PhenoDB and CAVATICA: Container-Based Annotation and Analysis

L. Vail[1], S. Griffith[1], R. Martin[1], M. Brown[2], D. Miller[2], Y. Guo[2], A. Heath[2], Y. Zhu[2], N. Sobreira[1]

[1] Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[2] Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, USA

Introduction

PhenoDB is a web-based application for investigators to run analyses on exome or genome VCF files from individuals or families with suspected Mendelian disease.

In partnership with Seven Bridges Genomics, The Center for Data-Driven Discovery in Biomedicine (D^3b) developed **CAVATICA**, a cloud-based data analysis and sharing platform that integrates access to pediatric genomics datasets.

PhenoDB and CAVATICA staff collaborated to add the PhenoDB analysis module to CAVATICA.

Benefits of Collaboration

- CAVATICA provides a HIPAA-compliant, secure, and scalable environment for storing data and collaborating to analyze those data. CAVATICA implementation allows current users to take advantage of some of the benefits of using PhenoDB without duplicating data across multiple platforms.
- While the Johns Hopkins-managed PhenoDB instances meet a broad range of security requirements, they might not directly serve the needs of all research studies that would benefit from the PhenoDB analysis module.

Workflow Details

Requirements to use PhenoDB on CAVATICA: a CAVATICA account, an exome or genome VCF file from a proband with suspected Mendelian disease, and (optionally) VCFs from other family members.

Workflow Steps

0. Copy the two PhenoDB public apps into a Project where you can run Tasks.

1. Run 'PhenoDB ANNOVAR Annotation.' Choose your VCF files and select whether to use Hg19 or Hg38. This step generates annotated files for the analysis step.

2. Run 'PhenoDB Analysis.' Select the annotated files from the first step, specifying the following for each one: the individual's affectedness status, gender, and relationship to the proband. Select analysis parameters (listed below).

3. Review the results. For each inheritance type selected, PhenoDB on CAVATICA provides the variants consistent with that inheritance mode. A log file is also available for each analysis performed listing the filtering steps done on the proband's VCF file to provided the final candidate variant list per mode of inheritance selected.

Analysis Parameters:

1. Minor Allele Frequency Cutoff
2. Refgene Gene Location
3. Inheritance Types:
   - Autosomal Dominant – Inherited Mutation (the affected proband and one or more affected or unaffected family members)
   - Autosomal Dominant – New Mutation (the proband and their parents)
   - Autosomal Recessive – Compound Heterozygous
   - Autosomal Recessive – Homozygous

Technical Implementation Details

CAVATICA

- Common Workflow Language (CWL) definitions provide the interface between CAVATICA and app's Docker containers for accepting inputs and returning output files.
- Graphical user interfaces in CAVATICA are automatically generated based on the inputs defined in an app's CWL.

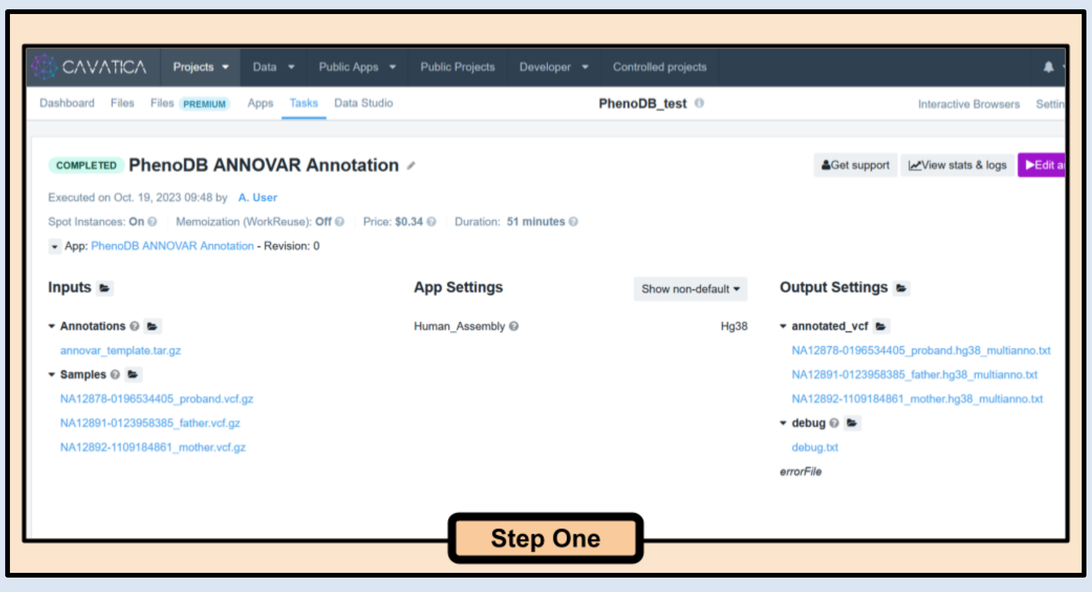
PhenoDB Apps for CAVATICA

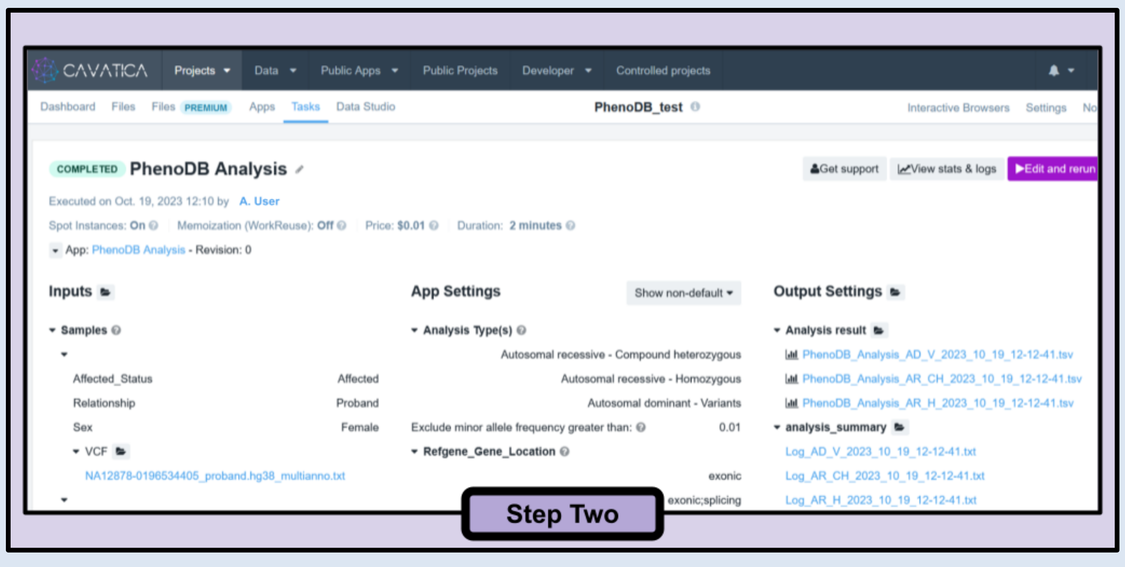
- Built into Docker containers with an Ubuntu 20.04 base image.
- Written in Python 3, with no external libraries or frameworks.
- Encapsulate the ANNOVAR bioinformatics tool, written in Perl, in the annotation Docker container.
- Accesses ANNOVAR databases (100 GB in total) stored on CAVATICA's cloud storage, copying these data into the container at run time.

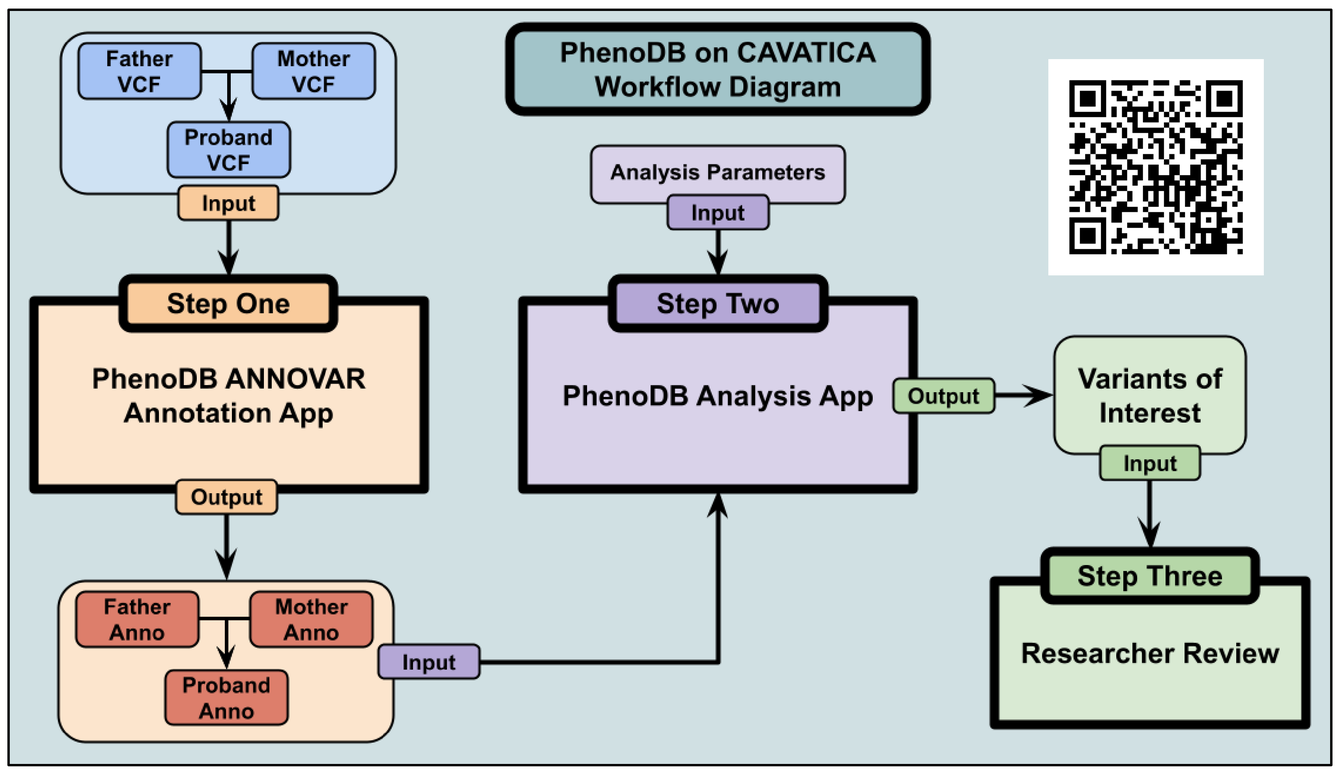
Implementation Challenges

- The cloud environment on CAVATICA was difficult to replicate locally for rapid tests. Developing apps for CAVATICA involves variables such as the different Amazon Web Services (AWS) resources that host the running container and its reference data files, the Docker image in a remote repository and CWL version on AWS that are updated independently and need to work in concert, and the CWL then Docker layers a run command passes through before executing.
- The databases required by ANNOVAR are of considerable size (100 GB uncompressed) requiring careful consideration of storage location and whether to compress these files.
- CAVATICA public apps have different restrictions from apps in individual Projects, in particular public apps can only copy over individual files, not folders. This meant that a simpler initial implementation with the ANNOVAR reference data stored as 98 files could not be used for the public app.

CAVATICA App User Interfaces







Next Steps

- Implement additional analysis features from the PhenoDB web app: X-linked dominant mode of inheritance, maternal and paternal imprinting mode of inheritance, gene exclusion, coordinate restriction, and cohort analysis.
- Explore collaboration with other cloud-based genomics analysis platform providers to bring the PhenoDB analysis module to those platforms via the extant containers created for the CAVATICA collaboration.

References

1 - PhenoDB, GeneMatcher and VariantMatcher, tools for analysis and sharing of sequence data. Wohler E, Martin R, Griffith S, Rodrigues EDS, Antonescu C, Posey JE, Coban-Akdemir Z, Jhangiani SN, Doheny KF, Lupski JR, Valle D, Hamosh A, Sobreira N. Orphanet J Rare Dis. 2021 Aug 18;16(1):365.

2 - New tools for Mendelian disease gene identification: PhenoDB variant analysis module; and GeneMatcher, a web-based tool for linking investigators with an interest in the same gene. Sobreira N, Schiettecatte F, Boehm C, Valle D, Hamosh A. Hum Mutat. 2015 Apr;36(4):425-31.