

# Getting to know GREGoR Data in AnVIL

Dr. Ben Heavner  
GREGoR Data Coordinating Center

November 2, 2023



*Genomics Research to Elucidate the Genetics of Rare Diseases*

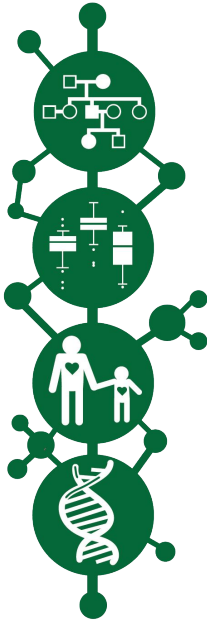
---

 @GREGoR\_research

 [www.gregorconsortium.org](http://www.gregorconsortium.org)

# The first release of the GREGoR Consortium Data Set is registered with dbGaP: study phs003047

Participants	2512 (1130 affected)
Families	990
Short read whole exome sequence	997
Short read whole genome sequence	2438
RNA-seq	183
Genome Build	GRCh38
Size (TB)	72.9



# Accessing the GREGoR Consortium Dataset

- The GREGoR Consortium is among the first NHGRI efforts to release Consortium Data via the NHGRI Analysis Visualization and Informatics Lab-space (AnVIL)

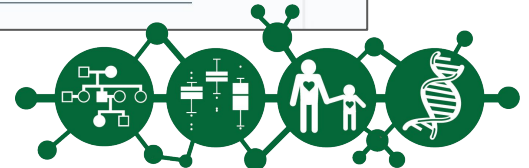


## Welcome to AnVIL

The NHGRI AnVIL (Genomic Data Science Analysis, Visualization, and Informatics Lab-space) is a project powered by Terra for biomedical researchers to access data, run analysis tools, and collaborate. [Learn more about Terra.](#)

If you are a new user or returning user, click sign in to continue.

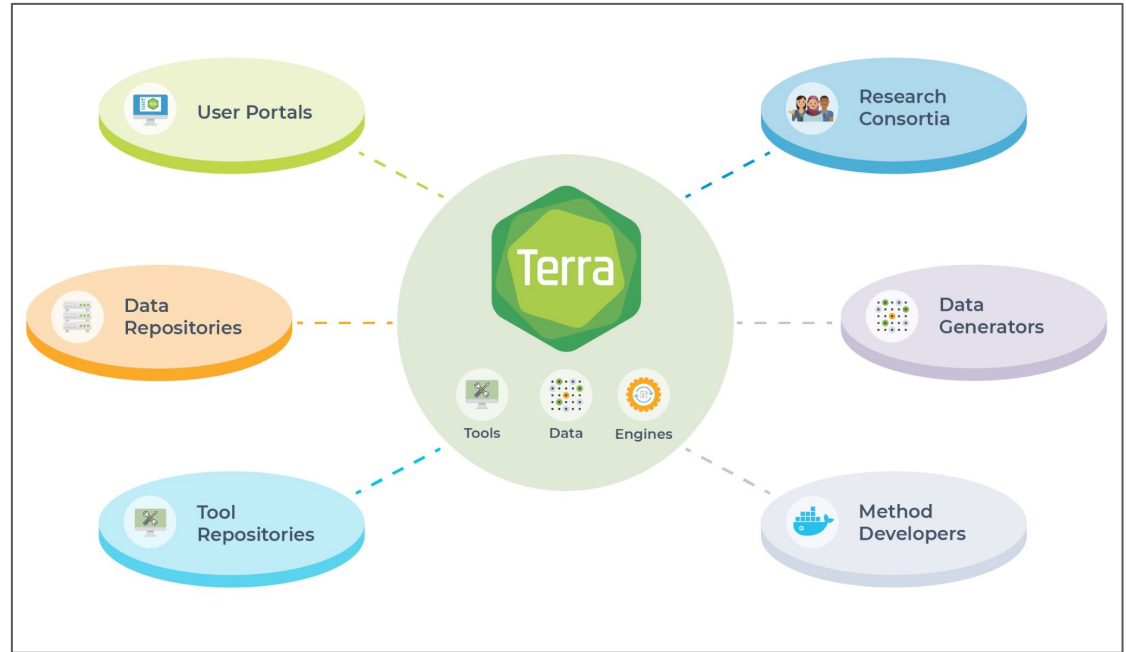
SIGN IN



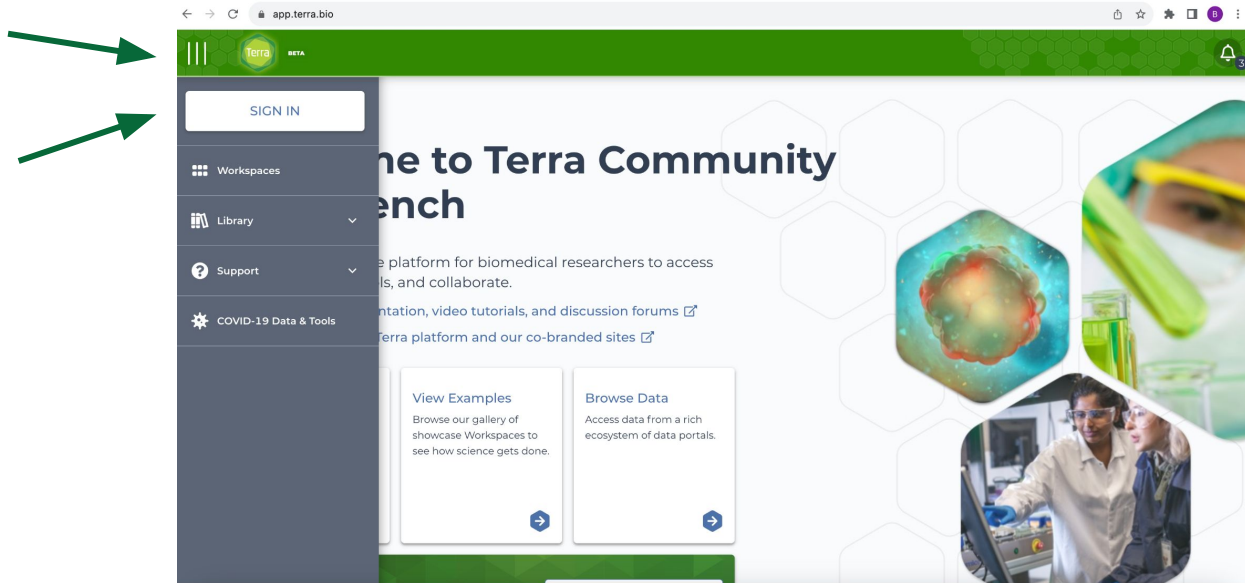
# What is AnVIL?

The [Analysis, Visualization and Informatics Labspace \(AnVIL\)](#) is a data commons platform funded by the National Human Genome Research Institute.

It leverages [Terra](#), a cloud-native, open platform that connects researchers to each other and to the datasets and tools they need to achieve scientific breakthroughs.



# Logging in to AnVIL



<https://anvil.terra.bio>  
(blue color scheme)

OR

<https://app.terra.bio>  
(green color scheme,  
access to beta features)



# GREGoR Consortium Data Workspace Dashboard

← → ↻ app.terra.bio/#workspaces/gregor-dcc/GREGOR\_COMBINED\_CONSORTIUM\_U03

Terra BETA WORKSPACES Workspaces > gregor-dcc/GREGOR\_COMBINED\_CONSORTIUM\_U03 > Dashboard

COVID-19 Data & Tools

DASHBOARD DATA ANALYSES WORKFLOWS JOB HISTORY

## ABOUT THE WORKSPACE

Note: Data and links in this workspace have been migrated from their original upload workspaces to workspaces with a new architecture. Please notify the DCC if you have trouble accessing data.

The [GREGoR consortium](#) is using the NHGRI Analysis Visualization and Informatics Lab-space ([AnVIL](#)) for data sharing and collaborative analysis in the cloud.

This workspace includes data contributed by each of the [GREGoR Research Centers](#) in the first and second upload cycles as a combined data set which is designed to facilitate cross-RC analysis on the AnVIL.

GREGoR investigators are also encouraged to pose questions and engage in conversation in the #anvil channel on the [GREGoR Consortium Slack workspace](#). (Please [contact us](#) if you need an invitation to Slack.)

Here are some resources for more information: **from AnVIL:**

- Getting Started [Getting started on AnVIL](#) book

**Depositing data for consortium sharing and controlled-access release:**

- [GREGoR Consortium Data Submission SOP](#)
- [Data Upload Status update \(8/9/22\)](#)
- [Importing data to AnVIL](#)
- [Tutorial Video](#)
- [Example upload workspace](#)

### WORKSPACE INFORMATION

Last Updated	8/17/2023
Creation Date	7/7/2023
Access Level	Project Owner

### CLOUD INFORMATION


### OWNERS

### AUTHORIZATION DOMAIN

### TAGS

### NOTIFICATIONS

Rate: \$0.00 per hour



# Working with GREGoR Data on AnVIL

- Interactive analysis
  - Jupyter notebooks (Python, R)
  - RStudio (with Bioconductor installed)
- Workflows
  - WDL workflows for now, others on AnVIL development roadmap
- Other tools
  - [Seqr](#)
  - [IGV](#)
  - export/download (note: requester pays)
  - Command line/gsutils or GCP virtual machines



Google Cloud Platform



# Combined Data Workspace - Data Tab



BETA

WORKSPACES

Workspaces > gregor-dcc/GREGOR\_COMBINED\_CONSORTIUM\_U03 > Data



DASHBOARD

DATA

ANALYSES

WORKFLOWS

JOB HISTORY

+ IMPORT DATA



OPEN WITH...



0 rows selected

ADVANCED SEARCH

Search



Rate:  
< \$0.01  
per hour



TABLES



Search all tables



- aligned\_dna\_sh... (2438) ⓘ
- aligned\_dna\_sho... (264) ⓘ
- aligned\_rna\_shor... (231) ⓘ
- analyte (2639) ⓘ
- called\_variants\_d... (264) ⓘ
- experiment (2635) ⓘ
- experiment\_dn... (2443) ⓘ
- experiment\_rna\_... (192) ⓘ
- family (999) ⓘ

Data Tables



<input type="checkbox"/>	phenoty... ↓ ⓘ	additional_details ⓘ	additional_modifiers ⓘ	onset_age_range ⓘ	ontology
<input type="checkbox"/>	0004ef8afd0a...	NA	NA	NA	HPO
<input type="checkbox"/>			NA	NA	HPO
<input type="checkbox"/>	002a7d76d62...	NA	NA	NA	HPO
<input type="checkbox"/>	003067a8319...	NA	NA	NA	HPO
<input type="checkbox"/>	0038f532813...	NA	NA	NA	HPO
<input type="checkbox"/>	004f6073f1d0...	bilateral sensorineural hearing loss; ...	NA	HP:0011463	HPO
<input type="checkbox"/>	0055abc3058...	NA	NA	NA	HPO
<input type="checkbox"/>	006815d767e...	NA	NA	NA	HPO
<input type="checkbox"/>	007264de8ed...	NA	NA	NA	HPO
<input type="checkbox"/>	009dcd7eb86...	dyslexia	NA	NA	HPO

1 - 100 of 5806



1

2

3

4

5

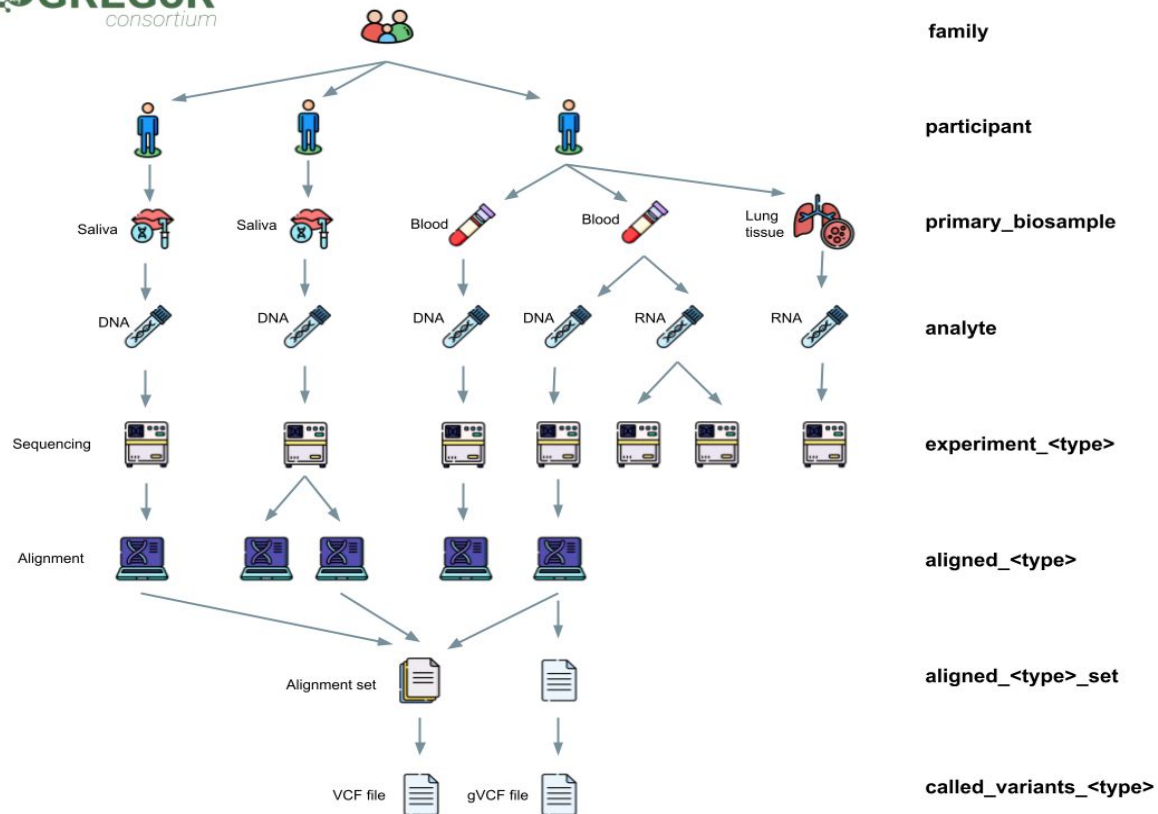


Items per page: 100





# Workspace Tables = GREGoR Data Model

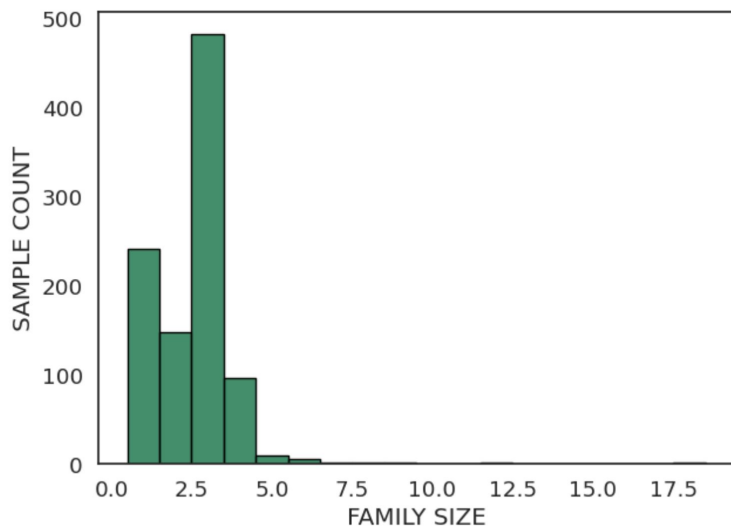


# To Summarize: How to gain access and begin analyzing GREGoR Data

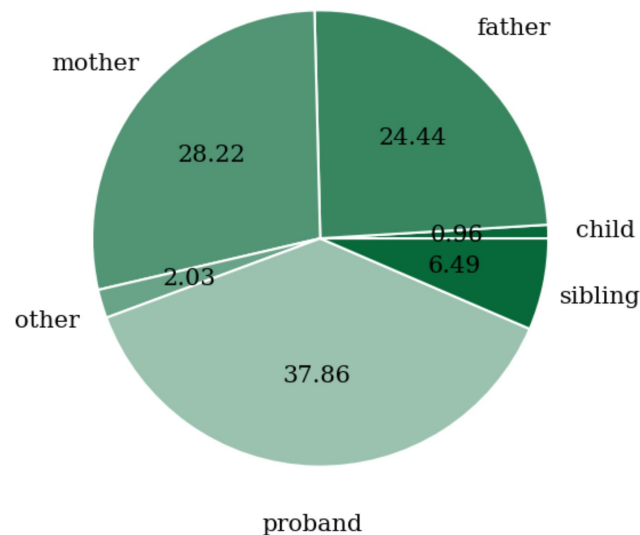
1. Submit Data Access Request to NHGRI
2. Log in to AnVIL with AnVIL account linked to eRA ID
3. Clone the most recent released GREGoR Dataset Workspace (e.g. AnVIL\_GREGOR\_RELEASE\_01\_GRU or AnVIL\_GREGOR\_RELEASE\_01\_HMB)
4. Conduct your analysis using your cloned workspace



# More about the GREGoR Dataset: family structure



Family Size Distribution



Proband relationship (%)



# More about the GREGoR Dataset: phenotypes

- **1,274** Participants with HPO-encoded phenotypes
- **1,457** Unique phenotype terms (HPO)

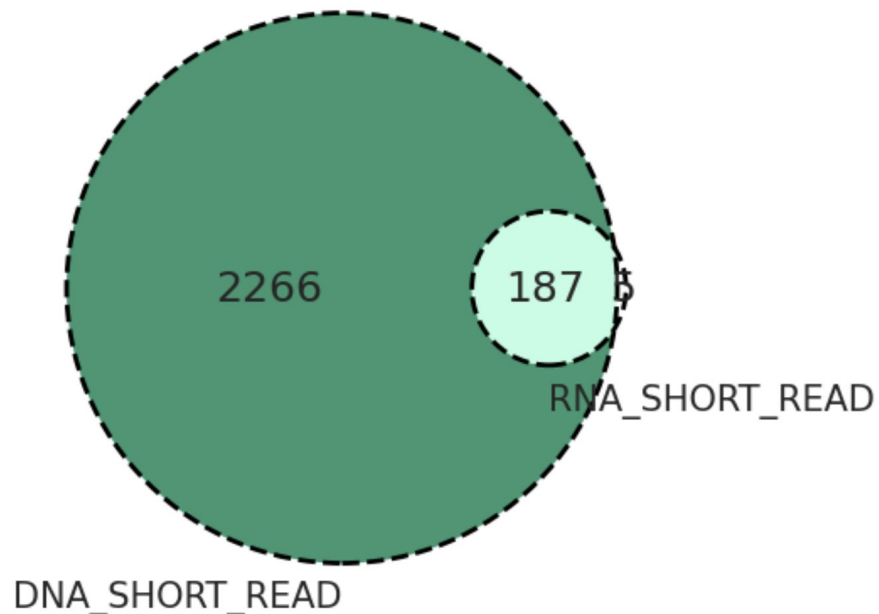
## Most common phenotype terms

HP: 0001263	Global developmental delay	N = 205
HP: 0001250	Seizure	130
HP: 0001252	Hypotonia	119
HP: 0001249	Intellectual Disability	110
HP: 0001324	Muscle weakness	93
HP: 0002011	Morphological central nervous system abnormality	91



# More about the GREGoR Dataset: molecular data

Experiment Type	Number of Participants
DNA short read only (WGS and/or WES)	2266
RNA-Seq only	5
DNA and RNA	187



Upload Cycle	R01	Future
Data Model	1.1	1.?
Included Data Types	Family structure,  Phenotype,  Short Read DNA,  Short Read RNA	Short Read DNA,  Short Read RNA,  Expanded Genetic Findings,  Long Read DNA (Nanopore),  Long Read DNA (Pac Bio),  ATAC-Seq,  Joint Callset,  Optical Mapping,  Functional Data

## The future: Expanding the GREGoR Dataset

- **Additional participants**
- **Additional experimental data types**



# Housekeeping notes



Session Feedback and Interest Survey  
<https://tinyurl.com/gregor-post-session-survey>

- Please **register your attendance** at this session on the sign-in sheets on the back table
- Learn more about GREGoR at <https://gregorconsortium.org/>
- Visit GREGoR Posters at ASHG!
- Contact the GREGoR Data Coordinating Center via email at [gregorconsortium@uw.edu](mailto:gregorconsortium@uw.edu)



# Acknowledgement

- Particular thanks to the members of the GREGoR Consortium for their contributions to this work
- The GREGoR Consortium is funded by the National Human Genome Research Institute of the National Institutes of Health, through the following grants: U01HG011758, U01HG011755, U01HG011745, U01HG011762, U01HG011744, and U24HG011746. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.







**Questions  
or  
Comments**