

PhenoDB on Cavatica

User Guide

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 <https://phenodb.org/>. Not all PhenoDB analysis functionality in the web version is implemented yet in the Cavatica app.

PhenoDB Apps on Cavatica

PhenoDB on Cavatica is run in two steps, each in a separate Cavatica app available in this Project

1. PhenoDB ANNOVAR Annotation: annotates VCF files
2. PhenoDB Analysis: finds variants consistent with user-selected inheritance patterns and other parameters.

The example inputs used below are in the project's Files in folder 'PhenoDB_Sample_Files.' For assistance, email phenodb@jhmi.edu and Laura Vail (lvail1@jhmi.edu).

PhenoDB ANNOVAR Annotation: Steps

1. Begin with individual VCF files from one or more people
2. Enter your inputs, descriptions below

The screenshot shows the Cavatica web interface. At the top, there's a navigation bar with 'CAVATICA' logo and various menu items like 'Projects', 'Data', 'Public Apps', etc. Below this is a sub-header 'PhenoDB_dev'. The main content area is titled 'PhenoDB ANNOVAR Annotation run - 07-07-23 12:42:55'. It includes a 'DRAFT' status, a 'Last update by lvail1 on July 7, 2023 08:42' timestamp, and a dropdown for 'App: PhenoDB ANNOVAR Annotation - Revision: 10'. There are buttons for 'Get support', 'Discard', and 'Run'. The interface is divided into three main sections: 'Inputs', 'App Settings', and 'Output Settings'. The 'Inputs' section shows 'Batching' set to 'Off' and a list of VCF files under 'Annotations' and 'Samples'. The 'App Settings' section has 'Human_Assembly' set to 'Hg38'. The 'Output Settings' section shows 'annotated_vcf', 'debug', and 'errorFile' all set to 'No value'.

- 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

The screenshot shows the CAVATICA web interface. At the top, there's a navigation bar with 'CAVATICA' logo and tabs for Projects, Data, Public Apps, Public Projects, Developer, and Controlled projects. Below this is a sub-navigation bar with Dashboard, Files, Apps, Tasks (selected), and Data Studio. The main header shows 'PhenoDB_dev' and links for Interactive Browsers, Settings, and Notes.

The main content area displays a completed task: 'PhenoDB ANNOVAR Annotation run - 07-12-23 12:33:46'. It includes buttons for 'Get support', 'View stats & logs', and 'Edit and rerun'. Below the task title, it states 'Executed on July 12, 2023 08:34 by lvail1'. Further details include 'Spot Instances: On', 'Memoization (WorkReuse): On', 'Price: \$0.28', and 'Duration: 38 minutes'. The app used is 'PhenoDB ANNOVAR Annotation - Revision: 19'.

The interface is divided into three main sections: Inputs, App Settings, and Output Settings.

- Inputs:** Contains two expandable sections:
 - Annotations:** Shows 'Annovar_Template'.
 - Samples:** Lists two VCF files: 'NA12878-0196534405_proband.vcf' and 'NA12892-1109184861_mother.vcf.gz'.
- App Settings:** Shows 'Human_Assembly' and 'Hg38'.
- Output Settings:** Shows 'annotated_vcf' with two output files: '_1_NA12878-0196534405_proband.hg38_multianno.txt' and '_1_NA12892-1109184861_mother.hg38_multianno.txt'. There is also an 'errorFile' field with the value 'No value'.

- These output files can be analyzed with the PhenoDB Analysis app

PhenoDB ANNOVAR Annotation: Inputs

- 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 to use for the annotation. Select the folder ‘Annovar_Template’ in the project’s Files

PhenoDB Analysis: Steps

- Begin with annotated VCF files generated by the 'PhenoDB ANNOVAR Annotation' app
- Enter your inputs, descriptions below

CAVATICA Projects Data Public Apps Public Projects Developer Controlled projects

DRAFT PhenoDB Analysis run - 07-12-23 15:07:13 [Get support](#) [Discard](#) [Run](#)

Last update by [lvail1](#) on July 12, 2023 11:07
App: [PhenoDB Analysis](#) - Revision: 21

Task Inputs Execution Settings

Inputs

Batching [Off](#)

Samples

Affected_Status *

Affected

Relationship *

Proband

Sex *

Male

VCF * [Change selection](#)

NA12878-0196534405_proband.hg38_multianno.txt

Affected_Status *

Unaffected

Relationship *

Mother

Sex *

Female

VCF * [Change selection](#)

NA12892-1109184861_mother.hg38_multianno.txt

App Settings

[Edit parameters](#) [Show editable](#)

Analysis Type(s)

Autosomal recessive - Compound heterozygous

Autosomal recessive - Homozygous

Exclude minor allele frequency greater than: *

0.01

Refgene_Gene_Location

exonic

exonic;splicing

splicing

Output Settings

Analysis result No value

analysis_summary No value

exceptions No value

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 of the variants that led to that result

CAVATICA Projects Data Public Apps Public Projects Developer Controlled projects

Dashboard Files Apps **Tasks** Data Studio **PhenoDB_dev** Interactive Browsers Settings Notes

COMPLETED PhenoDB Analysis run - 07-12-23 12:35:37 [Get support](#) [View stats & logs](#) [Edit and rerun](#)

Executed on July 12, 2023 08:37 by [lvail1](#)
Spot Instances: [On](#) Memoization (WorkReuse): [On](#) Price: \$0.01 Duration: 2 minutes
App: [PhenoDB Analysis](#) - Revision: 20

Inputs

Samples

Affected_Status Affected

Relationship Proband

Sex Male

VCF [Change selection](#)

NA12878-0196534405_proband.hg38_multianno.txt

Affected_Status Unaffected

Relationship Mother

Sex Female

VCF [Change selection](#)

NA12892-1109184861_mother.hg38_multianno.txt

App Settings

[Show non-default](#)

Analysis Type(s)

Autosomal recessive - Compound heterozygous

Autosomal recessive - Homozygous

Exclude minor allele frequency greater than: 0.01

Refgene_Gene_Location

exonic

exonic;splicing

splicing

Output Settings

Analysis result

[PhenoDB_Analysis_AR_CH_2023_07_12_08-38-55.tsv](#)

[PhenoDB_Analysis_AR_H_2023_07_12_08-38-55.tsv](#)

analysis_summary

[Log_AR_CH_2023_07_12_08-38-55.txt](#)

[Log_AR_H_2023_07_12_08-38-55.txt](#)

debug

exceptions No value

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, dynamically sort the results by clicking on column headers

CAVATICA Projects Data Public Apps Public Projects Developer Controlled projects

Dashboard Files Apps Tasks Data Studio PhenoDB_dev Interactive Browsers Settings Note

Files

PhenoDB_Analysis_AD_NM_2023_07_07_13-11-12.tsv Edit Metadata Copy Move Tags Download

227.9 KiB (233,385 bytes) · Produced on July 7, 2023 13:11 (Eastern Daylight Time), by PhenoDB Analysis run - 07-07-23 17:08:19 · Hosted on AWS (us-east-1)

Metadata Raw View **Preview**

Search

Chr	Start	End	Ref	Alt
19	501715	501762	GGAGTCTCCCGACACACCTCCCGGAGTCTCCCGACACACCTCCCA	-
19	501717	501738	AGTCTCCCGACACACCTCCCG	0
19	501718	501755	GTCTCCCGACACACCTCCCGGAGTCTCCCGACACCA	0
19	501719	501791	TCTCCCGACACACCTCCCGGAGTCTCCCGACACACCTCCAGGAGCTCCCGACACACCTCCCGGAGC	0
19	501720	501743	CTCCCGACACACCTCCCGGAGT	0

PhenoDB Analysis: Inputs

- 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
- 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:

Inheritance 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