Last updated: February 14th 2023

Storage updates and release schedules: We intend to release updates each quarter to stored data as scheduled releases. This document describes the contents of the top level directories within the GP2 Tier 1 (gp2tier1) and Tier 2 (gp2tier2) accessible cloud storage buckets that contain public summary level and private participant level data, respectively.

For example, release 4 on February 14th 2023 would be in the top level directory /release4_14022023 in both the tier1 and tier2 storage buckets.

Contact: For questions relating to data processing, please email admin@gp2.org.

Release specific info follows below.

##################
Current Release
##################

Release4_14022023

For a current list of samples, studies, cohorts and geographic territories covered by GP2 please see the GP2 website here [https://gp2.org/cohort-dashboard/].

For more information regarding this release, please check out the GP2 blog post under the title 'Components of GP2's Fourth Data Release' : [https://gp2.org/blog/]

For this release, 49,754 consistently poor performing SNPs from the genotyping array were removed before the QC process. A list of these variants (underperforming_GP2_SNPs.csv) is now located in the meta_data directory of the Tier 2 buckets and on Github [https://github.com/GP2code/releases]. At processing, a less stringent sample call rate threshold was used, changing from 0.98 to 0.95. This way, more samples can be shared without sacrificing quality. The clinical metadata now includes more information about the non-PD diagnoses. An additional column has been added that includes additional information for anyone with a phenotype label of 'Other'. This additional information currently includes the following diagnoses: Genetically-enriched-cohort-Unaffected; Genetically-enriched-cohort-Affected; PSP; Not

Reported; DLB; MSA; SWEDD; Prodromal; Parkinsonism, Other; Other Neurological disorder(s), Unspecified; Essential Tremor; Vascular Parkinsonism.
The reference genotype and metadata for the ancestry inferences as part of the GenoTools pipeline [https://github.com/GP2code/GenoTools] can be found under **gp2tier2** in the **utils** directory. There may be some minimal shuffling of participants between ancestry labels as new releases are shared, due to the continued improvement of our ancestry prediction pipeline. As reference series availability grows, we will include more granular ancestry estimates in future releases.

Probabilistic estimates for copy number variations (CNVs) have been updated. You can find this pipeline under active development in the GP2 github repository [https://github.com/GP2code]. Please see the **Bucket and Directory Overview** below as well as the release's companion blog post for more detailed information on the CNVs.

Complex Disease
General Information:

- 2,583 samples are added in this release, the total number of shared GP2 samples now equals 17,485 (as of this release we have shared 9,429 PD cases, 6,648 controls, and 1,408 'Other').
- New genotype samples were processed using GenoTools version 0.1 [https://github.com/GP2code/GenoTools]. All samples were imputed to TOPMed reference detailed in the GenoTools pipeline.
- All data provided is GRCh38 (hg38).

GDPR note:

- Currently, all data included in this release has been determined to comply with GDPR guidelines, as it comes from countries not governed by GDPR or participants who are no longer living.

Bucket and Directory Structure:

gp2tier1 @release4_14022023
├── utils/
└── summary_statistics/

gp2tier2 @release4_14022023
├── raw_genotypes/

├── imputed_genotypes/
├── cnvs/
├── meta_data/
├── clinical_data/
├── wgs/
├── utils/
└── summary_statistics/

Bucket and Directory Overview:

- **gp2tier1**, this is the bucket for summary statistics and other non-participant level data. The top level directories for this always correspond to each release, with a mirrored structure for each release.
    - **utils** - Currently contains two separate directories: **illumina_utils** and **ref_panel** - the same resources included in tier 2 and detailed in the overview below
    - **summary_statistics** - The file META5_no23_with_rsids2.txt contains open access summary statistics from the most recent Parkinson's GWAS can be found here as well as in the tier 2 storage bucket. (This excludes 23andMe samples, from Nalls et al 2019, https://pubmed.ncbi.nlm.nih.gov/31701892/) Column headers conform to the standard METAL meta-analysis output [https://genome.sph.umich.edu/wiki/METAL_Documentation].

- **gp2tier2**, this is the bucket for participant level data. The top level directories for this always correspond to each release, with a mirrored structure for each release. Its content is mirrored below.
    - **raw_genotypes** - PLINK binary files for each ancestry group for all samples passing quality control prior to imputation. Each PLINK binary includes all attempted variants from the array for that ancestry group. As a note, for flexibility in community analyses, all known duplicate samples were removed but related samples remain.
    - **imputed_genotypes** - All genotype data has been imputed using the TOPMed reference panel and is contained in PLINK2 files separated by chromosome. Prior to upload, these files have been filtered for minor allele count > 10 and imputation quality > 0.3 as is industry standard. Each file set is separated by genetically defined ancestry groups prior to imputation. Workflow for ancestry determination and all other QC processes are found under [https://github.com/GP2code/GenoTools]. All ancestry specific

datasets show lambda values of less than 1.05 suggesting a high level of basic quality control.

- **cnvs** - probabilistic estimates of copy number variation per gene and +/- 250kb flanking regions for deletions, duplications and insertions. Estimates are not made available for all ancestries due to sample size restrictions, but this will improve as we grow our diversity. Code for these estimates can be found here [https://github.com/GP2code/]. It is a useful tool to prioritize potential insertion, duplication and deletions in genes of interest for follow-up studies.
- **meta_data** - Information in the meta_data directory includes: QC metrics, ancestry counts, predictive ancestry labels, confusion matrix, new samples UMAP, projected principal components, pruned samples, reference principal components, reference UMAP, and total (samples and reference) UMAP. It also includes:
  - **previous_release_samples** - This directory contains ID lists per ancestry group of all participants included in release 1, 2, and 3
  - **Relatedness -** New to this release, instead of having a separate file indicating related individuals, the clinical data now has a separate column reporting "related" individuals. 0 denotes unrelated participants, 1 for participants related on a cousin level or closer. In the next release, we will provide a key to better detail family structure.
  - **lambdas_for_QC.csv** - lambda and lambda1000 values per ancestry group for QC checking purposes.
  - **underperforming_GP2_SNPs.csv** - list of poor performing SNPs from the array that are removed from the QC process.
- **utils** - Currently contains two separate directories:
  - **illumina_utils** - contains SNP manifest files in .bpm and .csv formats as follows: NeuroBooster_20042459_A1.{bpm, csv} and NeuroBooster_20042459_A2.{bpm, csv} with A1 being GRCh37 and A2 being GRCh38. The A2 version was used for ALL analyses. This directory also contains the custom cluster file described above and in the blog post: recluster_09092022.egt.
  - **ref_panel** - contains the plink files for the reference panel used in the ancestry method in GenoTools along with a ref_panel_ancestry.txt which contains ancestry labels for each sample in the reference panel
- **clinical_data** - The corresponding data dictionary for an explanation of the columns can be found in release4_14022023_data_dictionary.csv. Quality control and predicted ancestry information per-sample has been provided

in the clinical data master key. An additional phenotype column "other_pheno" now expands on "Other" diagnoses.
    - **wgs** - Whole genome sequencing data from the Monogenic Network in PLINK binary and PLINK2 format as well as related metadata. The README_MonogenicWGS.md file contains more detailed information on the available data.
    - **summary_statistics** - this includes basic summary statistics from gp2tier1. More details above.

- **Ancestry group definitions**
    - AAC - African American / Caribbean ancestry
    - AFR - African ancestry
    - AJ - Ashkenazi Jewish ancestry
    - AMR - Latino and indigenous Americas populations
    - EUR - General European ancestry
    - EAS - East Asian ancestry
    - SAS - South Asian ancestry
    - FIN - Finnish population isolate
    - CAS - Central Asian ancestry
    - MDE - Middle Eastern ancestry

This release does not contain FIN (Finnish population isolate) due to insufficient sample size for accurate estimates of imputation quality.

######################
Previous Releases
######################

<u>Release3_31102022</u>

For a current list of samples, studies, cohorts and geographic territories covered by GP2 please see the GP2 website here [https://gp2.org/cohort-dashboard/].

For more information regarding this release, please check out the GP2 blog post under the title 'Components of GP2's Third Data Release' : [https://gp2.org/blog/]

For this release, sample genotypes were re-clustered using a custom, ancestry-aware cluster file to improve sample and variant call rates. The custom cluster file is available in

the utils directory under gp2tier2. It includes 2,793 samples across 6 ancestries with 420 Gaucher disease for better calling of known GBA risk variants.

This release adds Middle Eastern genetically-predicted ancestry in addition to more samples of African, African Admixed, Ashkenazi Jewish, Latino and Indigenous people of the Americas, East Asian, European, Finnish European, South Asian, and Central Asian ancestry. The reference genotype and metadata for the ancestry inferences as part of the GenoTools pipeline [https://github.com/GP2code/GenoTools] can be found under **gp2tier2** in the **utils** directory. As a note, some samples included previously in the East Asian or South Asian dataset have been reclassified as Central Asian with the inclusion of that genetic ancestry reference set. As reference series availability grows, we will include more granular ancestry estimates in future releases.

Probabilistic estimates for copy number variations (CNVs) have been updated. You can find this pipeline under active development in the GP2 github repository [https://github.com/GP2code]. Please see the **Bucket and Directory Overview** below as well as the release's companion blog post for more detailed information on the CNVs.

Complex Disease
General Information:

- 6,258 samples are added in this release, bringing the number of shared GP2 samples now equals 14,902 total samples (as of this release we have shared 8,190 PD cases, 6,712 non-PD).
- New genotype samples were processed using GenoTools version 0.1 [https://github.com/GP2code/GenoTools]. All samples were imputed to TOPMed reference detailed in the GenoTools pipeline.
- All data provided is GRCh38 (hg38).

GDPR note:

- Currently, all data included in this release has been determined to comply with GDPR guidelines, as it comes from countries not governed by GDPR or participants who are no longer living.

Bucket and Directory Structure:

gp2tier1 @release3_31102022
├── utils/

└── **summary_statistics/**

gp2tier2 @release3_31102022

├── **raw_genotypes/**
├── **imputed_genotypes/**
├── **cnvs/**
├── **meta_data/**
├── **clinical_data/**
├── **wgs/**
├── **utils/**
└── **summary_statistics/**

Bucket and Directory Overview:

- **gp2tier1**, this is the bucket for summary statistics and other non-participant level data. The top level directories for this always correspond to each release, with a mirrored structure for each release.
    - **utils** - Currently contains two separate directories: **illumina_utils** and **ref_panel** - the same resources included in tier 2 and detailed in the overview below
    - **summary_statistics** - The file META5_no23_with_rsids2.txt contains open access summary statistics from the most recent Parkinson's GWAS can be found here as well as in the tier 2 storage bucket. (This excludes 23andMe samples, from Nalls et al 2019, https://pubmed.ncbi.nlm.nih.gov/31701892/) Column headers conform to the standard METAL meta-analysis output [https://genome.sph.umich.edu/wiki/METAL_Documentation].

- **gp2tier2**, this is the bucket for participant level data. The top level directories for this always correspond to each release, with a mirrored structure for each release. Its content is mirrored below.
    - **raw_genotypes** - PLINK binary files for each ancestry group for all samples passing quality control prior to imputation. Each PLINK binary includes all attempted variants from the array for that ancestry group. As a note, for flexibility in community analyses, all known duplicate samples were removed but related samples remain.
    - **imputed_genotypes** - All genotype data has been imputed using the TOPMed reference panel and is contained in PLINK2 files separated by chromosome. Prior to upload, these files have been filtered for minor allele

count > 10 and imputation quality > 0.3 as is industry standard. Each file set is separated by genetically defined ancestry groups prior to imputation. Workflow for ancestry determination and all other QC processes are found under [https://github.com/GP2code/GenoTools]. All ancestry specific datasets show lambda values of less than 1.05 suggesting a high level of basic quality control.

- **cnvs** - probabilistic estimates of copy number variation per gene and +/- 250kb flanking regions for deletions, duplications and insertions for all samples. Code for these estimates can be found here [https://github.com/GP2code/]. It is a useful tool to prioritize potential insertion, duplication and deletions in genes of interest for follow-up studies.
- **meta_data** - Information in the meta_data directory includes: QC metrics, ancestry counts, predictive ancestry labels, confusion matrix, new samples UMAP, projected principal components, pruned samples, reference principal components, reference UMAP, and total (samples and reference) UMAP. It also includes:
  - **GP2_[ancestry]_release2_samples** - ID lists per ancestry group of all participants included in release 2
  - **GP2_round2_[ancestry]_release3.related** - ID lists per ancestry group of related participants
- **utils** - Currently contains two separate directories:
  - **illumina_utils** - contains SNP manifest files in .bpm and .csv formats as follows: NeuroBooster_20042459_A1.{bpm, csv} and NeuroBooster_20042459_A2.{bpm, csv} with A1 being GRCh37 and A2 being GRCh38. The A2 version was used for ALL analyses. This directory also contains the custom cluster file described above and in the blog post: recluster_09092022.egt.
  - **ref_panel** - contains the plink files for the reference panel used in the ancestry method in GenoTools along with a ref_panel_ancestry.txt which contains ancestry labels for each sample in the reference panel
- **clinical_data** - The corresponding data dictionary for an explanation of the columns can be found in release3_31102022_data_dictionary.csv. Quality control and predicted ancestry information per-sample has been provided in the clinical data master key.
- **wgs** - Whole genome sequencing data from the Monogenic Network in PLINK binary and PLINK2 format as well as related metadata. The README_MonogenicWGS.md file contains more detailed information on the available data.

    - **summary_statistics** - this includes basic summary statistics from gp2tier1. More details above.

- **Ancestry group definitions**
    - AAC - African American / Caribbean ancestry
    - AFR - African ancestry
    - AJ - Ashkenazi Jewish ancestry
    - AMR - Latino and indigenous Americas populations
    - EUR - General European ancestry
    - EAS - East Asian ancestry
    - SAS - South Asian ancestry
    - FIN - Finnish population isolate
    - CAS - Central Asian ancestry
    - MDE - Middle Eastern ancestry

This release does not contain FIN (Finnish population isolate) due to insufficient sample size for accurate estimates of imputation quality

Plink user note, GWAS-type analyses will be based on dosages, these analyses will have no missingness on the imputed genotype level as they treat imputed genotypes as a continuum from 0-2 copies of the effect allele per SNP. These dosages are non-integer as a means to account for uncertainty inherent in imputation. In some cases, integer genotypes are needed for analyses such as linkage calculations and similar. In analyses where integers are needed, you may encounter some degree of missigness. This is due to the allele dosage probabilities being outside the default tolerances of plink2 for calling an integer genotype. Consider this keeping your analyses cautious. If you wish to go against the suggestions of plink, there are methods built into plink2 to fill in any integer genotypes that did not pass the threshold.

Release2_06052022 (beta)

For a current list of samples, studies, cohorts and geographic territories covered by GP2 please see the GP2 website here [https://gp2.org/cohort-dashboard/].

For more information regarding this release, please check out the GP2 blog post under the title ‘Components of GP2’s Second Data Release’ : [https://gp2.org/blog/]

Complex Disease
General Information:

- 3,736 samples are added in this release, the number of shared GP2 samples now equals 8,644 (5,249 PD cases, 3,395 non-PD).
- New genotype samples were processed using GenoTools version 0.1 [https://github.com/dvitale199/GenoTools]. All samples were imputed to TOPMed reference detailed in the GenoTools pipeline.
- All data provided is GRCh38 (hg38).

GDPR note:

- Currently, all data included in this release has been determined to comply with GDPR guidelines, as it comes from countries not governed by GDPR or participants who are no longer living.

Bucket and Directory Structure:

gp2tier1 @release2_06052022
└── summary_statistics/

gp2tier2 @release2_06052022
├── raw_genotypes/
├── imputed_genotypes/
├── cnvs/
├── meta_data/
├── clinical_data/
├── wgs/
└── summary_statistics/

Bucket and Directory Overview:

- **gp2tier1**, this is the bucket for summary statistics and other non-participant level data. The top level directories for this always correspond to each release, with a mirrored structure for each release.
    - **summary_statistics** - The file META5_no23_with_rsids2.txt contains open access summary statistics from the most recent Parkinson's GWAS can be found here as well as in the tier 2 storage bucket. (This excludes 23andMe samples, from Nalls et al 2019, https://pubmed.ncbi.nlm.nih.gov/31701892/) Column headers conform to

the standard METAL meta-analysis output [https://genome.sph.umich.edu/wiki/METAL_Documentation].

- **gp2tier2**, this is the bucket for participant level data. The top level directories for this always correspond to each release, with a mirrored structure for each release. Its content is mirrored below.
    - **raw_genotypes** - PLINK binary files for each ancestry group for all samples passing quality control prior to imputation. Each PLINK binary includes all attempted variants from the array for that ancestry group. As a note, for flexibility in community analyses, all known duplicate samples were removed but related samples remain.
    - **imputed_genotypes** - All genotype data has been imputed using the TOPMed reference panel and is contained in PLINK2 files separated by chromosome. Prior to upload, these files have been filtered for minor allele count > 10 and imputation quality > 0.3 as is industry standard. Each file set is separated by genetically defined ancestry groups prior to imputation. Workflow for ancestry determination and all other QC processes are found under [https://github.com/dvitale199/GenoTools].
    - **cnvs** - probabilistic estimates of copy number variation per gene and +/- 250kb flanking regions for deletions, duplications and insertions for all samples. Code for these estimates can be found here [https://github.com/GP2code/GenoTools/tree/main/CNV]. This is currently "hypothesis generating" data and will be improved for next release.
    - **meta_data** - Information in the meta_data directory includes: QC metrics, ancestry counts, predictive ancestry labels, confusion matrix, new samples UMAP, projected principal components, pruned samples, reference principal components, reference UMAP, and total (samples and reference) UMAP. It also includes:
        - **GP2_[ancestry]_release1_samples** - ID lists per ancestry group of all participants included in release 1
        - **GP2_round2_[ancestry]_release2.related** - ID lists per ancestry group of related participants
    - **clinical_data** - The corresponding data dictionary for an explanation of the columns can be found in release2_26042022_data_dictionary.csv.
    - **wgs** - Whole genome sequencing data from the Monogenic Network in PLINK binary and PLINK2 format as well as related metadata. The README_MonogenicWGS.md file contains more detailed information on the available data.
    - **summary_statistics** - this includes basic summary statistics from gp2tier1

- **Ancestry group definitions**
    - AAC - African American / Caribbean
    - AFR - African ancestry
    - AJ - Ashkenazi Jewish
    - AMR - Latino and indigenous Americas populations
    - EUR - general European ancestry
    - EAS - East Asian ancestry
    - SAS - South Asian ancestry
    - FIN - Finnish population isolate
    - CAS - Central Asian

This release does not contain FIN (Finnish population isolate) due to insufficient sample size for accurate estimates of imputation quality

Release1_29112021 (alpha)

For a current list of samples, studies, cohorts and geographic territories covered by GP2 please see the GP2 website here [https://gp2.org/cohort-dashboard/].

General Information:

- 4908 samples are added in this release, the number of available GP2 samples now equals 4908 (3,434 PD cases, 1,474 non-PD).
- New genotype samples were processed using GenoTools version 0.1 [https://github.com/dvitale199/GenoTools]. All samples were imputed to TOPMed reference detailed in the GenoTools pipeline.
- All data provided is GRCh38 (hg38).

GDPR note:

- Currently, all data included in this release has been determined to comply with GDPR guidelines, as it comes from countries not governed by GDPR or participants who are no longer living.

Bucket and Directory Structure:

Gp2_tier1 @release1_29112021

```
└── summary_statistics/
```

Gp2_tier2 @release1_29112021

```
├── raw_genotypes/
├── imputed_genotypes/
├── meta_data/
├── clinical_data/
└── summary_statistics/
```

Bucket and Directory Overview:

- **gp2tier1**, this is the bucket for summary statistics and other non-participant level data. The top level directories for this always correspond to each release, with a mirrored structure for each release.
    - **summary_statistics** - The file META5_no23_with_rsids2.txt contains open access summary statistics from the most recent Parkinson's GWAS (excluding 23andMe samples, from Nalls et al 2019, https://pubmed.ncbi.nlm.nih.gov/31701892/) can be found here as well as in the tier 2 storage bucket. Column headers conform to the standard METAL meta-analysis output [https://genome.sph.umich.edu/wiki/METAL_Documentation].

- **gp2tier2**, this is the bucket for participant level data. The top level directories for this always correspond to each release, with a mirrored structure for each release. Its content is mirrored below.
    - **raw_genotypes** - PLINK binary files for each ancestry group for all samples passing quality control prior to imputation. Each PLINK binary includes all attempted variants from the array for that ancestry group. As a note, for flexibility in community analyses, all known duplicate samples were removed but related samples remain.
    - **imputed_genotypes** - All genotype data has been imputed using the TOPMed reference panel and is contained in PLINK2 files separated by chromosome. Prior to upload, these files have been filtered for minor allele count > 10 and imputation quality > 0.3 as is industry standard. Each file set is separated by genetically defined ancestry groups prior to imputation.
    - **meta_data** - Meta data included in the HDF5 file GP2_round1.QC.metrics.h5 is currently comprised of QC, ancestry counts, ancestry labels, confusion matrix, new samples UMAP, projected principal

components, pruned samples, reference principal component, reference UMAP, total (samples and reference) UMAP.
    - **clinical_data** - The corresponding data dictionary for an explanation of the columns can be found in release1_29112021_data_dictionary.csv.
    - **summary_statistics** - this includes basic summary statistics from gp2_tier1

- **Ancestry group definitions**
    - AAC - African American / Caribbean
    - AFR - African ancestry
    - AJ - Ashkenazi Jewish
    - AMR - Latino and indigenous Americas populations
    - EUR - general European ancestry
    - EAS - East Asian ancestry
    - SAS - South Asian ancestry
    - FIN - Finnish population isolate

This release does not contain AFR or FIN due to insufficient sample size for imputation quality