GREGoR R04 Data Summary

GREGoR DCC

2025-09-24

Contents

List of Figures

Overview

This report provides data summaries for the fourth release of the GREGoR Dataset (R04) which is available on AnVIL. Graphical and tabular summaries of participant, family, experiment, and phenotype information are generated from information provided by member Research Centers (RCs) and uploaded to AnVIL data tables using the GREGoR data model (https://github.com/UW-GAC/gregor_data_models).

Abbreviations:
RCs:
BCM = Baylor College of Medicine Research Center
BROAD = Broad Institute
UCI = University of California, Irvine
GSS = GREGoR Stanford Site
UW-CRDR = University of Washington Center for Rare Disease Research
Consent codes:
GRU = General research use and clinical care
HMB = Health/medical/biomedical research and clinical care

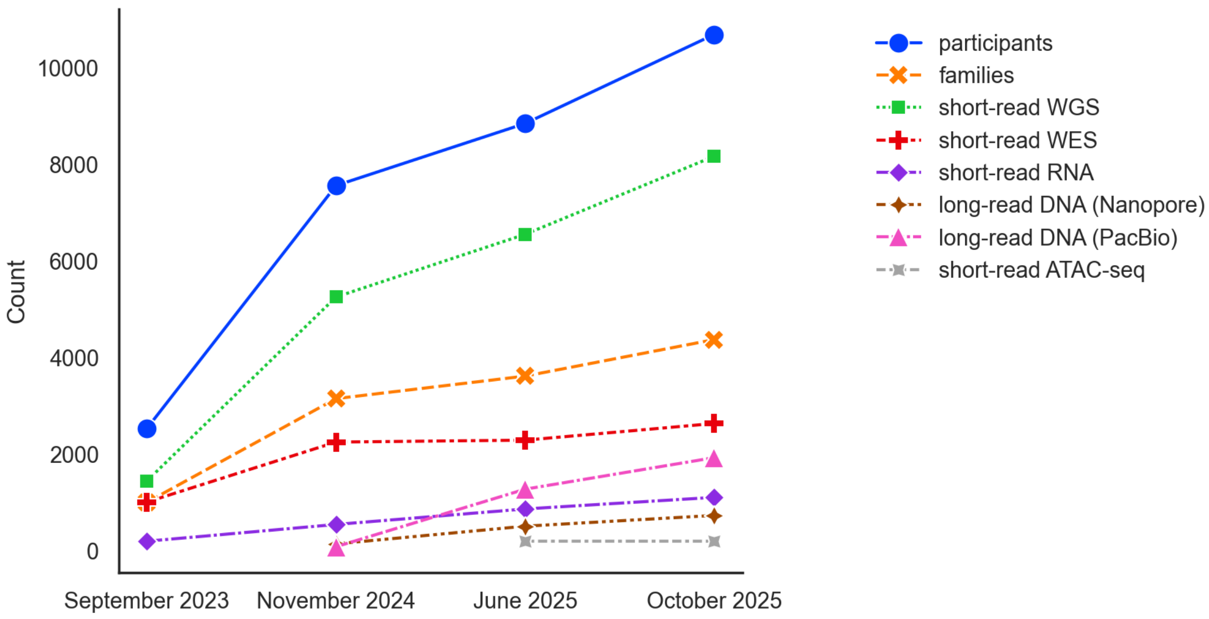

Figure 1: Overview of the GREGoR Dataset across data releases.

Table 1: The number of participants, families and experiments in the GREGoR Dataset

	Number of entries
PARTICIPANTS	10683
FAMILIES	4366
SHORT-READ WGS	8161
SHORT-READ WES	2629
SHORT-READ RNA	1100
LONG-READ DNA (NANOPORE)	726
LONG-READ DNA (PACBIO)	1922
SHORT-READ ATAC-SEQ	189

Summary of solve status for probands in the GREGoR Dataset

Table 2: Summary of solve status for probands in the GREGoR Dataset

	No. of probands	%
Partially solved	24	0.01
Probably solved	111	0.03
Solved	537	0.13
Unsolved	3581	0.84

Solve status definitions:
Solved:

Dominant: Pathogenic/Likely Patthogenic (P/LP) variant with matching inheritance pattern in a gene that also matches the phenotype where there is at least evidence of moderate gene-disease validity (with at least 1 prior publication or preprint or submission to GenCC by any submitter, including GREGoR center)

Recessive: Biallelic P/LP variants with good phenotype and inheritance mode match in gene with at least moderate evidence of gene-disease validity. Can include cases where phase is unknown if the phenotype match is strong (otherwise downgrade to probably solved)

Dual diagnosis/blended phenotype: Can include cases where some components of the phenotype are not explained, particularly phenotypes that may be non-Mendelian or familial

Partially (phenotype) solved: P/LP variant(s) with matching inheritance pattern in a gene with at least moderate gene-disease validity that only accounts for part of the phenotype (i.e. a P variant to explain hearing loss in a patient with hearing loss and intellectual disability); If multiple partial solves are discovered that together explain the majority of the phenotype, the case would be considered solved

Probably solved: i.e. high chance that causal variants have been identified but need more support to reach P/LP

Unsolved: includes cases with a low or moderate candidate listed in the genetics findings table; these are cases where full analysis effort should still be put forth

Unaffected: Any unaffected participant

GREGoR participant and family summaries

Table 3: The number of participants and families in the GREGoR Dataset by consent group

Consent	Participants	Families
GRU	7938	3316
HMB	2745	1050
Total	10683	4366

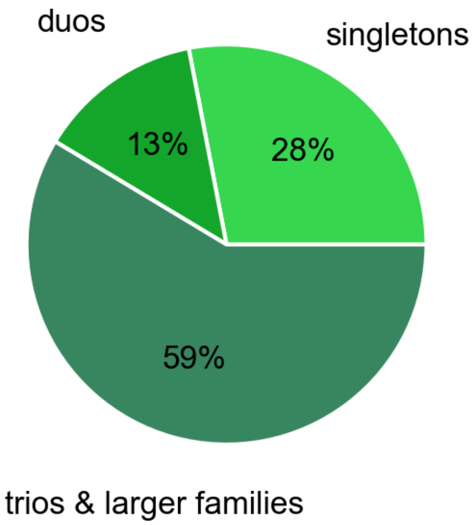

Figure 2: Pie chart summary of family structure in the GREGoR Dataset

Table 4: Table summary of family structure in the GREGoR Dataset

Family Structure	No. of Families
Singletons	1224
Duos	583
Trios & larger families	2559
Total	4366

Phenotype Summaries

Table 5: Summary of affected status in the GREGoR Dataset.

	No. of participants	%
Affected	4933	46.2
Possibly affected	20	0.2
Unaffected	5370	50.3
Unknown	360	3.4

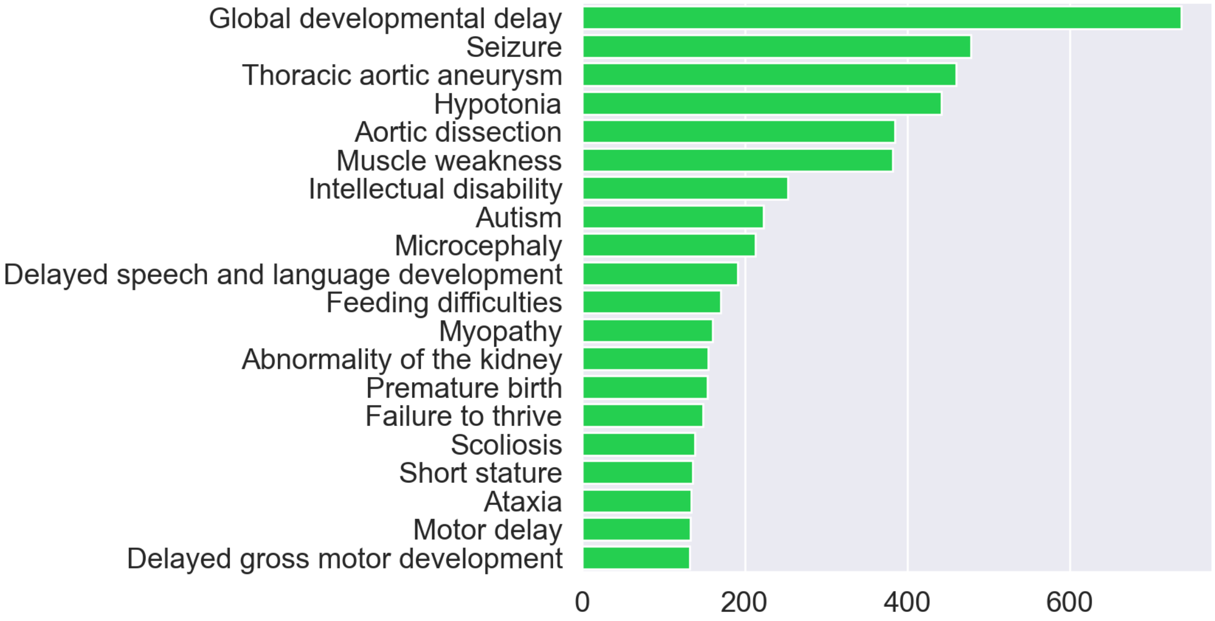

Figure 3: Common phenotypes (HPO) in the GREGoR Dataset. Phenotypes (HPO names) are on the y-axis, in descending order and shown if family count > 120 (x-axis).

Experiment Summaries

Short-read DNA

Table 6: The number of unique participants with short-read DNA sequencing experiments in the GREGoR Dataset.

Consent	Exome	Genome	Targeted
GRU	1907	6141	0
HMB	722	2020	1
Total	2629	8161	1

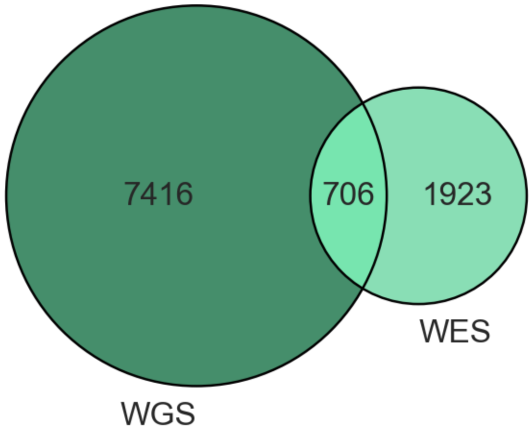

Figure 4: Venn diagram showing participants with whole genome (WGS) and whole exome (WES) sequencing data in the GREGoR Dataset.

Short-read RNA

Table 7: The number of unique participants with short-read RNA in the GREGoR Dataset

Consent	paired-end	paired-end & untargeted	untargeted
GRU	431	518	0
HMB	52	98	1
Total	483	616	1

Table 8: Short-read RNA sequencing experiments by primary biosample

Primary_biosample	No. of experiments
UBERON:0000178 (blood)	893
CL:0000057 (fibroblast)	115
UBERON:0002385 (muscle tissue)	73
UBERON:0000479 (tissue)	7
UBERON:0019306 (nose epithelium)	7
CL:0000542 (lymphocyte)	3
CL:0000034 (stem cell)	1

Table 8: Short-read RNA sequencing experiments by primary biosample

Primary_biosample	No. of experiments
UBERON:0001003 (skin epidermis)	1

Short-read ATAC-seq

Table 9: The number of unique participants with short-read ATAC-seq experiments.

Consent	No. of Participants
GRU	189
Total	189

Table 10: The number short-read ATAC-seq experiments by primary biosample.

Primary_biosample	No. of experiments
CL: 0000576	189

Long-read DNA

Table 11: The number of unique participants with long-read whole genome experiments in the GREGoR Dataset.

Consent	Nanopore	PacBio_FiberSeq	PacBio
GRU	707	5	1913
HMB	19	4	0
Total	726	9	1913

Participants and probands with multiple data types

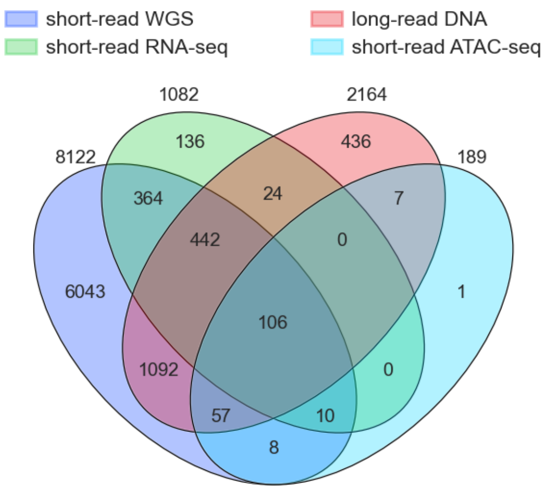

Figure 5: Venn diagram showing **participants** with multi-omic data in the GREGoR Dataset.

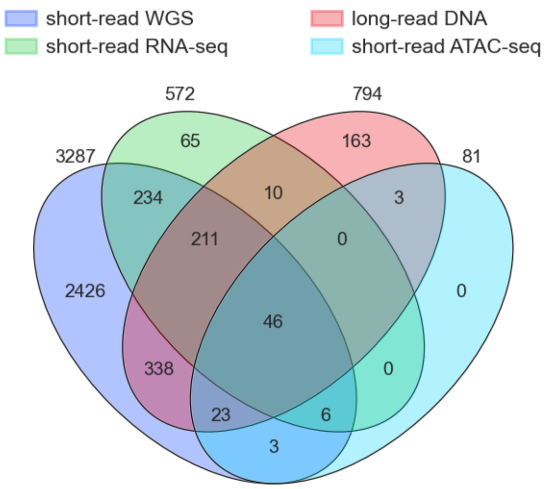

Figure 6: Venn diagram showing **probands** with multi-omic data in the GREGoR Dataset.

Summary of genetic findings in the GREGoR Dataset

Table 12: The number of participants with genetic findings by variant classification.

Variant_classification	No. of entries
Benign	1
Curation in progress	568
Likely benign	2
Likely pathogenic	248
Pathogenic	236
Uncertain significance	407
Uncertain significance - high	37
Uncertain significance - moderate	8
Well-established P/LP	113
nan	66
Total	1686

Table 13: Variant type(s) listed in the GREGoR genetic findings table.

Variant_type	No. of entries
SNV	966
SNV/INDEL	390
INDEL	226
SV	82
RE	14
CNV	8
Total	1686

RE = repeat element; SNV/INDEL = single nucleotide variant OR insertion/deletions; SV = structural variant; CNV=copy number variant

Table 14: Method of discovery for genetic finding entries.

	No. of entries
SR-GS	1170
SR-ES	473
SR-ES & SR-GS	21
SR-GS-reanalysis	6
SR-GS & LR-GS	4
SR-GS & LR-GS	3
LR-GS	2
LR-GS & SR-GS	2
SR-ES-reanalysis	2
SR-ES & SR-GS & LR-GS	1
SR-GS & LR-GS & SNP array	1
nan	1
Total	1686

SR-GS = short-read genome; SR-ES = short-read exome; LR-GS = long-read genome