The Components of GP2’s Fifth Data Release

Tags: Research Operations; Research Collaboration; Complex Disease Genetics

Authors:

Hampton Leonard
Data Tecnica International/National Institutes of Health | USA

Hampton has a background in data science and machine learning, which she applies to large multi-omic datasets in the neurodegenerative disease space. She is passionate about investigating differences on both clinical and omic levels and how these differences can affect clinical trial outcomes.

Mike Nalls
Data Tecnica International/National Institutes of Health | USA

Mike founded Data Tecnica in early 2017 after over a decade of experience in large dataset analytics and methods research in healthcare and other scientific fields. 400+ peer-reviewed publications in the field of applied statistics in large datasets, brain diseases, and genomics. He is a strong advocate of open science, collaboration, and transparency in science.

Dan Vitale
Data Tecnica International/National Institutes of Health | USA

Dan is a data science consultant for Data Tecnica International, consulting primarily for the Laboratory of Neurogenetics and CARD at the National Institute on Aging of the National Institutes of Health. His work is focused on open science, automation and development of genetic analytic pipelines and software, and machine learning.

Mathew Korestsky
National Institutes of Health | USA

Mat is a post-baccalaureate student at the National Institutes of Health. He is passionate about pipeline development and meaningful applications of computer science in the biomedical research space.

Kristin Levine
Data Tecnica International/National Institutes of Health | USA

Kristin works with the Data Tecnica and National Institute on Aging (NIA) teams on data and code sharing plus real-world data analysis of biobanks and healthcare systems. She is also an accomplished writer, now applying her communication skills to scientific domains.

Mary B Makarious
National Institutes of Health | USA

Mary is a graduate student participating in the NIH graduate partnership program in collaboration with the University College London. She is a rising star in biomedical data science, with a background in genomics and machine learning. She is also an experienced user of open science platforms like Terra.

Zih-Hua Fang
German Center for Neurodegenerative Diseases | Germany

The lead of the monogenic data analysis efforts in GP2, they are making significant contributions to GP2's efforts to study monogenic and familial Parkinson's disease.

On behalf of the GP2 Complex Disease Data Analysis, Monogenic Data Analysis, and Data and Code Dissemination Working Groups.

Blog post:

Overview:
In May 2023, GP2 announced the fifth data release on the Terra platform in collaboration with AMP® PD.

This release includes 7,462 additional new complex disease participants and 487 new monogenic disease participants, adding to the previous releases from the Complex and Monogenic Networks.

- **The complex disease data (genotypes) now consists of a total of 24,935 genotyped participants (12,728 PD cases, 10,533 Controls, and 1,674 'Other' phenotypes).**
- **The monogenic disease data (whole genome sequences) now consists of a total of 722 sequenced participants.**

Individual-Level Data:
New cohorts added to GP2 in this release are:

- Arizona Brain Bank Brain and body donation programme (BBDP), a US-based brain bank
- Innate and Adaptive Immunity in Parkinson Disease (IMMUNEPD), a US-based study from the University of Alabama at Birmingham
- IPDGC Africa (IPDGCAF-NG), a Nigerian cohort
- PDGENEration (PDGNRTN), a cohort from the Parkinson's Foundation
- Genetic study of Parkinson's disease and related movement disorders (STELLENBOS), a South African cohort from Stellenbosch University

- Malaysian Parkinson's Genetics Cohort (UMKLM), a Malaysian cohort from the University of Malaya
- Genotype-Phenotype Correlations in Parkinson’s Disease and Related Movement Disorders (KUL), a monogenic cohort from the University of Malaysia
- Parkinson Disease and Movement Disorders Center Biorepository (PDMDC), a monogenic cohort from Northwestern University

Genetically-determined ancestry of complex disease GP2 participants is broken into ten ancestry groups; the table below details the genetically-determined ancestry of complex disease participants in this release that have passed quality control and been imputed. These numbers include samples from previous releases that have been reclustered using the new cluster file and gone through quality control along with the newly genotyped and shared samples unique to this current release.

Complex Disease - GP2 Release 5				
Ancestry	**Total**	**PD**	**Control**	**Other**
African	2,353	855	1,488	10
African Admixed	1,331	289	1,031	11
Ashkenazi Jewish	1,546	740	386	420
Latino and Indigenous people of the Americas	573	345	205	23
East Asian	2,913	878	2,022	13
European	15,356	9,230	4,966	1,160
South Asian	220	80	135	5
Central Asian	478	195	279	4
Middle Eastern	149	106	18	25
Finnish	16	10	3	3
Total	**24,935**	**12,728**	**10,533**	**1,674**

Important to note in this release: STELLENBOS is a South African cohort and displays genetic ancestry admixture that reflects that region and its history. Some of our ancestry predictions for members of this study may change as we adjust for this new population in future releases. In this release we have provided admixture estimates for all participants so that researchers can make choices about if and how to include high admixture in their analyses as part of the tier 2 data. We have also included a list of samples that we recommend removing from analyses at this stage as their high levels of admixture lead to lower confidence in their ancestry predictions from our current model.

Genetically-determined ancestry of monogenic disease GP2 participants is estimated from the same pipeline as the complex participants. The table below details the genetically-determined ancestry of monogenic disease participants in GP2 release 5.

Monogenic Disease - GP2 Release 5				
Ancestry	**Total**	**PD**	**Control**	**Other**
African	169	109	60	0
African Admixed	37	32	5	0
Latino and Indigenous people of the Americas	12	12	0	0
European	432	401	21	8
Central Asian	74	71	3	0
Total	**722**	**625**	**89**	**8**

Future data releases will continue to grow the diversity of participants available. You can check out our dashboard to see our progress [https://gp2.org/cohort-dashboard/].

An important update for this release is a minor change to how we determine which participants are related. We are now using KING software [https://www.kingrelatedness.com/]. This change may have updated estimates of relatedness for some of our previously released participants. With this change, a file that details the estimated relatedness thresholds will now be made available at the release for researchers to use for their own custom filtering purposes. In addition to the projected principal components, we now include principal components calculated per ancestry available for downstream analyses.

Summary Statistics:
We are pleased to announce that new summary statistics are also available through tier 1 access for this release. Kim et al 2023 are also available through our ongoing collaboration with Neurodegenerative Disease Knowledge Portal [https://ndkp.hugeamp.org/].

- Nalls et al 2019 without 23andme in hg38 (largest European PD GWAS)
- Loesch et al 2021 (largest Latin American/Amerindian PD GWAS)
- Kim et al 2023 without 23andme (Multi-ancestry meta-analysis for PD risk including European, East Asian, Latin American/Amerindians, and African Americans)
- Rizig et al 2023 without 23andme in hg38 (largest African and African admixed PD GWAS; using GP2 release 5 data)

Copy Number Variants:
Copy number variant (CNV) calls for all genotyped samples passing quality control (gene-level plus 250kb flanking regions) have been updated to include all samples in release 5. This data has been clustered using a custom GP2 genotype clustering file (available in the **utils** directories under both tier 1 and tier 2 data access). Both the cluster file and the pipeline used to predict the probabilistic CNV calls can be found on the GP2 Github [https://github.com/GP2code] for use with data outside of GP2. For more information regarding clustering using the custom GP2 genotyping clustering file and the copy number variant probabilistic calls, please see the "The Components of GP2's Third Data Release" on the GP2 blog.

More information on the structure of the complex disease genotype and clinical data is detailed in the blog post 'The Components of GP2's First Data Release' [https://gp2.org/the-components-of-gp2-first-data-release/] as well as in the README that has been updated for this release and is available on the official GP2 Terra workspaces. The monogenic PD WGS data is also detailed in the same README.

As always, please refer to the README that accompanies each GP2 release for further details regarding pipelines, data, and analyses!