Global Parkinson's Genetics Program (GP2) Complex Network Protocol: Defining the causes of sporadic Parkinson's disease

Clodagh Towns [1], Madeleine Richer [1], Simona Jasaityte [1], Eleanor Stafford [1], Julie Joubert [1], Tarek Antar [2], Alejandro Carrasco-Martinez [1], Mary B. Makarious [1,5], Bradford Casey [8], Dan Vitale [3,4], Kristin Levine [3,4], Hampton Leonard [3,4,5], Claire E. Wegel [7], Justin Solle [8], Mike A. Nalls [2,3,4], Cornelis Blauwendraat [2,4], Andrew B. Singleton [2,4], Manuela M. X. Tan [6], Hirotaka Iwaki [2,3,4], Huw R. Morris [1]* *and the Global Parkinson's Genetics Program (GP2)*

1. Department of Clinical and Movement Neurosciences, Queen Square Institute of Neurology, University College London, London, UK
2. Integrative Genomics Unit, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA
3. Data Tecnica International, Washington, DC, USA
4. 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
5. Molecular Genetics Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA
6. Department of Neurology, Oslo University Hospital, Oslo, Norway
7. Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA
8. Department of Clinical Research, Michael J. Fox Foundation for Parkinson's Research, New York City, NY, USA

Corresponding author:

Professor Huw R Morris, Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, Royal Free Hospital, Rowland Hill Street, London NW3 2PF, h.morris@ucl.ac.uk.

Abstract

The Global Parkinson's Genetics Program (GP2) will genotype over 150,000 participants from around the world, and integrate genetic and clinical data for use in large-scale analyses to dramatically expand our understanding of the genetic architecture of PD. This report details the workflow for cohort integration into the complex arm of GP2, and together with our outline of the monogenic hub in a companion paper, provides a generalizable blueprint for establishing large scale collaborative research consortia.

Main Text

Parkinson's disease (PD) is a multifactorial disorder with complex etiology. The largest genome-wide association study (GWAS) to-date included 37,688 cases and 18,618 proxy cases (unaffected first-degree relatives) from European ancestry, and identified 90 independent risk signals across 78 genomic regions; 38 of which were novel signals[1]. Despite these advances, PD GWAS are currently limited by scale, a focus on European populations, and limited integration with clinical phenotype data.

A power calculation based on the 2019 GWAS data indicates that inclusion of an additional ~99K cases would enable variants of smaller effect size that contribute to polygenic risk (*p*-value cut off: 1.35×10^{-3}) to reach genome-wide significance. Therefore, expanding PD GWAS to at least this size will result in identification of additional risk loci and improve genetic prediction of PD occurrence[1]. The variability of phenotypes observed in PD are likely to have a genetic basis.[2,3,4] Knowledge of associations between genotype and clinical outcomes will enable clinicians to provide patients with a more accurate prognosis. Understanding the gene-to-phenotype pathways responsible for specific PD features would provide an opportunity to develop treatments targeting phenotype-specific disease pathways, resulting in more efficient and personalized treatment with fewer side effects. We aim to capture the diversity of PD

outcomes, including Parkinson's itself but also related conditions such as dementia with Lewy bodies and parkinsonisms, and to perform large-scale analyses of clinical-genetic data with sufficient power for gene discovery. The focus of PD GWAS on individuals of European ancestry has left gaps in our knowledge of PD-associated genetic variants in underrepresented populations and limited our ability to resolve GWAS loci [5]. To advance understanding of the genetic determinants of PD on a global scale, we need to ensure representation of diverse ancestries in future GWAS, with sample sizes sufficient to detect ancestry-specific signals.

The Global Parkinson's Genetics Program (GP2, http://gp2.org/), funded by the Aligning Science Across Parkinson's initiative (ASAP, https://parkinsonsroadmap.org/) will recruit PD cohorts from across the world [6]. We will genotype >150,000 participants and integrate genetic data with harmonized clinical data for use in case-control and genotype-phenotype association studies. This figure will include a minimum of 50,000 individuals from ancestries currently underrepresented in PD research: including Black American, African, Middle Eastern, Central/East/South Asian, Indian, Caribbean and Central/South American. The remaining participants are expected to be of European ancestry.[6] Cleaned genetic and clinical data will be harmonized across cohorts and made available to PD researchers via a controlled-access online repository with a unified user agreement. Contributing investigators are encouraged to play an active role in the project by suggesting and leading analyses, and will receive authorship on GP2 publications as well as support for their analyses. GP2 also provides comprehensive training opportunities for researchers from contributing institutions via an online learning management system. The outcome of GP2 will be an open access resource which will integrate clinical and genetic data from the largest, most diverse sample of PD cases to-date, and facilitate discovery of novel genetic determinants of PD occurrence and phenotypic variation in multiple ancestries. A global network of PD researchers will be established, facilitating future collaboration and advancement of PD research. The project began in January 2020 and current funding extends to 2029.

Coordinating cohorts around the world to include 150,000 participants is a considerable logistical challenge. As of May 2022, we have established a workflow for cohort integration, a standardized set of clinical data elements, a recommended consent template, a protocol for evaluating cohorts joining the study and a process for integrating and harmonizing the incoming clinical data. More broadly we have established the complex hub framework, which created a code of conduct and publication policy, created a DNA quality and shipping protocol, and created the NeuroBooster Array (NBA) genotyping workflow.

The size and depth of the dataset which is being generated by GP2 will provide unprecedented opportunities to discover novel genotype-phenotype associations. Core analyses, analyses that will be continuously updated by the analysis teams at GP2 with the inclusion of new samples, will include case-control, age at onset, and progression GWAS, as well as within-ancestry analyses of previously underrepresented populations. Beyond this, GP2 is supporting additional analytical projects proposed by contributing collaborators. The GP2 network will also facilitate collaboration on auxiliary studies between investigators sharing the same area of interest (e.g., biomarkers of different modalities) to address outstanding questions in PD research.

GP2 is conducted according to overarching principles of democratization of data, collaboration and cooperation, safe and responsible data sharing, commitment to diversity in research, transparency and reproducibility, and production of an actionable resource in accordance with the Findable, Accessible, Interoperable and Reusable (FAIR) scientific data management principles.[7] Here, we report the specific steps undertaken by the Cohort Integration Working Group (CIWG) to identify, recruit, and harmonize clinical cohorts within the complex disease arm of GP2. GP2 also has a monogenic arm which focuses on cases with potential monogenic

causes of PD, i.e. those with early age at onset or a family history of PD (*monogenic protocol reference*).

Methods

PD investigators and cohorts are identified through relevant publications, or through existing PD consortia. We welcome contact from interested investigators around the world (Email: cohort@gp2.org). Prospective collaborators complete a Site Interest Form (SIF) (Supplementary Material) to provide a brief overview of their cohort. Data availability, type of cohort (e.g., brain bank, drug study etc.), ethnic diversity and sample/data sharing restrictions are all considered by the CIWG when reviewing SIF submissions. Cohorts' consent documents are reviewed by the Operations and Compliance Working Group (OCWG) to ensure consent for sample/data sharing was obtained, and compliance with regional and cohort-specific data sharing restrictions. The OCWG assists the investigator if revisions are needed, and template consent language is provided online (https://gp2.org/resources/consent-guidelines/ [Supplementary Material]). We have an inclusive approach; to-date >95% of cohorts evaluated have been accepted. Following cohort approval, a concise Collaboration Agreement is signed by the collaborating PI and the Michael J. Fox Foundation (MJFF), and the necessary material and data transfer agreements are completed. The cohort's samples are then transferred to one of GP2's genotyping centers, and data is transferred through secure upload to a cloud server.

Once the cohort's clinical data has been transferred, harmonization and quality control (QC) are performed by the CIWG, following a standard framework.[8] Core clinical data were decided on the basis of clinical scales available in existing PD cohorts (primarily the Tracking Parkinson's [9], PPMI[10], PDBP[11], NET PD-LS1[12] and GEoPD[13] cohorts) as well as on the basis of recent proposals for a modular set of assessment criteria to characterize longitudinal PD cohorts.[14,15] The core clinical data elements are the clinical data categories of interest which will be

harmonized across cohorts and used for analysis (Table 1). We have generated a common dataset for brain banks, and have added data elements for time to event analysis (Supplementary Table 1). We are also collecting information on availability of other data and samples, including, plasma, serum, RNA, fibroblasts, CSF, skin biopsy, PBL, iPS to expand awareness of the resources that are available at contributing sites (Supplementary Material). The common dataset is defined in a common data dictionary (Supplementary Table 2). The harmonization of raw data uses a set of coding rules for re-coding and handling of missing data, using a modified custom script for each cohort.

DNA samples from each cohort are genotyped on the Neuro Booster Array (NBA), developed in collaboration by Data Tecnica International, NIA, NINDS, Illumina Inc and GP2. The NBA consists of the Illumina Infinium Global Diversity Array (GDA), a high-density (1.9M total variants) global backbone optimized for cross-population imputation coverage of the human genome, and additional custom content (>95K variants) which includes known causal variants for various neurodegenerative diseases and imputation boosters for underrepresented populations. QC of genotype data by the Data Analysis Working Group (DAWG) follows a standard pipeline. A custom genotype clustering file is used to ensure representation of the diverse genetic ancestries within the GP2 data. The clustering file used for the latest data release is based on 2,793 samples across 6 ancestry groups, and includes 420 Gaucher disease cases to capture variants of interest in the *GBA* risk gene. This file is available for download via the GP2 Github repository. Cleaned genetic data is returned to the contributing investigator to use as they wish, and is uploaded to the GP2 repository for use in combined analyses (Figure 1). GP2 covers the costs of genotyping and sample shipment for all contributing cohorts, and can assist investigators with the analysis of their cohort's data.

The dataset which we are currently aggregating and harmonizing contains a wide variety of demographic and clinical factors, thanks to the diversity of contributing cohorts. As of May 2022,

the CIWG has approved 113 cohorts for inclusion in GP2, of which 102 have been approved by the OCWG, and 43 have completed all necessary agreements and are in the process of transferring samples and clinical data. Samples that have been transferred are currently being genotyped and passed through the data QC pipeline. So far, the approved cohorts span over 50 different countries and territories. A map showing the geographic distribution of these cohorts, as well as current expected and completed sample numbers, can be found on the GP2 website (https://gp2.org/cohort-dashboard/).

Data Availability

GP2 has partnered with the Accelerating Medicines Partnership - Parkinson's Disease (AMP-PD; https://amp-pd.org) to share data generated by GP2, and in December 2021 the first GP2 genotyping data was released on the AMP-PD platform. As of 2022, the data consists of 14,902 samples (8,190 PD cases), representing a broad range of diverse ancestries defined directly from the genotyping data from these cohorts. Genotyping and data QC is ongoing, and there will be regular data releases (2-4 times per year) as the project progresses.

The Data and Code Dissemination Working Group (DCDWG) at GP2 is committed to making these data, resources, and training materials accessible to the scientific community. The DCDWG actively supports efforts to consolidate analysis scripts with the necessary analytical tools and decisions via GitHub (https://github.com/GP2code), consolidate training materials with several translations in collaboration with the Training and Networking Working Group (https://gp2.org/training/), and consolidate additional analytical pipelines on Terra (https://amp-pd.org/tools). All contributing investigators have access to GP2 data, and external researchers can also gain access by following instructions on the GP2 website (https://gp2.org/applying-for-gp2-data-access-on-the-amp-pd-platform/). There are two tiers of data access. Tier 1 consists

of summary statistics and any researcher can gain access by completing an online application. Tier 2 access includes deidentified, individual level genetic data, and to gain access researchers must sign a Data Usage Agreement co-signed by their institution.

Acknowledgements:

This research is supported by the Aligning Science Across Parkinson's Initiative, the Intramural Research Program, National Institute on Aging, National Institutes of Health, Department of Health and Human Services, project ZO1 AG000949, and the Michael J. Fox Foundation for Parkinson's Research.

Data used in the preparation of this article were obtained from 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.

Author Contributions

C. Towns, M. Richer, S. Jasaityte, E. Stafford, J. Joubert, T. Antar, A. Carrasco-Martinez, M. Tan, H. Iwaki., and H. Morris are members of the GP2 Cohort Integration Working Group (CIWG), of which H.Morris is the lead, and M. Tan and H. Iwaki are co-leads. C.Towns was the primary contributor in the drafting of this manuscript, assisted by M. Richer, and J. Joubert. The draft was reviewed by all CIWG members prior to circulation to other contributing authors for review. M. Makarious is the co-lead of the GP2 Data and Code Dissemination Working Group and drafted the data availability section of this manuscript, together with B. Casey who is senior associate director at the Michael J. Fox Foundation's research division. D. Vitale, K. Levine, H. Leonard and M. Nalls are members of the GP2 Complex Disease Data Analysis Working Group (DAWG), of

which M. Nalls is the lead and H. Leonard is a co-lead. DAWG are responsible for the data cleaning and analysis described in this manuscript. C. Wegel and J. Solle are members of the GP2 Operations and Compliance Working Group of which J. Solle is a co-lead. A. Singleton is the lead of the GP2 Complex Disease Network and C. Blauwendraat is the co-lead. All listed authors reviewed the manuscript and provided comments and revisions prior to submission.

Competing Interests statement

A. Singleton and C. Blauwendraat are supported by the Intramural Research Program of the National Institute on Aging and have received grant support from the Michael J. Fox Foundation for Parkinson's Research and the Aligning Science Across Parkinson's Initiative. A. Singleton has received royalty payments related to a diagnostic for stroke.

H. Leonard, H. Iwaki, D. Vitale, K. Levine and M. Nalls are consultants employed by Data Tecnica International. Data Tecnica is engaged in a consulting agreement with the US National Institutes of Health.

H. Morris is employed by UCL. In the last 24 months, he reports paid consultancy from Biogen, Biohaven, Lundbeck; lecture fees/honoraria from Wellcome Trust, Movement Disorders Society; Research Grants from Parkinson’s UK, Cure Parkinson’s Trust, PSP Association, CBD Solutions, Drake Foundation, Medical Research Council, Michael J Fox Foundation. H. Morris is also a co-applicant on a patent application related to C9ORF72—Method for diagnosing a neurodegenerative disease (PCT/GB2012/052140).

B. Casey and J. Solle are employed by the Michael J. Fox Foundation for Parkinson's Research.

References:

1. 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).
2. Iwaki, H. *et al.* Genetic risk of Parkinson disease and progression: An analysis of 13 longitudinal cohorts. *Neurol. Genet.* **5,** 1–14 (2019).
3. Latourelle, J. C. *et al.* Large-scale identification of clinical and genetic predictors of Parkinson's disease motor progression in newly-diagnosed patients: a longitudinal cohort study and validation. *Lancet Neurol.* **16,** 908–916 (2017).
4. Tan, M. M. X. *et al.* Genome-Wide Association Studies of Cognitive and Motor Progression in Parkinson's Disease. *Mov. Disord.* **36,** 424–433 (2021).
5. Heinzel, S., Lerche, S., Maetzler, W. & Berg, D. Global, Yet Incomplete Overview of Cohort Studies in Parkinson's disease. *J. Parkinsons. Dis.* **7,** 423–432 (2017).
6. The Global Parkinson's Genetics Program (GP2). GP2: The Global Parkinson's Genetics Program. *Mov. Disord.* **36,** 842–851 (2021).
7. Wilkinson M.D., *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. **3,** 160018 (2016).
8. Bennett, S. N. *et al.* Phenotype harmonization and cross-study collaboration in GWAS consortia: The GENEVA experience. *Genet. Epidemiol.* **35,** 159–173 (2011).
9. Malek, N. *et al.* Tracking Parkinson's: study design and baseline patient data. *J. Parkinsons. Dis.* **5,** 947–959 (2015).
10. Parkinson Progression Marker Initiative. The Parkinson Progression Marker Initiative (PPMI). *Prog. Neurobiol.* **95,** 629–635 (2011).
11. Rosenthal, L. S. *et al.* The NINDS Parkinson's disease biomarkers program. *Mov. Disord.* **31,** 915–923 (2016).
12. Kieburtz, K. *et al.* Effect of Creatine Monohydrate on Clinical Progression in Patients With

Parkinson Disease. *JAMA.* **313,** 584–593 (2015).

13. Bardien, S., Mellick, G. D., Hattori, N., Ross, O. A. & Lesage, S. *Celebrating the Diversity of Genetic Research to Dissect the Pathogenesis of Parkinson's Disease* (Frontiers Media SA, Lausanne, 2021).

14. Lerche, S. *et al.* Aiming for study comparability in Parkinson's disease: Proposal for a modular set of biomarker assessments to be used in longitudinal studies. *Front. Aging Neurosci.* **8,** 1–4 (2016).

15. Lerche, S. *et al.* Methods in Neuroepidemiology Characterization of European Longitudinal Cohort Studies in Parkinson's Disease-Report of the JPND Working Group BioLoC-PD. *Neuroepidemiology* **45,** 282–297 (2015).

16. Grinnon, S. T. *et al.* National Institute of Neurological Disorders and Stroke Common Data Element Project - Approach and methods. *Clin. Trials* **9,** 322–329 (2012).

17. Postuma, R. B. *et al.* MDS clinical diagnostic criteria for Parkinson's disease. *Mov. Disord.* **30,** 1591–1601 (2015).

18. Gibb W.R.G., Lees A.J. The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease. *J Neurol Neurosurg Psychiatry* **51,** 745-752 (1988).

19. Martínez-Martín, P. *et al.* Global versus factor-related impression of severity in Parkinson's disease: A new clinimetric index (CISI-PD). *Mov. Disord.* **21,** 208–214 (2006).

20. Goetz, C. G. *et al.* Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): Process, format, and clinimetric testing plan. *Mov. Disord.* **22,** 41–47 (2007).

21. Fahn S., Elton R., Members of the UPDRS Development Committee. In: Fahn S., Marsden C.D., Calne D.B., Goldstein M., Editors. *Recent Developments in Parkinson's Disease*, *Vol* ***2****, p. 153-163, 293-304*. (Macmillan Health Care Information, New Jersey, 1987).

22. Folstein, M. F., Folstein, S. E. & McHugh, P. R. 'Mini-mental state'. A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* **12,** 189–198 (1975).

23. Nasreddine, Z. S. *et al.* The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* **53,** 695–699 (2005).

24. Hoehn, M. M. & Yahr, M. D. Parkinsonism: onset, progression and mortality. *Neurology* **17,** 427–442 (1967).

25. Visser, M., Marinus, J., Stiggelbout, A. M. & van Hilten, J. J. Assessment of autonomic dysfunction in Parkinson's disease: The SCOPA-AUT. *Mov. Disord.* **19,** 1306–1312 (2004).

26. Stiasny-Kolster, K. *et al.* The REM sleep behavior disorder screening questionnaire - A new diagnostic instrument. *Mov. Disord.* **22,** 2386–2393 (2007).

27. Johns, M. W. A new method for measuring daytime sleepiness: The Epworth sleepiness scale. *Sleep* **14,** 540–545 (1991).

28. Yesavage, J. A. & Sheikh, J. I. Geriatric Depression Scale (GDS) recent evidence and development of a shorter version. *Clin. Gerontol.* **5,** 165–173 (1986).

29. Doty, R. L., Shaman, P. & Dann, M. Development of the University of Pennsylvania Smell Identification Test: a standardized microencapsulated test of olfactory function. *Physiol. Behav.* **32,** 489–502 (1984).

30. Wolfensberger, M. Sniffin'Sticks: a New Olfactory Test Battery. *Acta Otolaryngol.* **120,** 303–306 (2000).

31. Doty, R. L., Marcus, A. & Lee, W. W. Development of the 12-item Cross-Cultural Smell Identification Test (CC-SIT). *Laryngoscope* **106,** 353–356 (1996).

32. Schwab, R.S. & England, A.C. Projection Technique for Evaluating Surgery in Parkinson's Disease. In: Billingham, F.H. and Donaldson, M.C., Editors. *Third Symposium on Parkinson's Disease p. 152-157*. (Churchill Livingstone, Edinburgh, 1969).

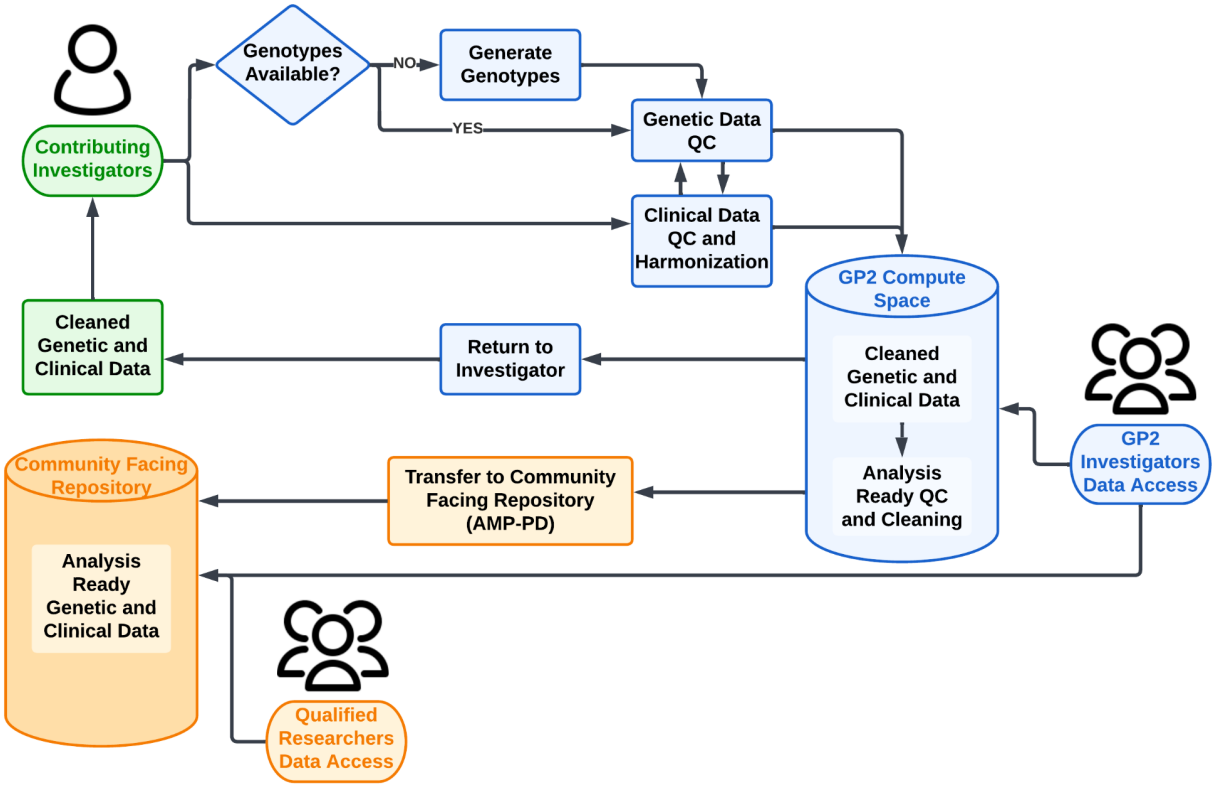
33. Jenkinson, C., Fitzpatrick, R., Peto, V., Greenhall, R. & Hyman, N. The PDQ-8: Development and validation of a short-form Parkinson's disease questionnaire. *Psychol. Health* **12,** 805–814 (1997).

34. Chaudhuri, K. R. *et al.* King's Parkinson's disease pain scale, the first scale for pain in PD: An international validation. *Mov. Disord.* **30,** 1623–1631 (2015).

Table 1. Recommended Core Clinical Data

	MINIMUM	MINIMUM PLUS	CORE	EXTENDED
Demographics *(e.g., NINDS CDE General Core)*[16]	✓	✓	✓	✓
Recruitment category *(Case / Control)*	✓	✓	✓	✓
Family History	✓	✓	✓	✓
Diagnostic checklist *(MDS clinical diagnosis criteria*[17] *or UK Parkinson's Disease Society Brain Bank Diagnostic Criteria)*[18]		✓	✓	✓
Primary diagnosis and PD certainty *(PPMI Primary Clinical Diagnosis)*[10]		✓	✓	✓
PD history *(e.g., NINDS CDE for Parkinson's Disease)* [16]		✓	✓	✓
Global PD severity *(CISI-PD)*[19]		✓	✓	✓
Behavioral and Environmental History *(PD-RFQ-U)*[16]			✓	✓
Medical History			✓	✓
Current Medication Status			✓	✓
nM-EDL *(MDS-UPDRS Part I [or UPDRS])*[20,21]			✓	✓
M-EDL *(MDS-UPDRS Part II [or UPDRS])*[20,21]			✓	✓
Motor Assessment *(MDS-UPDRS Part III [or UPDRS])* [20,21]			✓	✓
Complications *(MDS-UPDRS Part IV [or UPDRS])* [20,21]			✓	✓
Cognitive Assessment *(MMSE, MoCA)*[22, 23]			✓	✓
Motor *(Hoehn and Yahr stage)*[24]			✓	✓
Autonomic function assessment *(SCOPA-AUT)*[25]			✓	✓
pRBD *(RBDSQ)*[26]			✓	✓
Day time sleepiness *(ESS)*[27]				✓
Depression *(GDS-15)*[28]				✓
Orthostatic hypotension *(Standing/sitting/supine HR and BP)*				✓
Olfactory function *(UPSIT or Sniffin-stick or BSIT)*[29,30,31]				✓
General ADL *(Schwab & England ADL)*[32]				✓
PD EQL *(PDQ-8)*[33]				✓
Pain *(Kings PD Pain Scale)*[34]				✓

Figure 1. Cohort Integration Workflow



Data is contributed and returned to the local PI (green); new genetic data, data cleaning and harmonization is carried out by GP2 and made available for analysis within the GP2 consortium (blue); and data is released to qualified investigators via AMP-PD (orange).

AMP-PD: Accelerating Medicines Partnership - Parkinson's Disease (https://amp-pd.org)