Release Notes

Release 3 (R03):

To see what's new in this Release, please refer to the Release Notes below and [GREGoR Dataset characteristics by release version](#). Additionally, the GREGoR Consortium Data Coordinating Center maintains a list of errata for Release 3 at https://gregorconsortium.org/data/release-notes/r03-errata

Release Date: July 2025

AnVIL Workspaces:
- [AnVIL_GREGoR_R03_GRU](#)
- [AnVIL_GREGoR_R03_HMB](#)

Summary of changes since previous release:
- R03 adds an additional 1,489 participant IDs as well as short-read ATAC-seq data to the GREGoR Dataset. Additionally, 43 participant IDs in R02 have been retired (see below). In total, R03 includes a total of 8,840 participants.
- The version of the GREGoR Data Model used in R03 includes data tables that support short-read ATAC-seq data.
- A subset of aligned DNA short-read files were uniformly processed by the GREGoR Data Coordinating Center (DCC). The R03 dataset contains these reprocessed, aligned DNA short-read files instead of the aligned DNA short-read files separately processed by GREGoR Research Centers. The `aligned_dna_short_read_id` for each reprocessed file begins with "GREGoR_DCC_A1".
- Jointly called, multisample VCFs (split by chromosome) and single-sample genomic VCFs (gVCFs) are available for DCC reprocessed aligned DNA short-read files. These VCFs contain genotype data for single-nucleotide variants and short indels and can be found in the `called_variants_dna_short_read` table.
- R03 fixes the following errors identified in R02:
  - The subject UW_CRDR-sub-1443008 has been removed from the R03 dataset as it was inappropriately included in the R02 GRU release consent group.
  - R03 corrects metadata for family PMGRC-1067.
  - The subject Broad_RGP_2076_2 was removed from the R03 dataset due to being part of an incomplete family with no proband.
  - A small subset of participants were present in the dataset R02 with multiple, unique participant_ids. These were entered erroneously because the participant

has multiple tissue (analyte) or experiment types. We identified this issue in R02 release for the following families:

- Broad_BON_B09_24
- Broad_BON_B11-09
- Broad_BON_B16-87
- Broad_BON_B17-01
- Broad_BON_B17-54
- Broad_BON_B17-77
- Broad_BON_B17-97
- Broad_BON_B17_15
- Broad_BON_B18-37
- Broad_BON_B18-97
- Broad_BON_B19-52
- Broad_BON_B20_32
- Broad_BON_UC35
- Broad_BON_UC355
- Broad_BON_UC522
- Broad_HK079
- Broad_HK080
- Broad_HK081
- Broad_HK085
- UW_CRDR-fam-1590041
- UW_CRDR-fam-1590067

R03 corrects this issue by retaining one unique participant_id for each individual and mapping both analyte_ids to this participant_id. The file R03_remapped_participant_ids.tsv has a table with the mapping of participant_ids retired in R03 to new participant ids.