

# Loading data from an AnVIL workspace into *seqr*

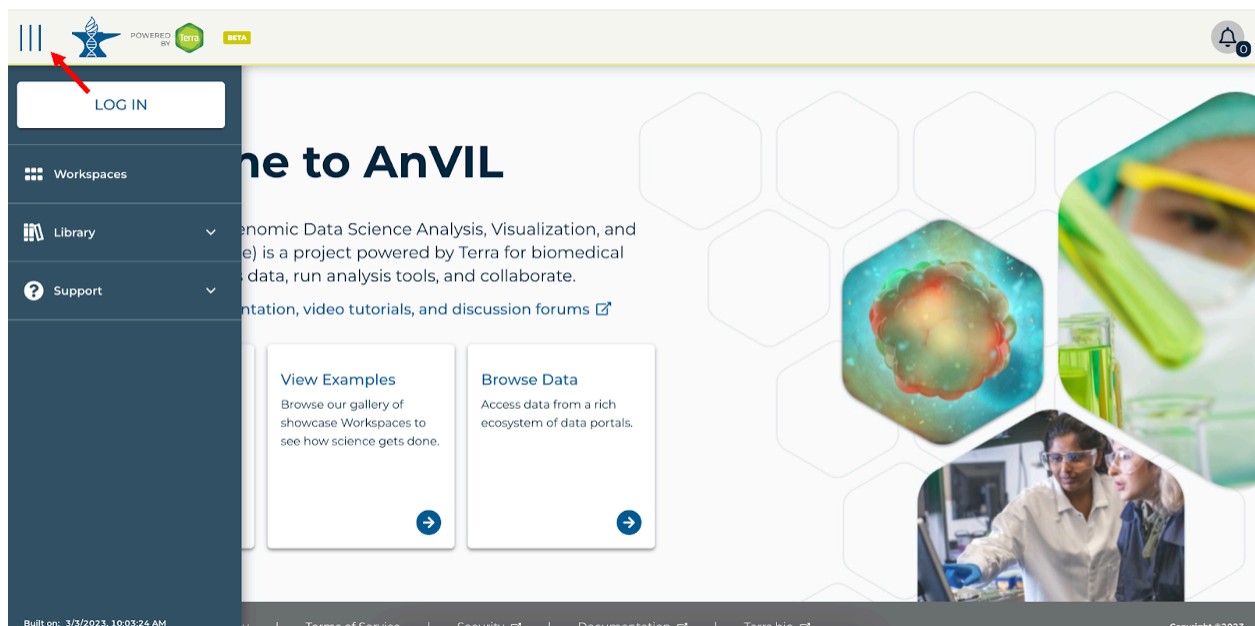
[Watch [video tutorial](#)]

## 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

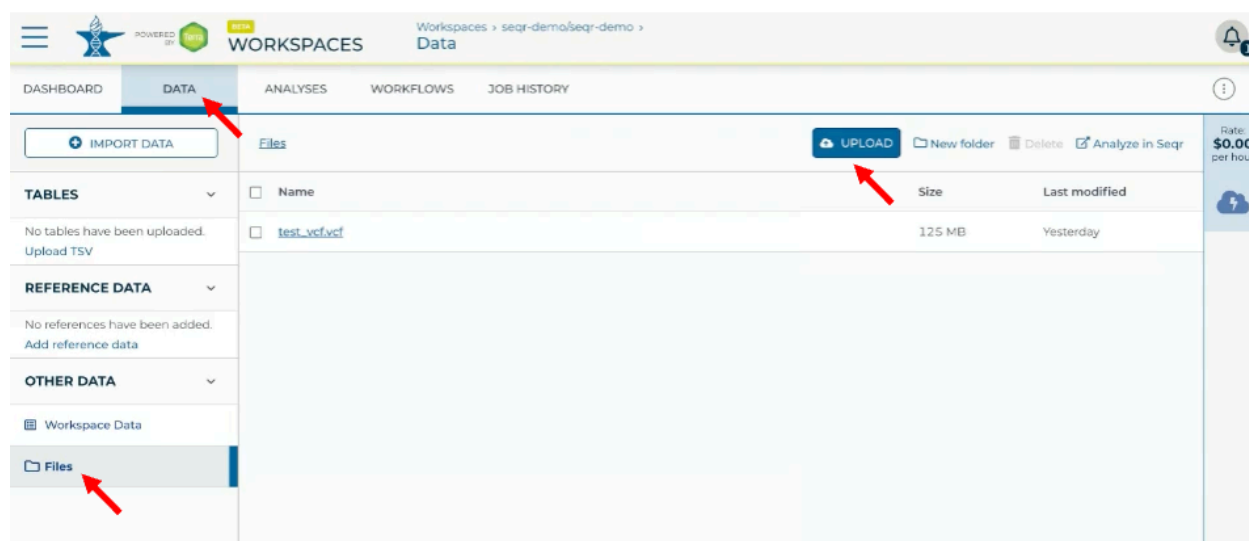
Analysis in *seqr* is optimized for loading of joint-called VCF files generated by the GATK pipeline. If you need to generate a joint-called file, you can use [GATK tooling](#). For more information about generating and validating a joint called file, read this [documentation](#).

The joint-called VCF file have a .gz extension, provided they are internally bgzipped. Here is more information on the bgzip tool: <http://www.htslib.org/doc/bgzip.html>.

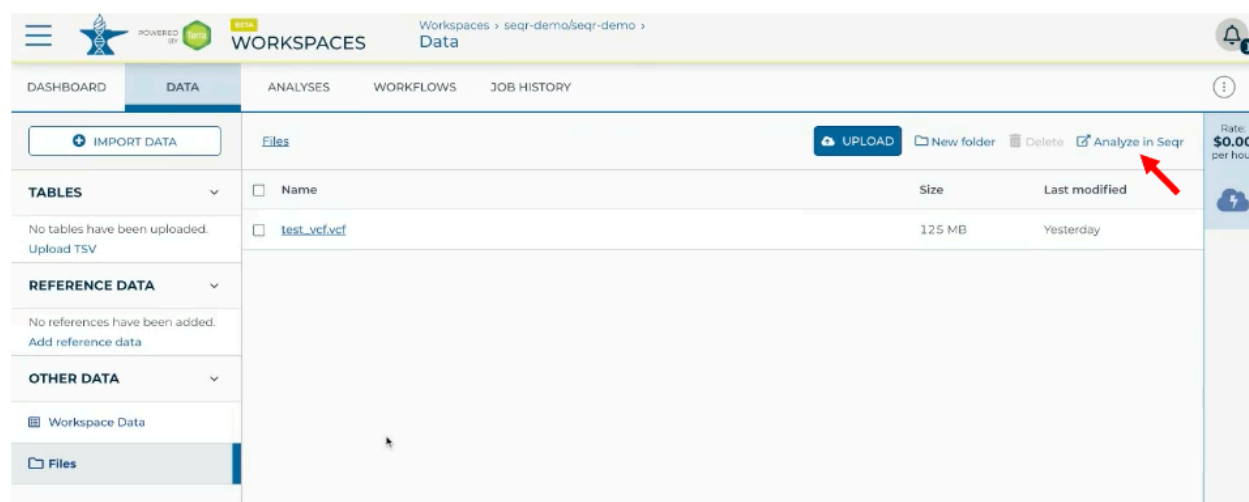
The joint-called VCF file must be stored in a workspace in which you have [Writer or Owner](#) level access and have the [Can Share](#) permissions. Additionally, the workspace must not be associated with any [Authorization Domains](#) in order for *seqr* to access it. If the workspace does not meet 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 AnVIL. To do this, log in to your AnVIL account, navigate to the Data section in your workspace, select Files, and then select Upload.



After the joint-called VCF is uploaded to a valid workspace in AnVIL, select *Analyze in seqr*.



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

The screenshot shows the 'Load data to seqr from AnVIL Workspace "seqr-demo/seqr-demo"' page. At the top, there's a navigation bar with 'seqr' and 'Summary Data' links, a search bar, and user information 'Logged in as Hana Snow' with a 'Log out' link. Below the navigation bar, the title 'Load data to seqr from AnVIL Workspace "seqr-demo/seqr-demo"' is centered. A light blue box contains the text: '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](#)'. Below this, there's a checkbox labeled 'By proceeding with seqr loading, I agree to grant seqr access to the data in this workspace', which is checked. At the bottom right, there's a blue 'Next >>' button. At the bottom left, there's a link: 'Need help? please submit: [GitHub Issues](#), or [Email Us](#)'.

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

The screenshot shows the 'Load data to seqr from AnVIL Workspace "seqr-demo/seqr-demo"' page, Step 2. The navigation bar is the same as in the previous screenshot. The title is the same. The light blue box with the documentation link is still present. Below it, there's a section 'Path to the Joint Called VCF' with a dropdown menu showing '/test\_vcfvcf'. Below the dropdown, there are two radio buttons for 'Sample Type': 'Exome' (selected) and 'Genome'. Below that, there are two radio buttons for 'Genome Version': 'GRCh37' and 'GRCh38' (selected). At the bottom right, there are two buttons: '<< Back' and a blue 'Next >>' button. At the bottom left, there's a link: 'Need help? please submit: [GitHub Issues](#), or [Email Us](#)'.

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.

The screenshot shows the 'Load data to seqr from AnVIL Workspace "seqr-demo/seqr-demo"' page, Step 3. The navigation bar is the same. The title is the same. The light blue box with the documentation link is still present. Below it, there's a section 'Project Description' with a text input field labeled 'Description'. Below that, there's a section 'Upload Pedigree Data' with a text area containing the text: '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.' At the bottom right of this text area, there's a blue 'I Agree' button. At the bottom right of the page, there are two buttons: '<< Back' and a blue 'Submit' button. At the bottom left, there's a link: 'Need help? please submit: [GitHub Issues](#), or [Email Us](#)'.

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

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 (.xls) 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 or Female, leave blank if unknown

**Affected Status** Affected or Unaffected, leave blank if unknown

**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

<< 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.

FAM36

VCGS\_FAM36\_115 VCGS\_FAM36\_116

VCGS\_FAM36\_114

Analysis Groups

Family Description

Analysis Status: Waiting for data

Assigned Analyst

Analysed By:

WES/WGS: +

gCNV/SV: +

RNAseq: +

Mitochondrial: +

STR: +

Case Notes Add Note +

Analysis Notes Add Note +

Matchmaker Notes Add Note +

Coded Phenotype

Post-discovery OMIM #

Age: Unknown

Age of Onset

Individual Notes

Consanguinity: Unknown

Other Affected Relatives: Unknown

Expected Mode of Inheritance

Assisted Reproduction:

Maternal Ancestry:

Paternal Ancestry:

Features:

Pre-discovery OMIM disorders:

Previously Tested Genes:

No Saved Variants

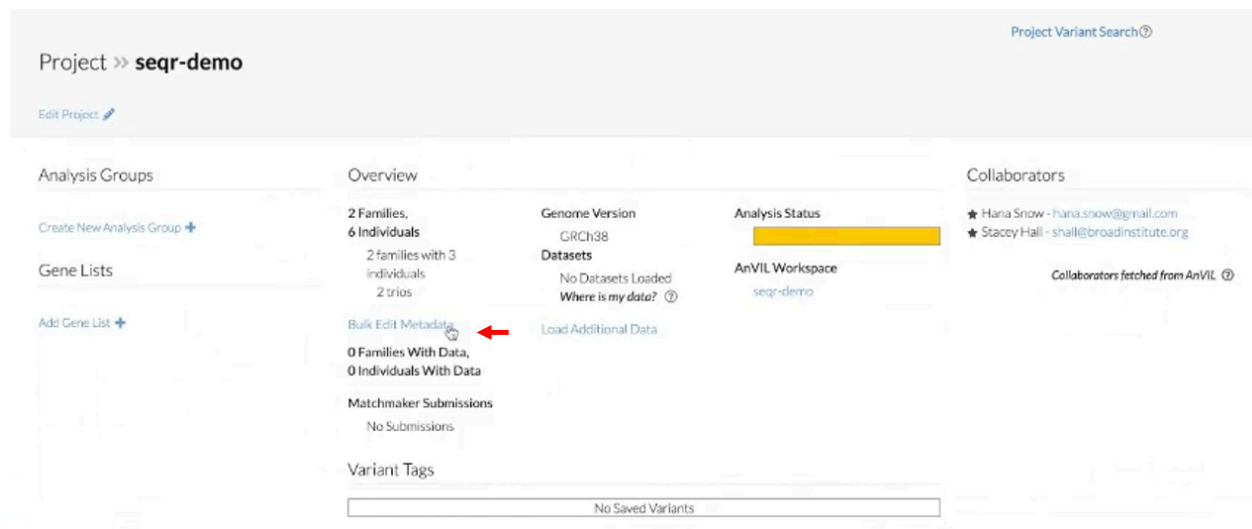
Variant Search

MatchMaker Exchange

VCGS\_FAM36\_114

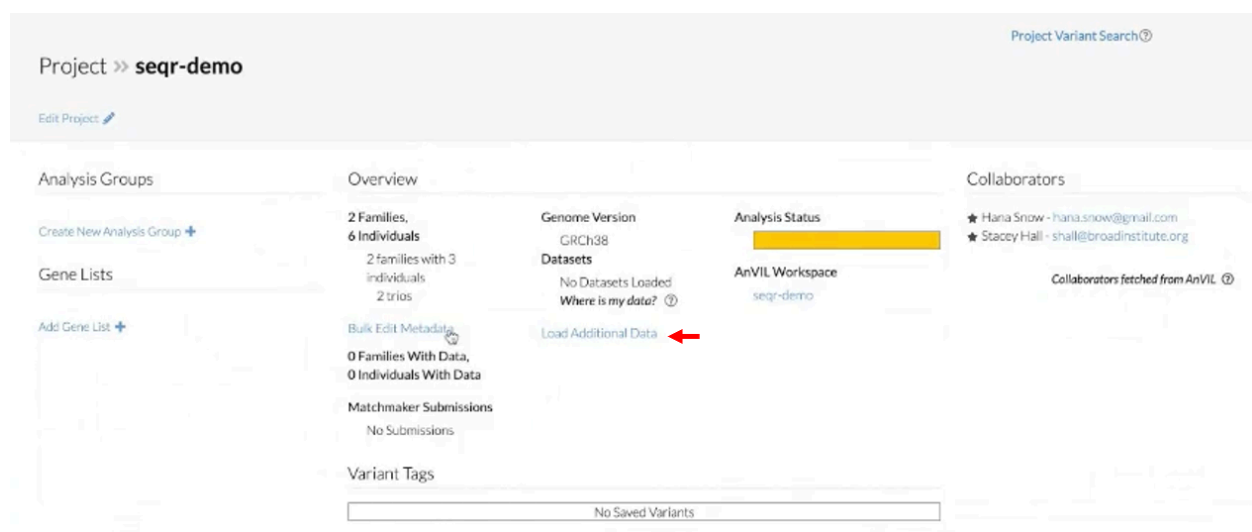
ADDED 2/8/2023

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



## 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.



The VCF you submit must be joint-called with all the data previously loaded in the project along with the new samples. This joint-called VCF must be in the same workspace associated with the *seqr* project. All notes and tags on the existing data will be maintained, if the samples are joint-called in the new VCF you upload.

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!