

Loading data from an AnVIL workspace into *seqr*

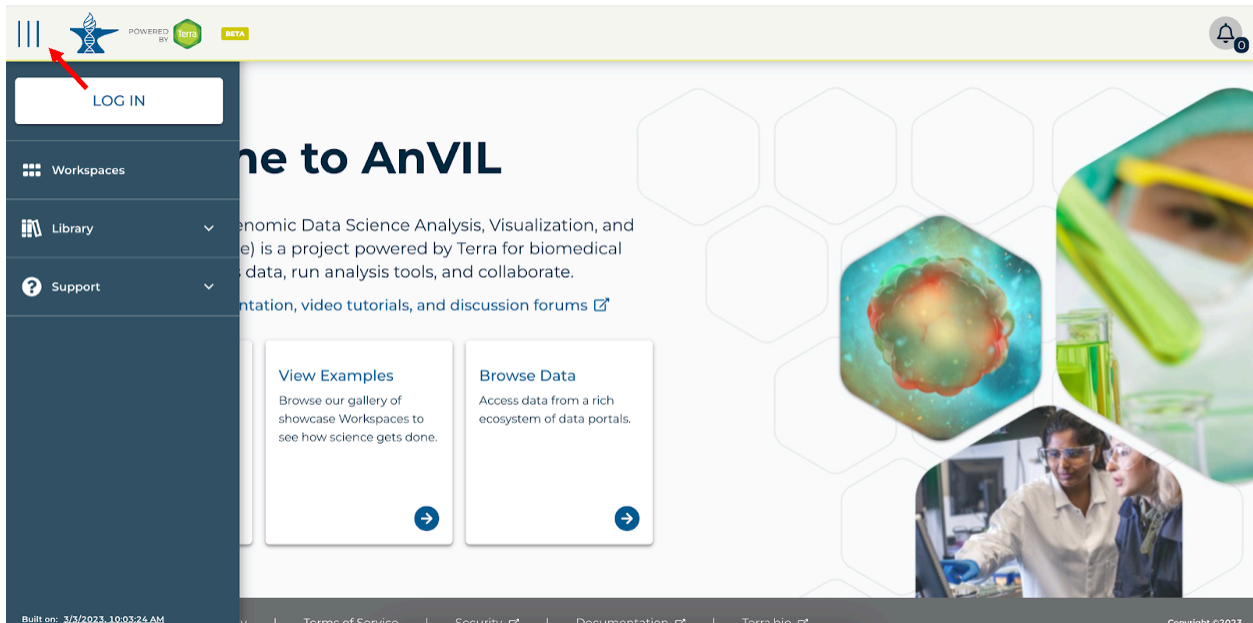
[Watch [video tutorial](#)]

1. Register for a Terra account

The NHGRI's AnVIL project is powered by Terra to access data, run analysis, and collaborate.

To use AnVIL, users must register for a Terra account, using a Gmail or other email (an institutional email, for example) associated with a Google identity.

Navigate to the [Terra](#) home page and click on the hamburger menu at the top left to sign in and register a new account.



Read Terra's information page on [setting up billing in Terra](#) and [understanding costs](#).

See [registering for a Terra account](#) for additional information on registering an account.

2. Prepare your files

Analysis in *seqr* is optimized for loading of joint-called VCFs generated using GATK or DRAGEN pipelines and joint-called using WARP (WDL Analysis Research Pipelines) or GVS (Genomic Variant Store). Sharded VCFs are also accepted. If you need to generate a joint-called file, you can use [GATK tooling](#). For more information about generating and validating a joint called file, read this [documentation](#).

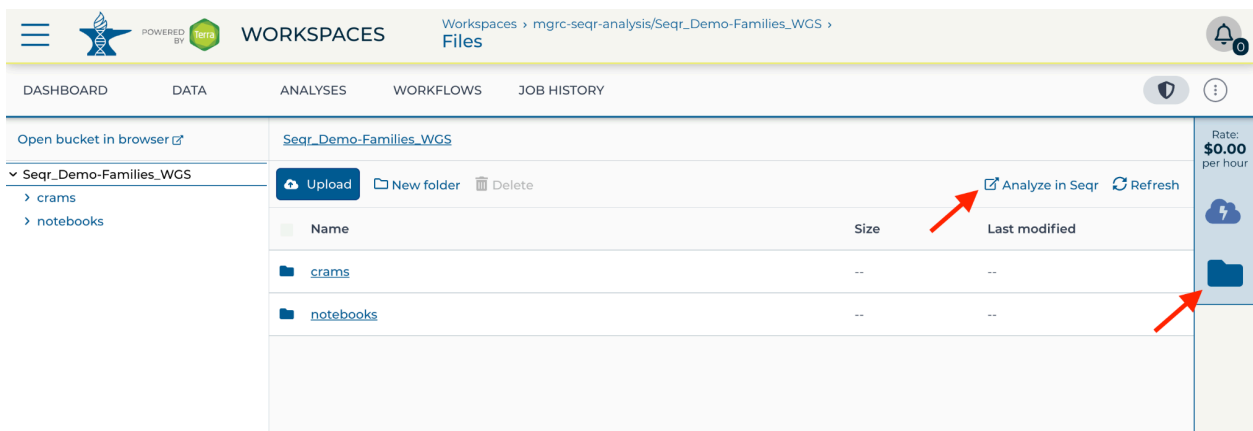
We also accept joint-called VCFs with a .gz extension provided they are internally bgzipped. Here is more information on the bgzip tool: <http://www.htslib.org/doc/bgzip.html>.

The joint-called VCF file must be stored in a workspace in which you have [Writer or Owner](#) level access and have the [Can Share](#) permissions. Additionally, the workspace must not be associated with any [Authorization Domains](#) in order for *seqr* to access it. If the workspace does not meet these requirements, we recommend you create a new workspace with the needed permissions and load your files from there.

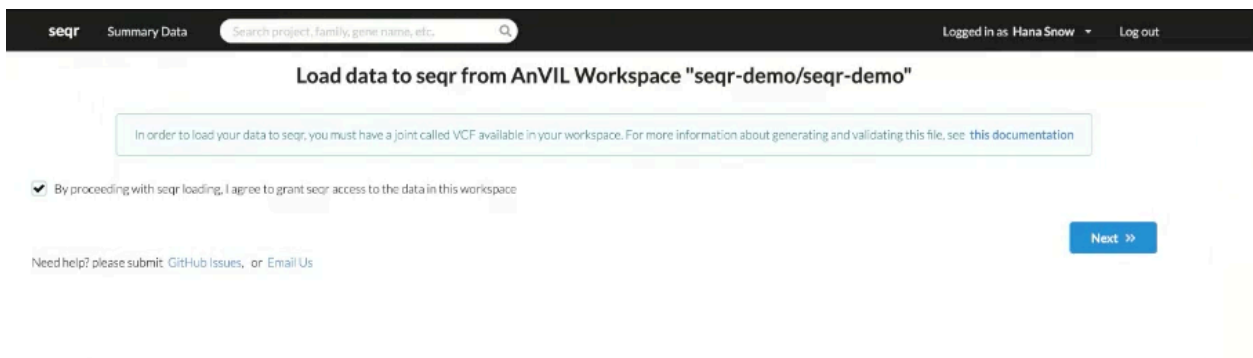
3. Upload files to *seqr*

Once you have a joint-called VCF on your local computer, you can upload the file to an AnVIL workspace. If you are using sharded VCFs, make sure all files are in one folder in the workspace.

To upload files, log in to your AnVIL account, select Browse Workspace files, upload a joint-called VCF, and then select Analyze in *seqr*.



This will prompt you to grant *seqr* access to your files in the workspace.



Select the joint-called VCF file you wish to load using the dropdown menu. Specify the Sample Type and Genome Version.

seqr Summary Data Search project, family, gene name, etc. Logged in as Hana Snow Log out

Load data to seqr from AnVIL Workspace "seqr-demo/seqr-demo"

In order to load your data to seqr, you must have a joint called VCF available in your workspace. For more information about generating and validating this file, see [this documentation](#)

Path to the Joint Called VCF

Sample Type Exome Genome

Genome Version GRCh37 GRCh38

Need help? please submit [GitHub Issues](#), or [Email Us](#)

<< Back Next >>

You can enter an optional Project Description which is especially useful if you have multiple projects loaded. You must agree to comply with federal regulations, which do not allow any protected health information (PHI) as seqr is not HIPAA-compliant and must not contain any identifiable information such as names or dates of birth in the pedigree or notes.

seqr Summary Data Search project, family, gene name, etc. Logged in as Hana Snow Log out

Load data to seqr from AnVIL Workspace "seqr-demo/seqr-demo"

In order to load your data to seqr, you must have a joint called VCF available in your workspace. For more information about generating and validating this file, see [this documentation](#)

Project Description

Upload Pedigree Data

seqr is not a HIPAA-compliant platform. By proceeding, I affirm that this pedigree file does not contain any protected health information (PHI), including in any of the IDs or in the notes. PHI includes names, contact information, birth dates, and any other identifying information

I Agree

Need help? please submit [GitHub Issues](#), or [Email Us](#)

<< Back Submit

Use the blank template or the example file provided to enter the Pedigree Data, and then hit Submit. This may take a few minutes. Do not hit refresh when the page is loading.

Project Description

Upload Pedigree Data

To load individual data from an AnVIL workspace to a new seqr project, upload a table in one of these formats:

Excel (.xlsx) [download blank template](#) or [an example pedigree](#)

Text (.tsv / .csv) [download blank template](#) or [an example pedigree](#)

The table must have a header row with the following column names.

Required Columns:

Family ID Family ID

Individual ID Individual ID (needs to match the VCF ids)

Sex Male, Female, or Unknown

Affected Status Affected, Unaffected, or Unknown

HPO Terms Semi-colon separated list of HPO terms. Required for affected individuals only.

Optional Columns:

Paternal ID Individual ID of the father

Maternal ID Individual ID of the mother

Notes free-text notes related to this individual

Click here to upload a table, or drag drop it into this box

Need help? please submit [GitHub Issues](#), or [Email Us](#)

<< Back Submit

Submitting the file sends a request to the *seqr* team to load your data. This can take up to a week to process. You will receive an email when your data is fully loaded to *seqr*.

When the data is available in your *seqr* project, you will see a summary of it in the Datasets section. At the bottom of the page, you will see the Families and Individuals based on the information in the Pedigree file submitted.

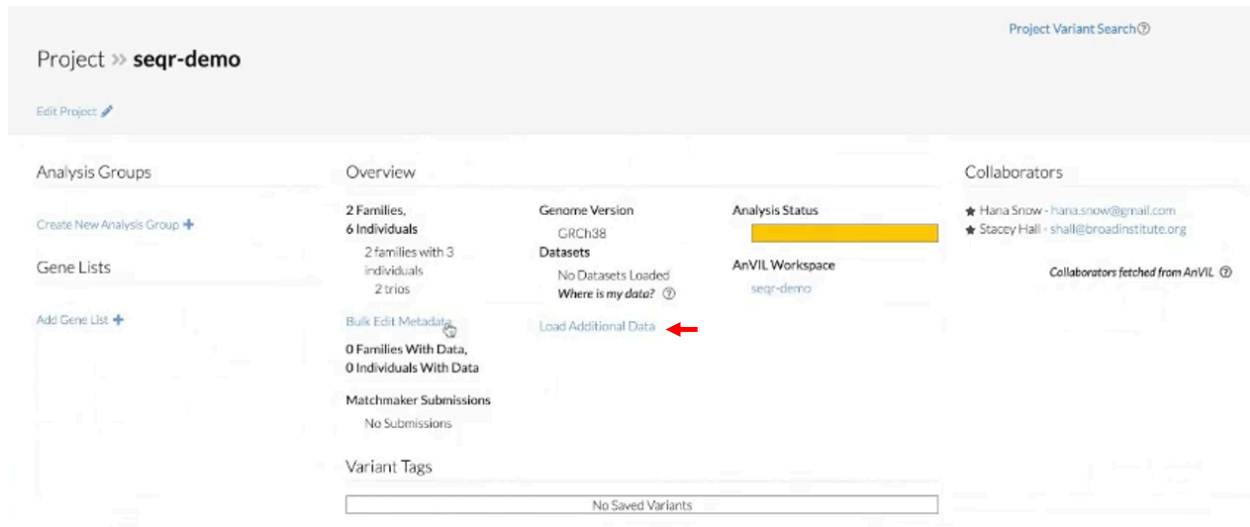
The screenshot shows a pedigree chart for family FAM36. The chart includes three individuals: VCGS_FAMILY_115 (male), VCGS_FAMILY_114 (female), and VCGS_FAMILY_114 (female). To the right of the chart is a detailed metadata panel for the individual VCGS_FAMILY_114. The metadata includes sections for Analysis Groups, Family Description, Assigned Analyst, Case Notes, Analysis Notes, Matchmaker Notes, Coded Phenotype, Post-discovery OMIM #, Age, Age of Onset, Individual Notes, Consanguinity, Other Affected Relatives, Expected Mode of Inheritance, Assisted Reproduction, Maternal Ancestry, Paternal Ancestry, Features, Pre-discovery OMIM disorders, and Previously Tested Genes. The 'Analysis Status' is 'Waiting for data'. The 'Age' is 'Unknown'. The 'Expected Mode of Inheritance' is 'Unknown'. The 'Other Affected Relatives' is 'Unknown'. The 'Maternal Ancestry' and 'Paternal Ancestry' are both 'Unknown'. The 'Features' section is empty. The 'Pre-discovery OMIM disorders' and 'Previously Tested Genes' sections are also empty. The individual was added on 2/8/2023.

You can enter additional case information by directly adding it to the individual or by using the Bulk Edit Metadata feature.

The screenshot shows the project overview page for 'seqr-demo'. The page is divided into several sections: Analysis Groups, Gene Lists, Overview, Collaborators, and Datasets. The Overview section provides a summary of the project's data status. It shows 2 Families and 6 Individuals. The data is broken down as follows: 2 families with 3 individuals and 2 trios. There are 0 Families With Data and 0 Individuals With Data. There are no Matchmaker Submissions. The Overview section also includes a 'Bulk Edit Metadata' link, which is highlighted with a red arrow. The Datasets section shows 'No Datasets Loaded' and a 'Where is my data?' link. The Collaborators section lists Hana Snow and Stacey Hall. The Analysis Status is 'Waiting for data'. The AnVIL Workspace is 'seqr-demo'. The Datasets section also includes a 'Load Additional Data' link. The bottom of the page shows a 'No Saved Variants' message.

4. Load additional data to a project

If at a later date you would like to load additional data to a project already in *seqr*, you can do so by using the Load Additional Data feature. The process is similar to the workflow used to create the original *seqr* project using an updated VCF and Pedigree file.



To add new data, create a new pedigree and a joint called VCF with all the samples you want to include in your update. This should include any new samples you want to add to the project and any of their family members which have been previously loaded. Load this VCF using the Load Additional Data feature on the Project Page. All notes and tags saved in previously analyzed cases will be maintained.

Note that a single Terra workspace corresponds to a specific project in *seqr*. You cannot load data from a new workspace into an existing project. If you would like to have a new project in *seqr*, you can submit a request to load a joint-called VCF from a new workspace.

Please reach out to the [seqr team](#) if you have any questions.

All the best with your analysis!