



Release Notes

Release 5 (R05):

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 5 at <https://gregorconsortium.org/data/release-notes/r05-errata>

Release Date: May 2026

AnVIL Workspaces:

- [AnVIL_GREGoR_R05_GRU](#)
- [AnVIL_GREGoR_R05_HMB](#)

Summary of changes since previous release:

- R05 adds an additional 1,634 participant IDs to the GREGoR Dataset, and retires 25. In total, R05 includes a total of 12,292 participants.
- A subset of aligned DNA short-read files have been uniformly processed by the GREGoR Data Coordinating Center (DCC). The R05 dataset contains these harmonized, aligned DNA short-read files in addition to the aligned DNA short-read files separately processed by GREGoR Research Centers. The `aligned_dna_short_read_id` for each harmonized file begins with "GREGoR_DCC_A1" and can be found in the `aligned_dna_short_read` table.
- Jointly called multisample VCFs (split by chromosome) and single-sample genomic VCFs (gVCFs) are available for DCC harmonized data. These VCFs contain genotype data for single-nucleotide variants and short indels and can be found in the `called_variants_dna_short_read` table.
- R05 fixes the following errors identified in R04:
 - The VEP-annotated VCFs from the GREGoR joint callset contain incorrect VRS identifiers. These remain correct in the corresponding unannotated VCFs.
 - R05 omits data associated with the following errors in R04:
 - The files associated with `aligned_pac_bio_ids` GSS228799-01-011-LG-2 and GSS128794-01-010-LG-2 are unintended duplicates and should not be used.
 - The following files are duplicates of other bams and may be deleted in the future:

- gs://fc-secure-a1f0d27d-c9d5-48bc-b4ba-5e0f81784fb1/GSS179904/SR_GS/GSS179904-3147-4e0bd01a-6bf2-4731-a054-0f8685bd8d62_reheadered.bam
 - gs://fc-secure-a1f0d27d-c9d5-48bc-b4ba-5e0f81784fb1/GSS179907/SR_GS/GSS179907-3148-05e96cd4-84d9-4274-a05f-cc9240b2f7ef_reheadered.bam
 - gs://fc-secure-a1f0d27d-c9d5-48bc-b4ba-5e0f81784fb1/GSS279909/SR_GS/GSS279909-3146-fc594df4-589c-446a-a6e3-f3af55ad237b_reheadered.bam
 - gs://fc-secure-ab18917b-873b-4082-8181-7653393999c9/GSS293189/LR_GS/pacbio/GSS293189_r64282e_20230419_021750.GC_A_000001405_15.haplotagged.bam
- PMGRC-1076-1065-3 and PMGRC-1079-1065-3 have been identified as swapped in the short read WGS data.
 - PMGRC-423-423-0 and PMGRC-425-423-1 have been identified as swapped in the PacBio data.
 - The following participants' PacBio data have been flagged for excess human source contamination: PMGRC-2109-2109-0, PMGRC-380-381-1, PMGRC-82-82-0
 - The PacBio data for PMGRC-418-418-0 has metadata discrepancies.
 - PMGRC-478-478-0 and PMGRC-479-478-2 have been identified as swapped in the short read RNA-seq data.
 - The RNA-seq data for PMGRC-838-839-2 have been flagged as potentially contaminated.
 - The short read RNA-seq data for PMGRC-104-104-0, PMGRC-1116-1116-0, and PMGRC-1229-1229-0 have metadata discrepancies.
 - The following short read RNA-seq data are flagged for QC failures (low RIN, insufficient percent human reads, insert size distribution, GC content, percent duplication): PMGRC-45-43-1, PMGRC-1203-1200-1, PMGRC-1076-1065-3, PMGRC-88-88-0, PMGRC-125-125-0, PMGRC-924-924-0, PMGRC-1065-1065-0, PMGRC-265-265-0.
 - The PacBio data for GSS210939 and GSS116494 are duplicates of other files and should not be used.
 - The PacBio data for GSS140604 and GSS240609 are swapped.
 - The PacBio data for GSS167887 and GSS167884 are swapped.
 - The PacBio data for GSS132377 and GSS132374 are swapped.
 - The PacBio data for GSS240249 does not match other data for this participant and should not be used.
 - The PacBio data for GSS265559 does not match other data for this participant and should not be used.
 - The ATAC data for GSS291239 is a duplicate of another file.
 - The ATAC data for GSS128794 and GSS228799 are swapped.
 - The short read RNA data for GSS130487 and GSS230489 are swapped.

- The short read RNA data for GSS152454 and GSS152457 are swapped.
- The short read RNA data for GSS176617 and GSS176614 are swapped.