10/19/2023

PhenoDB on Cavatica
User Guide

Contents

PhenoDB Cavatica Overview

PhenoDB is a clinical research tool to analyze VCF files from individuals or families with suspected Mendelian disease. It is now available on Cavatica so data can be analyzed while it remains within the platform's security and computation resources.

Additional information about PhenoDB and its creators is available at available at https://phenodb.org/. Not all PhenoDB analysis functionality in the web version is implemented yet in the Cavatica app.

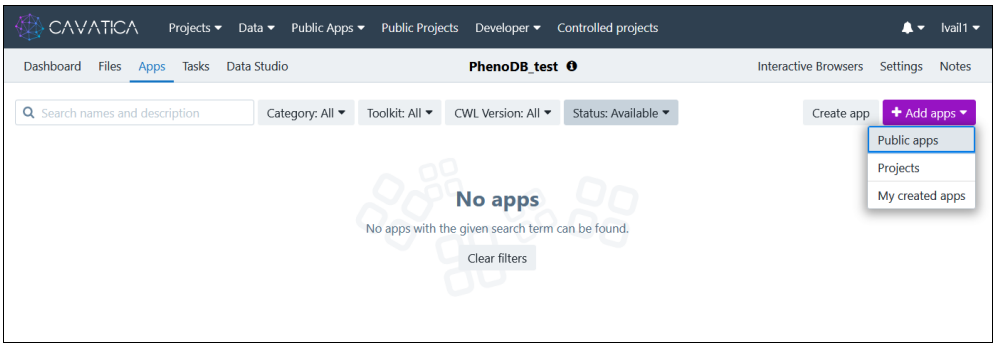
PhenoDB on Cavatica is run in two steps, each in a separate public Cavatica app.

1. PhenoDB ANNOVAR Annotation: annotates VCF files with parameters used for analysis
2. PhenoDB Analysis: reviews the annotated files, and finds variants consistent with user-selected inheritance patterns and other criteria.

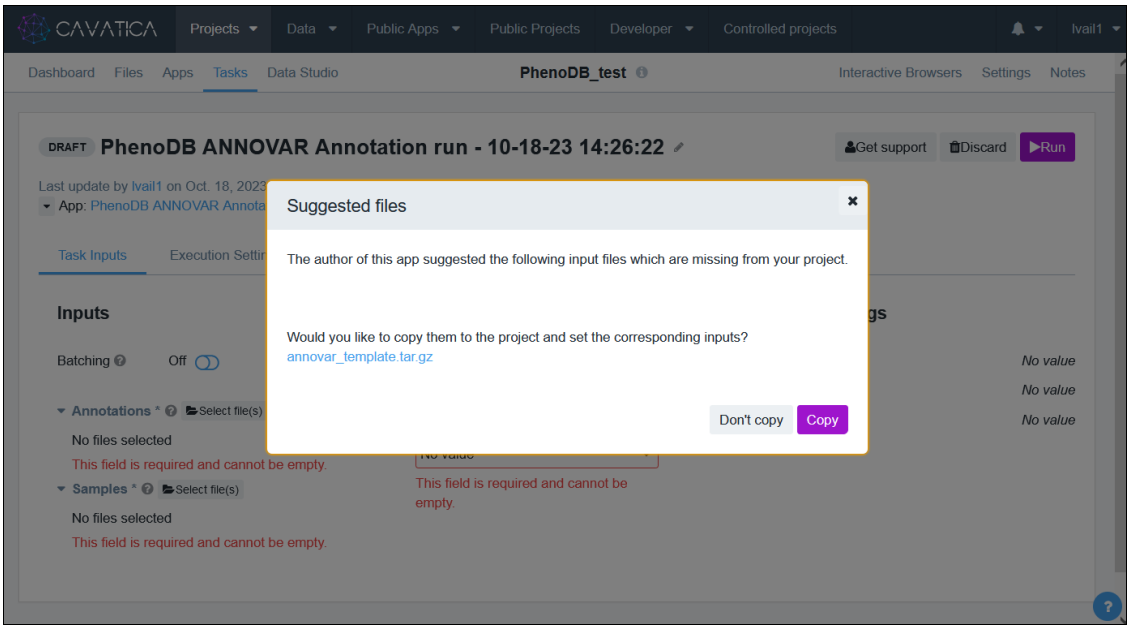
The example inputs used below are available in the public Cavatica project "Commit: PhenoDB App", which also stores the apps. The source code of the apps, its Dockerfile, and Common Workflow Language (CWL) specification are available at https://github.com/d3b-center/phenodb-cwl-app. For assistance, email phenodb@jhmi.edu and Laura Vail at lvail1@jhmi.edu.

Copy PhenoDB Public Apps Into Your Cavatica Project

1. In the project where you want to run PhenoDB analyses, open the Apps tab. In the upper left, click "Add Apps," and select "Public Apps" as the source.

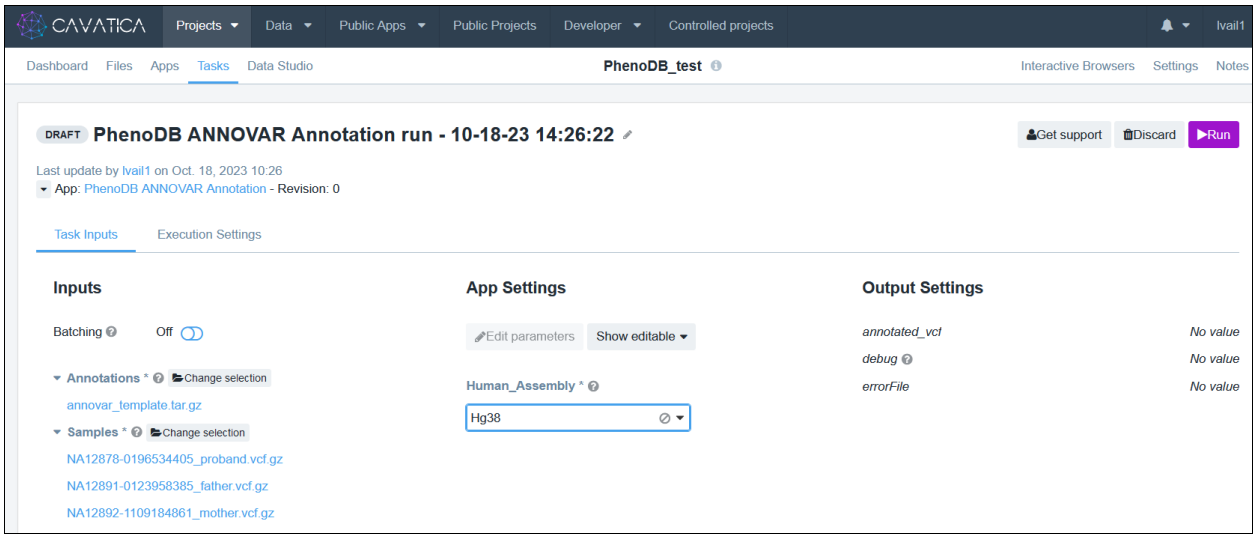


2. In the search bar, enter “PhenoDB” and two apps should show up. Click “Copy” for both of them. Close this search window.
3. Refresh your “Apps” page so the copied apps appear, and click “Run” on the PhenoDB ANNOVAR Annotation. A popup will ask whether to copy the ANNOVAR template file into your project: this is a required input, so click “Copy.” This will automatically populate the field “Annotations” for all runs.

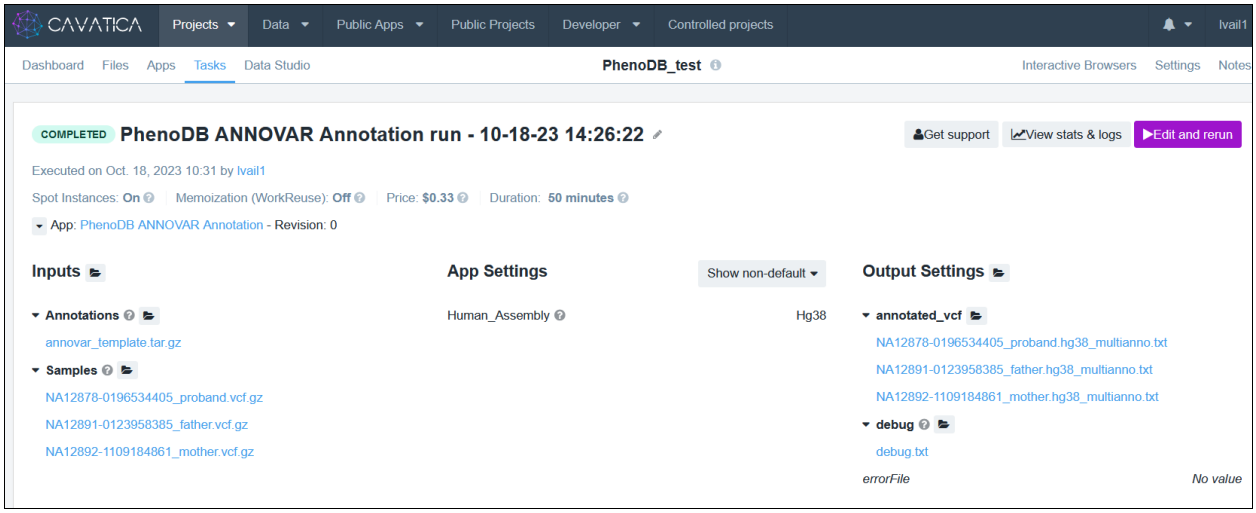


PhenoDB ANNOVAR Annotation Steps

1. Enter your inputs, descriptions below. The field “Annotations” should be automatically filled in when you open the app with a file from the public project. Each VCF can only contain data for a single sample.



2. Completing the task may take up to one hour or longer, depending on the size and number of the input files. When completed, there should be one new file for each VCF you entered, with "hg_38_multianno.txt" added to the end of the file's original name



3. These output files can then be analyzed with the PhenoDB Analysis app

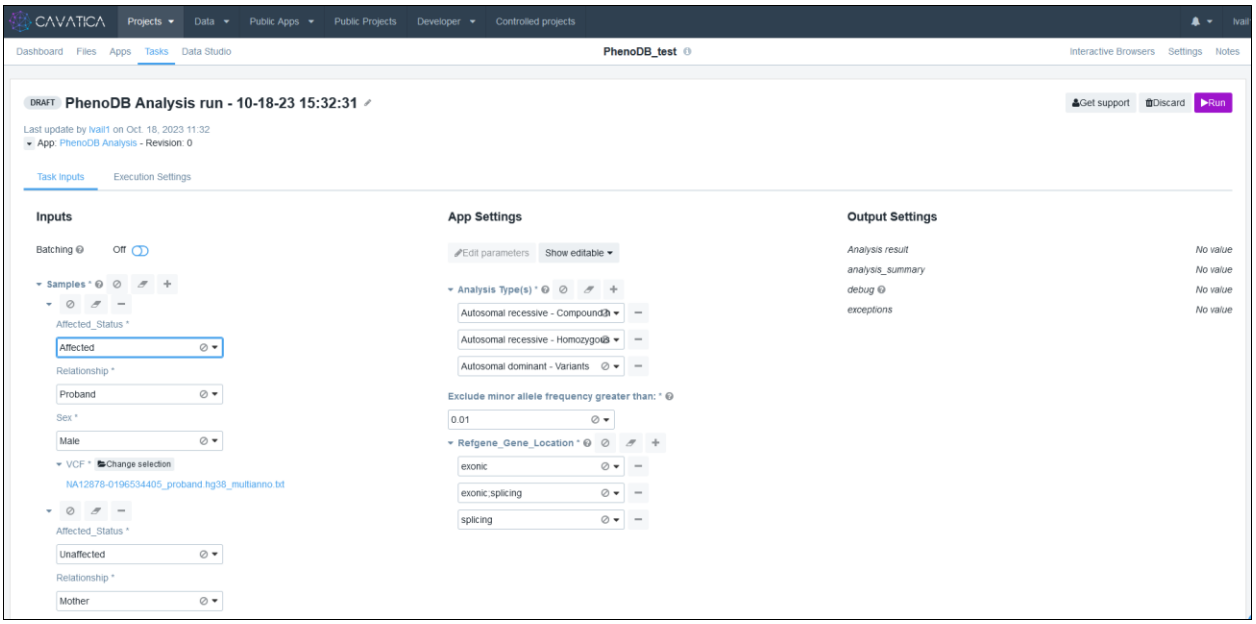
Annotation Inputs Description

- VCF File(s): Individual VCF files for all people included in the analysis
    - This step can be done for multiple families at once
    - The files can be either .gz compressed, or uncompressed
    - Set "Batching" to On for the VCFs, and select "Batch by: File." This helps the process run faster
- Human_Assembly: Select either Hg19 or Hg38
- Annotations: ANNOVAR reference data for the annotation and analysis. Only use the file "annovar_template.tar.gz" from the public project

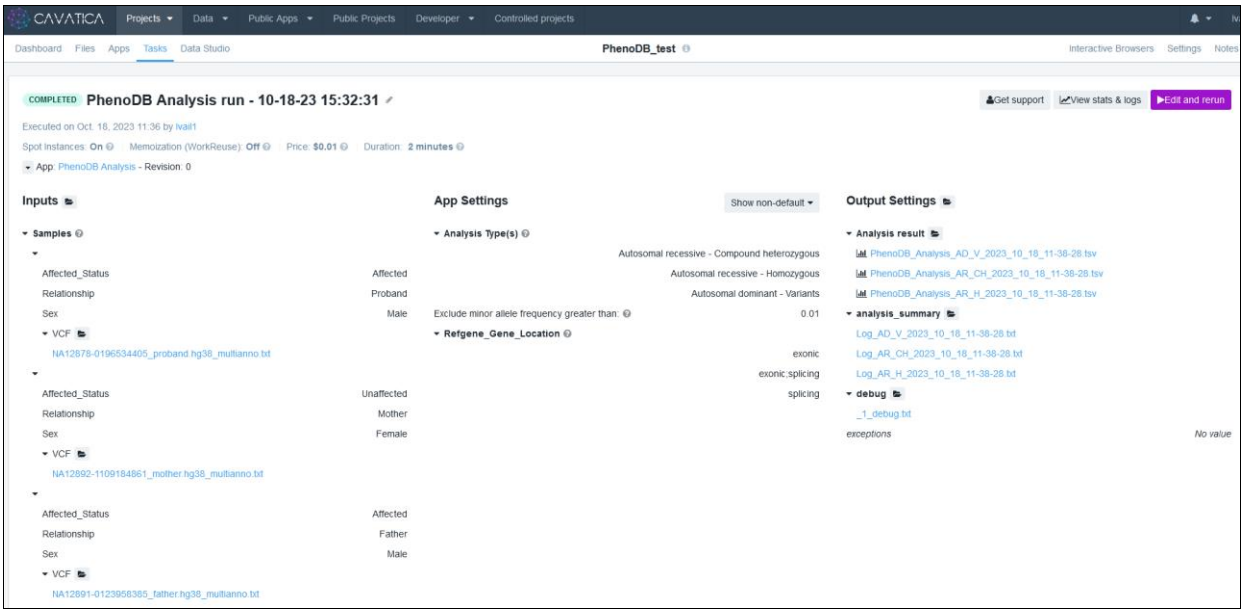
PhenoDB Analysis: Steps

1. Begin with annotated VCF files generated by the 'PhenoDB ANNOVAR Annotation' app

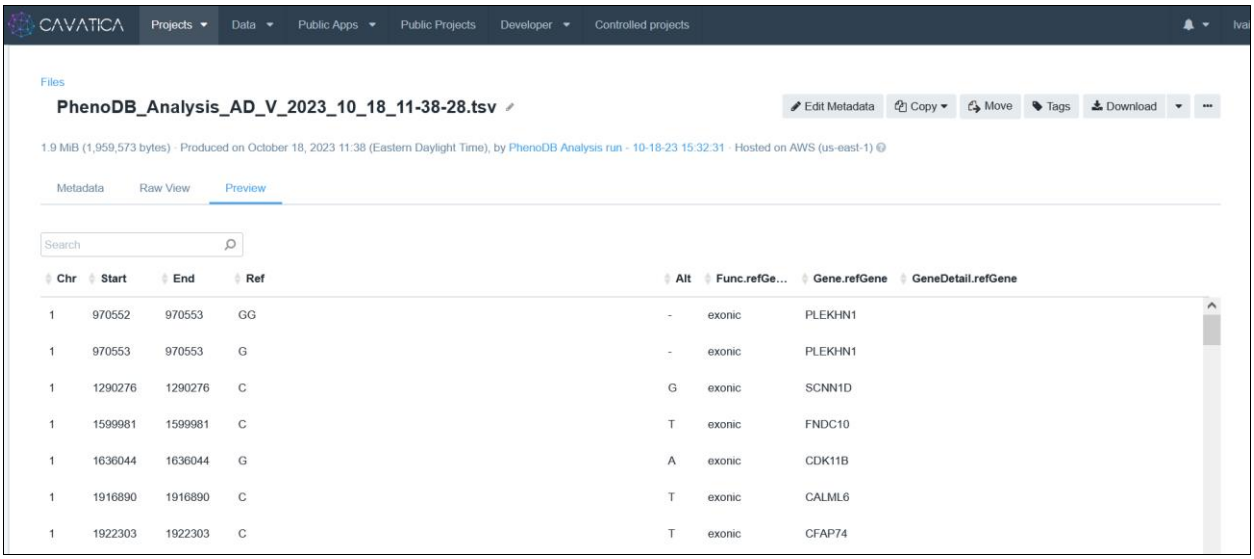
2. Enter your inputs, and descriptions of each item are below. Please note that there will be red text when you first click “+” to add a new Sample, and it will go away once all the fields for that Sample have been filled in. You can enter any number of Samples, but all should be related to the proband.



3. When complete, there should be two files for each selected analysis type: one ‘Analysis result’ tsv containing the variants of interest for that inheritance pattern, and one ‘Analysis summary’ file describing the filtering steps that led to the result.



4. Cavatica has built-in preview for text files, including column sort in the .tsv analysis results. Click on the names of the files, and select “Preview” to view and sort the .tsv files, and “Raw View” for the .txt files. In the .tsv preview, sort the results by clicking on column headers.



Analysis Inputs Description

- Samples: One or more individuals to be analyzed together. Each Sample has four attributes to be entered:
    - Affected Status: Affected, Unaffected, or Unknown
    - Relationship: Proband, or relationship to proband (Mother, Father, Other Relative)
    - Sex: Male, Female, Unknown or not XX/XY

    - VCF: Individual's annotated VCF file from the 'PhenoDB ANNOVAR Annotation' app, which end with "multianno.txt"
- Analysis Type(s): Choose one or more inheritance patterns for the analysis, described below. Recommended default: Autosomal recessive compound heterozygous and Autosomal recessive homozygous
- Exclude minor allele frequency greater than: Choose your cutoff point. Data sources are the ExAC, esp6500siv2, 1000g2014oct, and 1000g2015aug databases. Recommended default: 0.01
- RefGene gene location: Select one or more to include. Recommended default, three locations: "exonic", "exonic;splicing", and "splicing"
- Inheritance Types: Choose as appropriate for your analysis from the below descriptions. Recommended default:

Inhertiance Types: Description and Samples Required

AR_CH: Autosomal Recessive Compound Heterozygous

- Requires: At least proband
- Variants in Result: Includes only the heterozygous variants identified in the proband (assumed to be affected); if there is more than one affected family members, the analysis include only the variants that are identified in all affected members; next, the analysis includes only the genes that have more than one variant in the proband but if the same set of variants in a gene is found in one of the parents or in other unaffected family member then this gene (and its variants) is excluded of the analysis (Sobreira et al., 2015)

AR_H: Autosomal Recessive Homozygous

- Requires: At least proband
- Variants in Result: Identifies homozygous variants that are shared by all affected individuals and excludes variants that are homozygous in an unaffected individual (Sobreira et al., 2015)

AD_NM: Autosomal Dominant New Mutation

- Requires: Proband and two parents (trio), and neither parent is affected
- Variants in Result: Includes heterozygous variants that are identified in the proband but not in the parents (Sobreira et al., 2015)

AD_IM: Autosomal Dominant Inherited Mutation

- Requires: Proband plus at least one affected or unaffected relative, not necessarily a parent
- Variants in Result: Retains heterozygous variants that are shared by affected individuals and excludes those found in unaffected individuals (Sobreira et al., 2015)

AD_V: Autosomal Dominant Variant

- Requires: Should be used when only one individual is being analyzed.
- Variants in Result: Retains heterozygous variants with a minor allele frequency (MAF) less than the threshold selected for the ExAC, esp6500siv2, 1000g2014oct, and 1000g2015aug databases