Kids First Variant Workbench

## The Short Pitch

The Variant Workbench on CAVATICA fulfills the need of **rapid hypothesis testing** by allowing users to easily combine patient clinical data, genomic data, and variant annotation in a single workspace.

## The Long Pitch

The Kids First Data Resource has grown to more than 30 studies, 25,000 participants, and over 1.4 PB of genomic data. As the number of investigators who have learned about Kids First and applied for access increases, so too does the diversity in analytical skills of our user base. Some researchers are comfortable and confident with bioinformatic analysis and are able to intuitively use the data resource to answer biomedical research questions while others come from medical or “wet bench” backgrounds with relevant questions but not the skills to answer them quickly.

Investigators who contact the Kids First DRC for additional support often bring with them the same three questions they are looking to answer, with details adjusted for their unique research interests. The research process is slowed because the answers to these questions are found in different locations and must be manually connected together to make biologically meaningful findings.

1. ***Which Kids First participants have the condition I am researching?***
   1. *This question can be answered using the Kids First Portal’s Explore Data tool. The investigator can apply filters for HPO and MONDO terms to select participants with the conditions they study. At that point, the genomic files for those participants can be pushed to CAVATICA for analysis. Alternatively, the clinical data can be downloaded as a multi-tab Excel file which contains the conditions assigned to each participant; this file can then be used to filter a list of participants in CAVATICA.*
2. ***Of those participants, which have variants in genes I hypothesize to be causative of the given condition?***
   1. *This question can be answered using CAVATICA’s Data Studio environment. The researcher might query the VCFs with a command-line tool such as* bcftools*, or use a notebook-based analysis in* Python *or* R*. Either of these is dependent upon the investigator being comfortable using these methods for data analysis.*
3. ***Of those variants, which are predicted to affect the protein’s function and therefore may cause the condition?***
   1. *This answer can be answered using the Kids First Portal’s Variant Search tool, which compiles data from variant annotation databases such as CADD and ClinVar. While compiled in one place, this system is low-throughput and manual and would only be suitable for querying a handful of variants. Alternatively, the researcher could access these variant annotation sources directly and import the information into their CAVATICA project for direct comparison.*

The above workflow does not support rapid hypothesis testing using Kids First data [see appendix for general steps]. Because the information to answer these is in multiple locations across multiple sites, a user cannot easily explore the data resource and answer a simple research question quickly. Instead they must navigate to each of these locations and piece the data together to slowly unravel findings. What should be a simple exploratory question can take weeks to finally answer.

Compare this workflow to a tool such as [cBioPortal](https://www.cbioportal.org/). Participants can be quickly filtered based on diagnosis on the study page. Those with somatic mutations in a gene of interest are identified by using a text-based search box. This flows neatly to a list of the mutations with provided annotations from relevant databases OncoKB and Cancer Hotspots. cBioPortal is not the simplest tool to use, but once someone learns how to use it, answering multiple questions rapidly becomes very easy. The Kids First workflow is also not the simplest to learn, but once someone does understand this dataflow, they must then repeat this onerous process for each research question they wish to explore.

The Variant Workbench on CAVATICA fulfills the need of **rapid hypothesis testing** by allowing users to easily combine patient clinical data, genomic data, and variant annotation in a single workspace. By making it easy to answer the three questions described above, the Variant Workbench will empower more researchers to explore Kids First data and facilitate scientific discoveries in pediatric cancer and structural birth defects.

Example Research Questions:

1. Are there variants in the gene TNK1 present in probands with cardiovascular defects?
2. What are the phenotypic characteristics (HPO terms) of subjects with variants in the BSN gene?
3. Are variants in the gene *AMOTL1* present in participants with conditions that affect craniofacial development (orofacial clefting, craniosynotosis, craniofacial microsomia)? Are these variants potentially causal?
4. Are there variants in the gene MLRP5 present in participants that are mothers of probands with CADD>20 in the neuroblastoma study?

## Cloud Credits

Kids First is offering cloud credits to cover the cost of your analysis. Apply here: <https://kidsfirstdrc.org/news/kids_first-cloud_credits>

