while the number of participants/observations was visualized using a bar plot (Fig. 4d). Tabs were available for cross-sectional and longitudinal plots, with the option to download the plots, and finally, data was also made available in a tabular format.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

Both AMP PD and GP2 data were available for access through the AMP PD website (https://amp-pd.org/).

CODE AVAILABILITY

All code, pipelines, and applications that were used and produced are available on GitHub via the provided links in this manuscript.

Received: 7 July 2022; Accepted: 13 February 2023;
Published online: 04 March 2023

REFERENCES

1. International Parkinson Disease Genomics Consortium (IPDGC). Ten years of the International Parkinson Disease Genomics Consortium: progress and next steps. *J. Parkinsons. Dis.* **10**, 19–30 (2020).
2. Global Parkinson's Genetics Program. GP2: the global Parkinson's genetics program. *Mov. Disord.* **36**, 842–851 (2021).
3. Mc Cartney, A. M. et al. An international virtual hackathon to build tools for the analysis of structural variants within species ranging from coronaviruses to vertebrates. *F1000Res* **10**, 246 (2021).
4. Toga, A. W. & Crawford, K. L. The Alzheimer's disease neuroimaging initiative informatics core: a decade in review. *Alzheimers Dement.* **11**, 832–839 (2015).
5. Iwaki, H. et al. Accelerating medicines partnership: Parkinson's disease. Genetic resource. *Mov. Disord.* **36**, 1795–1804 (2021).
6. "Terra." n.d. Accessed 1 March 2023. https://app.terra.bio/.
7. Nalls, M. A. et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).
8. Foo, J. N. et al. Identification of risk loci for Parkinson disease in Asians and comparison of risk between Asians and Europeans: a genome-wide association study. *JAMA Neurol.* **77**, 746–754 (2020).
9. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
10. Yang, J. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).
11. 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).
12. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
13. Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, W199–W205 (2019).
14. Fisher, R. A. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edin.* **52**, 399–433 (1918).
15. Carlborg, O. & Haley, C. S. Epistasis: too often neglected in complex trait studies? *Nat. Rev. Genet.* **5**, 618–625 (2004).
16. Ritchie, M. D. & Van Steen, K. The search for gene-gene interactions in genome-wide association studies: challenges in abundance of methods, practical considerations, and biological interpretation. *Ann. Transl. Med.* **6**, 157 (2018).
17. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* **16**, 85–97 (2015).
18. Bandres-Ciga, S., Diez-Fairen, M., Kim, J. J. & Singleton, A. B. Genetics of Parkinson's disease: an introspection of its journey towards precision medicine. *Neurobiol. Dis.* **137**, 104782 (2020).
19. Bandres-Ciga, S. et al. Large-scale pathway specific polygenic risk and transcriptomic community network analysis identifies novel functional pathways in Parkinson disease. *Acta Neuropathol.* **140**, 341–358 (2020).
20. Quan, P. et al. Integrated network analysis identifying potential novel drug candidates and targets for Parkinson's disease. *Sci. Rep.* **11**, 13154 (2021).
21. Wallace, C. *coloc: Repo for the R package coloc*. (Github). (2022).
22. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
23. Sieberts, S. K. et al. Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions. *Sci. Data* **7**, 340 (2020).
24. "AD Knowledge Portal." n.d. Accessed 2 March 2023. https://adknowledgeportal.synapse.org/.
25. Drivas, T. G., Lucas, A. & Ritchie, M. D. eQTpLot: a user-friendly R package for the visualization of colocalization between eQTL and GWAS signals. *BioData Min.* **14**, 32 (2021).
26. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 1–12 (2019).
27. Traag, V. *leidenalg: Implementation of the Leiden algorithm for various quality functions to be used with igraph in Python*. (Github). (2020).
28. Kerimov, N. et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).
29. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
30. 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).
31. 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).
32. Pruim, R. J. et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
33. "Gas Power Calculator." n.d. Accessed 2 March 2023. https://csg.sph.umich.edu/abecasis/cats/gas_power_calculator/index.html.
34. "Google Analytics API into R." n.d. Accessed 2 March 2023. https://code.markedmondson.me/googleAnalyticsR/index.html.
35. Marek, K. et al. The Parkinson's progression markers initiative (PPMI) - establishing a PD biomarker cohort. *Ann. Clin. Transl. Neurol.* **5**, 1460–1477 (2018).
36. Rosenthal, L. S. et al. The NINDS Parkinson's disease biomarkers program. *Mov. Disord.* **31**, 915–923 (2016).
37. Kang, U. J. et al. The BioFIND study: characteristics of a clinically typical Parkinson's disease biomarker cohort. *Mov. Disord.* **31**, 924–932 (2016).
38. Web Application Framework for R [R package shiny version 1.7.1]. (2021).

ACKNOWLEDGEMENTS

This project was supported by the Global Parkinson's Genetics Program (GP2). GP2 is funded by the Aligning Science Against Parkinson's (ASAP) initiative and implemented by The Michael J. Fox Foundation for Parkinson's Research (https://gp2.org). For a complete list of GP2 members, see https://gp2.org. Data used in the preparation of this article were obtained from the AMP PD Knowledge Platform. For up-to-date information on the study, see https://www.amp-pd.org. AMP PD—a public-private partnership—is managed by the FNIH and funded by Celgene, GSK, the Michael J. Fox Foundation for Parkinson's Research, the National Institute of Neurological Disorders and Stroke, Pfizer, Sanofi, and Verily. This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging (NIA), National Institutes of Health, Department of Health and Human Services; project numbers ZO1 AG000535 and ZO1 AG000949, as well as the National Institute of Neurological Disorders and Stroke. Panel c of Fig. 3 was created with BioRender.com. Please see the supplemental material for IPDGC and GP2 member acknowledgements.

AUTHOR CONTRIBUTIONS

H.L.L., S.BC., A.J.N., M.A.N., S.D., and C.B. contributed to the concept and design of the study. H.L.L., R.M., A.MC., A.C.MN., AL.GM., A.I., A.M., B.I.B., B.J., B.H., C.S., C.T., D.V., D.C., E.Y., A.J., F.P.G., G.S., G.R., H.I., I.E., I.F.F., Z.JVR., J.J.K., J.Y., J.L., K.B., K.S., L.W., M.M.X.T., M.T.P., M.B.M., M.T., N.S.P., O.L.B., P.R.RP., P.A.J., P.S., R.AO., R.S., R.R., R.H.R., R.H., S.A., S.C.R., T.A., T.P.L., V.I., W.J.S., Y.S., A.S., M.A.N., S.D., S.BC., C.B., and A.J.N. were involved in the acquisition of data, data generation, and data cleaning. H.L.L., R.M., A.MC., A.C.MN., AL.GM., A.I., A.M., B.I.B., B.J., B.H., C.S., C.T., D.V., D.C., E.Y., A.J., F.P.G., G.S., G.R., H.I., I.E., I.F.F., Z.JVR., J.J.K., J.Y., J.L., K.B., K.S., L.W., M.M.X.T., M.T.P., M.B.M., M.T., N.S.P., O.L.B., P.R.RP., P.A.J., P.S., R.AO., R.S., R.R., R.H.R., R.H., S.A., S.C.R., T.A., T.P.L., V.I., W.J.S., Y.S., A.S., M.A.N., S.D., S.BC., C.B., and A.J.N. were involved in the creation of tools and pipelines. H.L.L., R.M., A.MC., A.C.MN., AL.GM., A.I., A.M., B.I.B., B.J., B.H., C.S., C.T., D.V., D.C., E.Y., A.J., F.P.G., G.S., G.R., H.I., I.E., I.F.F., Z.JVR., J.J.K., J.Y., J.L., K.B., K.S., L.W., M.M.X.T., M.T.P., M.B.M., M.T., N.S.P., O.L.B., P.R.RP., P.A.J., P.S., R.AO., R.S., R.R., R.H.R., R.H., S.A., S.C.R., T.A., T.P.L., V.I., W.J.S., Y.S., A.S., M.A.N., S.D., S.BC., C.B., and A.J.N. contributed to the drafting of the article and revising it critically.

FUNDING

Open Access funding provided by the National Institutes of Health (NIH).

COMPETING INTERESTS

H.L.L., D.V., H.I., Y.S., and M.A.N. declare no competing non-financial interests, but the following competing financial interests: H.L.L., D.V., H.I., Y.S., and M.A.N.'s participation in this project was part of a competitive contract awarded to Data Tecnica International LLC by the National Institutes of Health to support open science research. M.A.N. also currently serves on the scientific advisory board for Clover Therapeutics and is an advisor to Neuron23 Inc. R.M., A.MC., A.C.MN., AL.GM., A.I., A.M., B.I.B., B.J., B.H., C.S., C.T., D.C., E.Y., A.J., F.P.G., G.S., G.R., I.E., I.F.F., Z.JVR., J.J.K., J.Y., J.L., K.B., K.S., L.W., M.M.X.T.,

M.T.P., M.B.M., M.T., N.S.P., O.L.B., P.R.RP., P.A.J., P.S., R.AO., R.S., R.R., R.H.R., R.H., S.A., S.C.R., T.A., T.P.L., V.I., W.J.S., A.S., S.D., S.BC., C.B., and A.J.N. declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41531-023-00472-6.

Correspondence and requests for materials should be addressed to Hampton L. Leonard.

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.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023, corrected publication 2023