Loading data from an AnVIL workspace into *seqr*

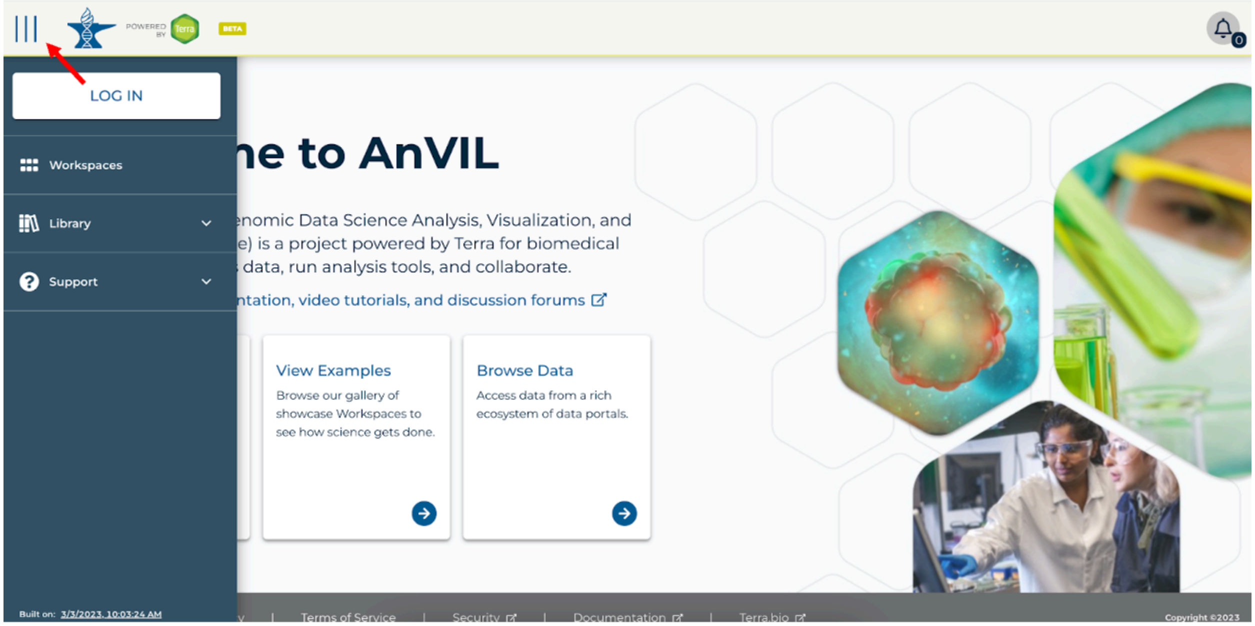
[Watch video tutorial]

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.



ne to AnVIL

enomic Data Science Analysis, Visualization, and
e) is a project powered by Terra for biomedical
data, run analysis tools, and collaborate.
ntation, video tutorials, and discussion forums

Browse our gallery of showcase Workspaces to see how science gets done.

Access data from a rich ecosystem of data portals.

Built on: 3/3/2023, 10:03:24 AM

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.

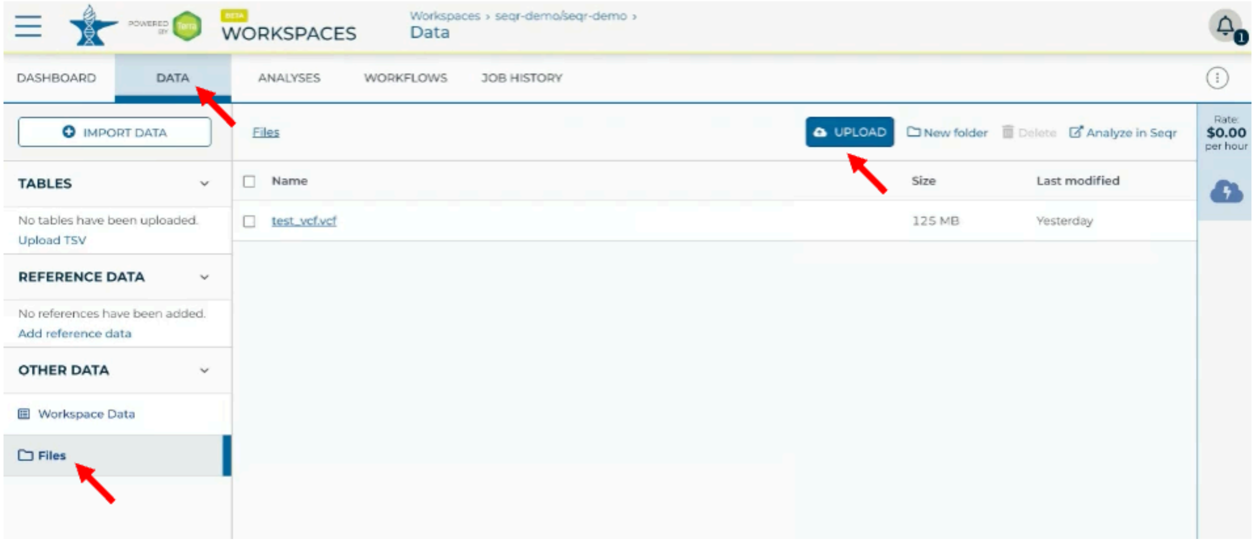
We also accept joint-called VCFs with a .gz extension provided they are internally bgzipped. Here is more information on the bgzip tool: http://www.htslib.org/doc/bgzip.html.

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, navigate to the Data section in your workspace, select Files, and then select Upload.

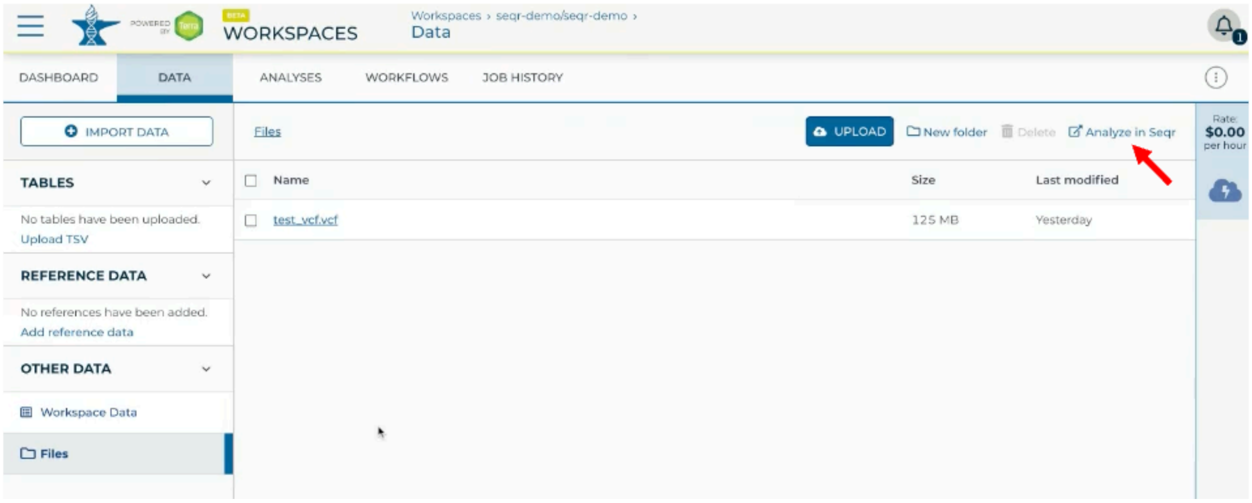


No tables have been uploaded.
Upload TSV

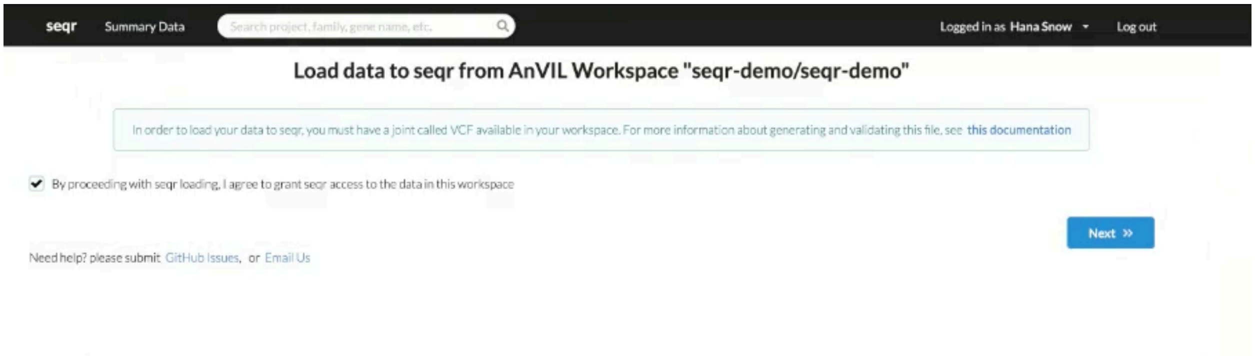
No references have been added.
Add reference data

	Name	Size	Last modified
□	test_vcf.vcf	125 MB	Yesterday

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.

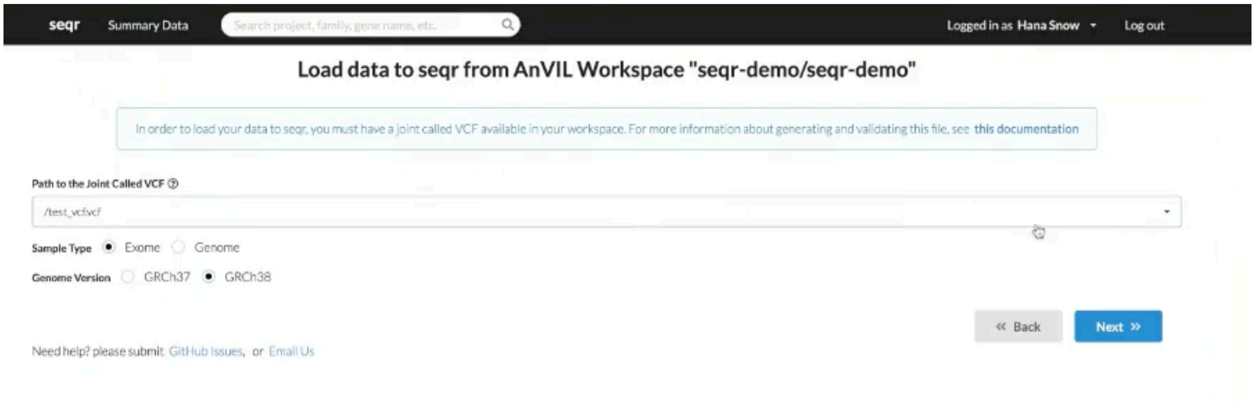

seqr
Summary Data
Search project, family, gene name, etc.
Logged in as Hana Snow
Log out
Next »
Need help? please submit GitHub Issues, or Email Us

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

☑ By proceeding with seqr loading, I agree to grant seqr access to the data in this 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
« Back
Next »
Need help? please submit GitHub Issues, or Email Us

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

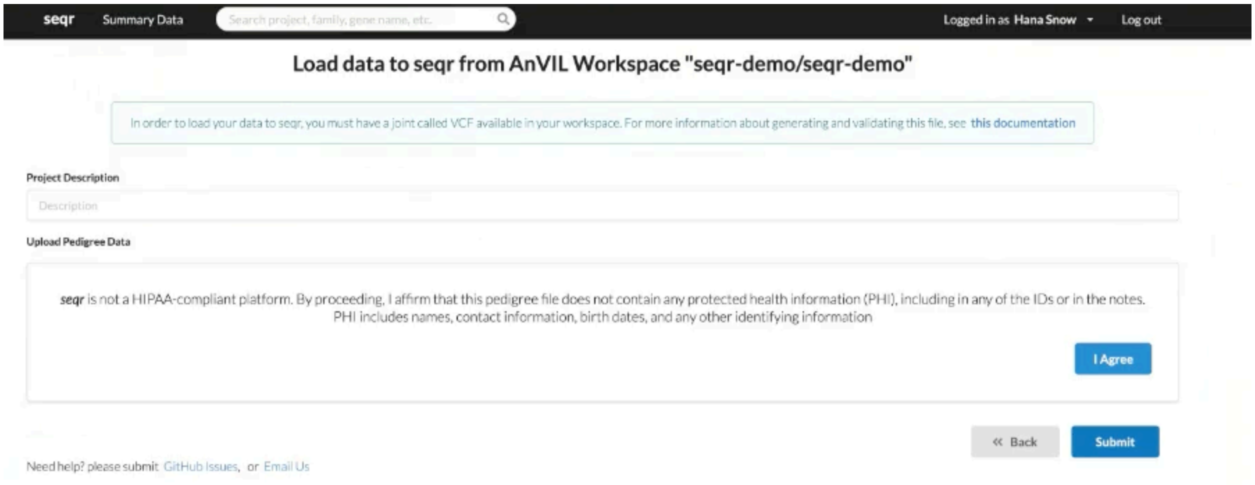
In order to load your data to seqr, you must have a joint called VCF available in your workspace. For more information about generating and validating this file, see this documentation

Path to the Joint Called VCF
/test_vcf.vcf

Sample Type ● Exome ○ Genome

Genome Version ○ GRCh37 ● GRCh38

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.



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

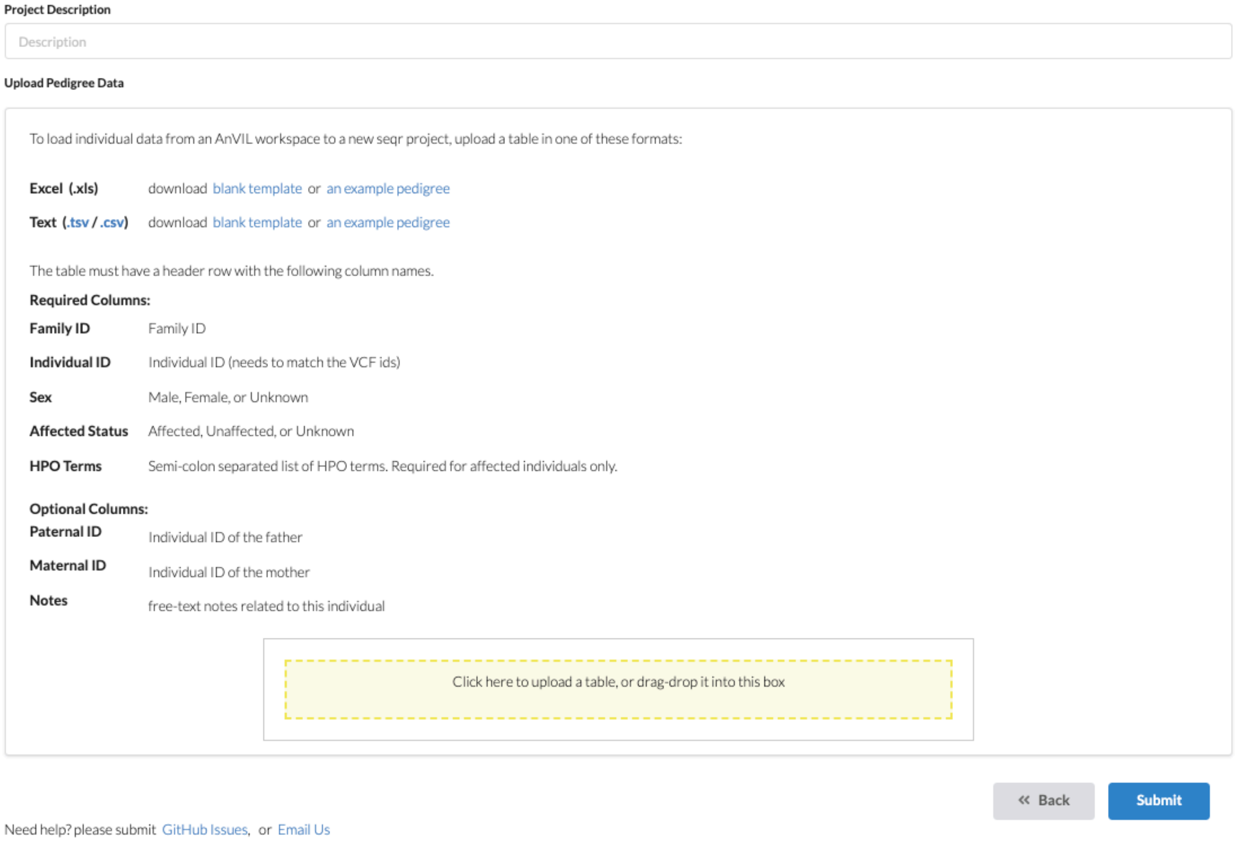
Description

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

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, 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:

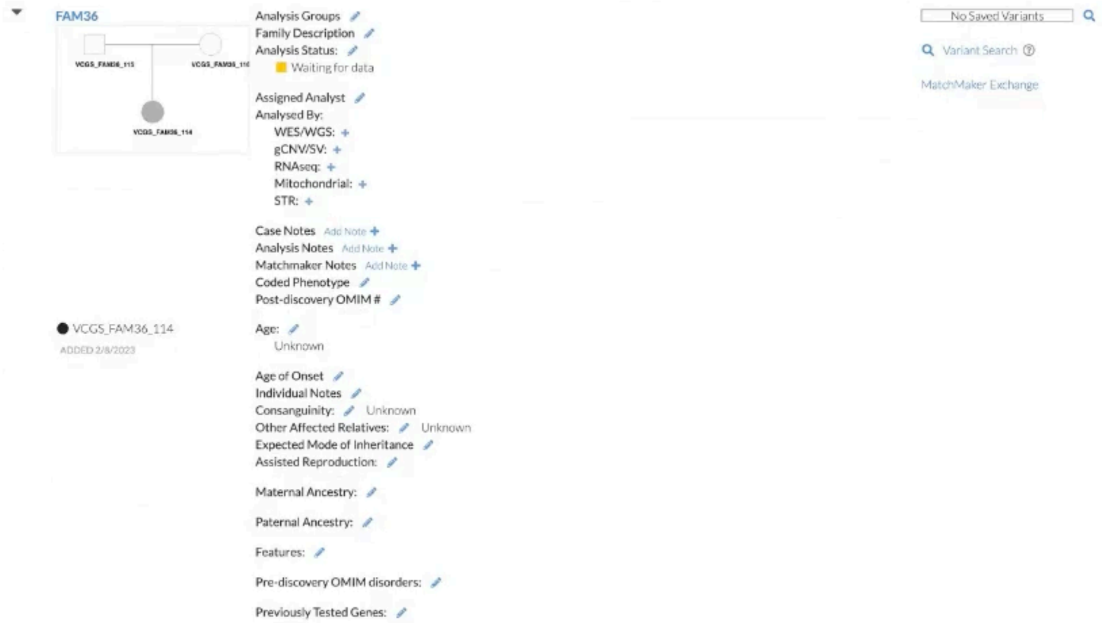
Paternal ID	Individual ID of the father
Maternal ID	Individual ID of the mother
Notes	free-text notes related to this individual

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

Need help? please submit GitHub Issues, or Email Us

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

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 #

● VCGS_FAM36_114
ADDED 2/8/2023

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:

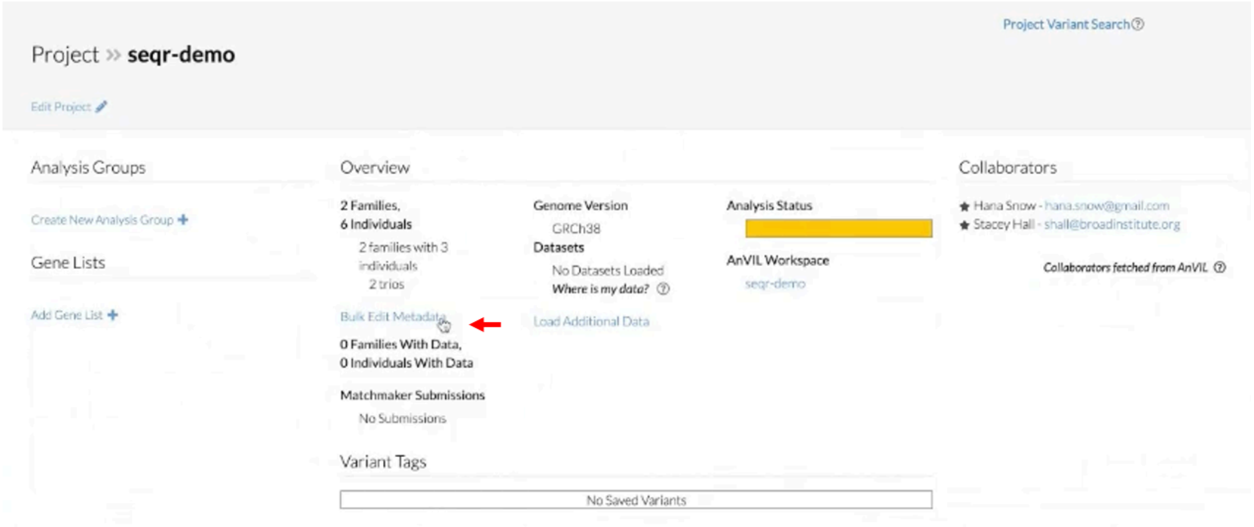
Previously Tested Genes:

No Saved Variants

Variant Search

MatchMaker Exchange

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



Project » **seqr-demo**

Edit Project

Project Variant Search

Analysis Groups

Create New Analysis Group +

Gene Lists

Add Gene List +

Overview

2 Families,
6 Individuals
2 families with 3 individuals
2 trios

Bulk Edit Metadata

0 Families With Data,
0 Individuals With Data

Matchmaker Submissions
No Submissions

Genome Version
GRCh38
Datasets
No Datasets Loaded
Where is my data?

Load Additional Data

Analysis Status

AnVIL Workspace
seqr-demo

Variant Tags

No Saved Variants

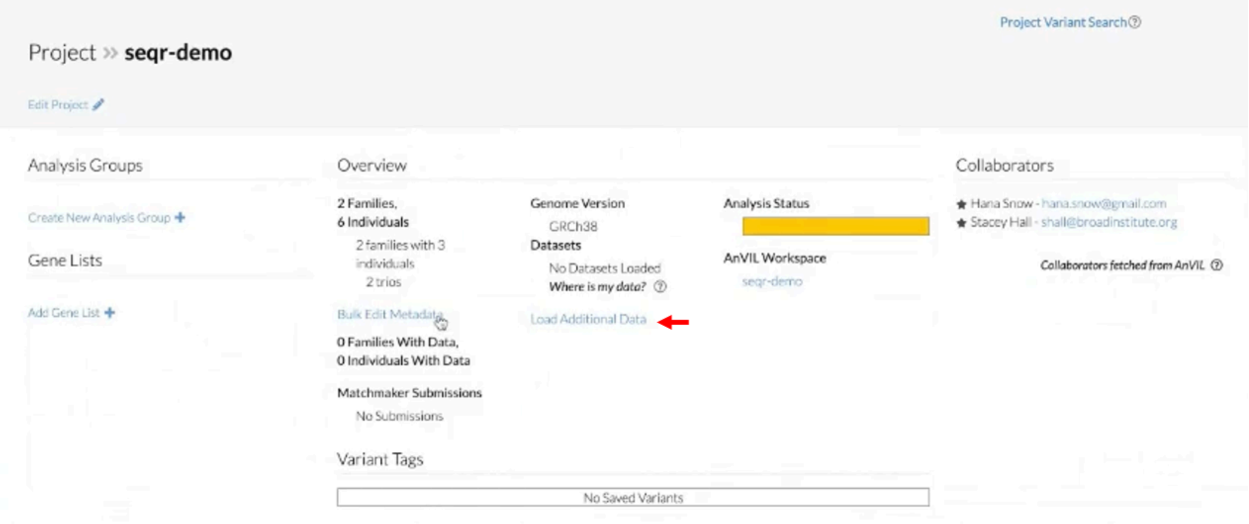
Collaborators

★ Hana Snow - hana.snow@gmail.com
★ Stacey Hall - shall@broadinstitute.org

Collaborators fetched from AnVIL

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.



Project » seqr-demo

Analysis Groups

Gene Lists

Overview

2 Families,
6 Individuals
2 families with 3 individuals
2 trios

0 Families With Data,
0 Individuals With Data

Matchmaker Submissions
No Submissions

Genome Version
GRCh38

Datasets
No Datasets Loaded

Analysis Status

AnVIL Workspace
seqr-demo

Collaborators
★ Hana Snow - hana.snow@gmail.com
★ Stacey Hall - shall@broadinstitute.org

Variant Tags
No Saved Variants

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!

Reviewed 07/18/2024