



# NHLBI BioData Catalyst BB-EIGHT (Penetrance API) User Guide

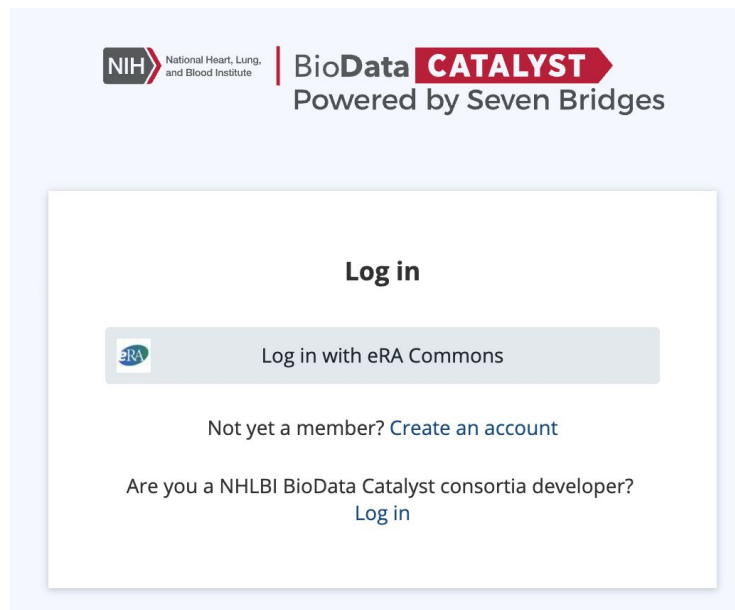
As part of the [BioData Catalyst](#) initiative, researchers at Harvard Medical School and Boston Children's Hospital have developed a computational tool called **BB-EIGHT: Beta-Binomial Estimation Interface for Genetic Risk across Human Traits**. BB-EIGHT addresses a ubiquitous challenge in the clinical use of genetic variation: the lack of quantitative risk estimates (penetrance) for pathogenic variation used in the clinic. While recent work has improved the consistency of variant classifications across laboratories and started to quantify penetrance in select instances, these efforts have largely focused on *genotype* information and molecular properties of individual variants (e.g. sequence conservation, deleteriousness). Complementing these efforts, BB-EIGHT allows investigators to bring *phenotypic* and *demographic* factors into focus which can influence estimates of risk just as strongly. Such factors include case and control disease definitions, demographic factors (e.g. age, gender, race/ethnicity), and the presence of relevant comorbidities.

BB-EIGHT enables investigators to understand genetic risk using multiple clinical and genomic datasets from TOPMed and TOPMed-related studies funded by the National Heart Lung and Blood Institute (NHLBI). The API allows an investigator to specify dynamic phenotype and genotype criteria and calculate penetrance in real time, map and visualize penetrance distributions across different populations across multiple cohorts, and identify individual variants that are high frequency in particular populations. As a focused example, we show how this approach can be used to quantify penetrance for hypertrophic cardiomyopathy (HCM) across several BDC cohorts.

# Authorization and Access

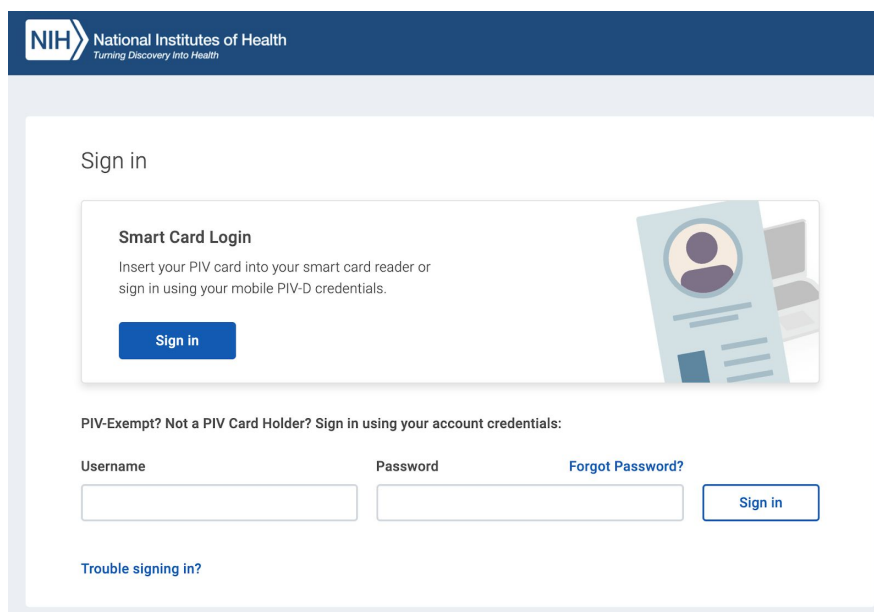
To use BB-EIGHT on BioData Catalyst:

1. You must have an NIH eRA commons account or an NIH username and password. [Please see these instructions.](#)
2. You must have an active dbGaP Data Access Request Approval, for more information on how to obtain access to data please visit the [BioData Catalyst Data Access webpage.](#)
3. Navigate to <https://accounts.sb.biodatacatalyst.nhlbi.nih.gov/>.
4. You will be directed to the log-in page where you can log-in using your NIH eRA commons account information.



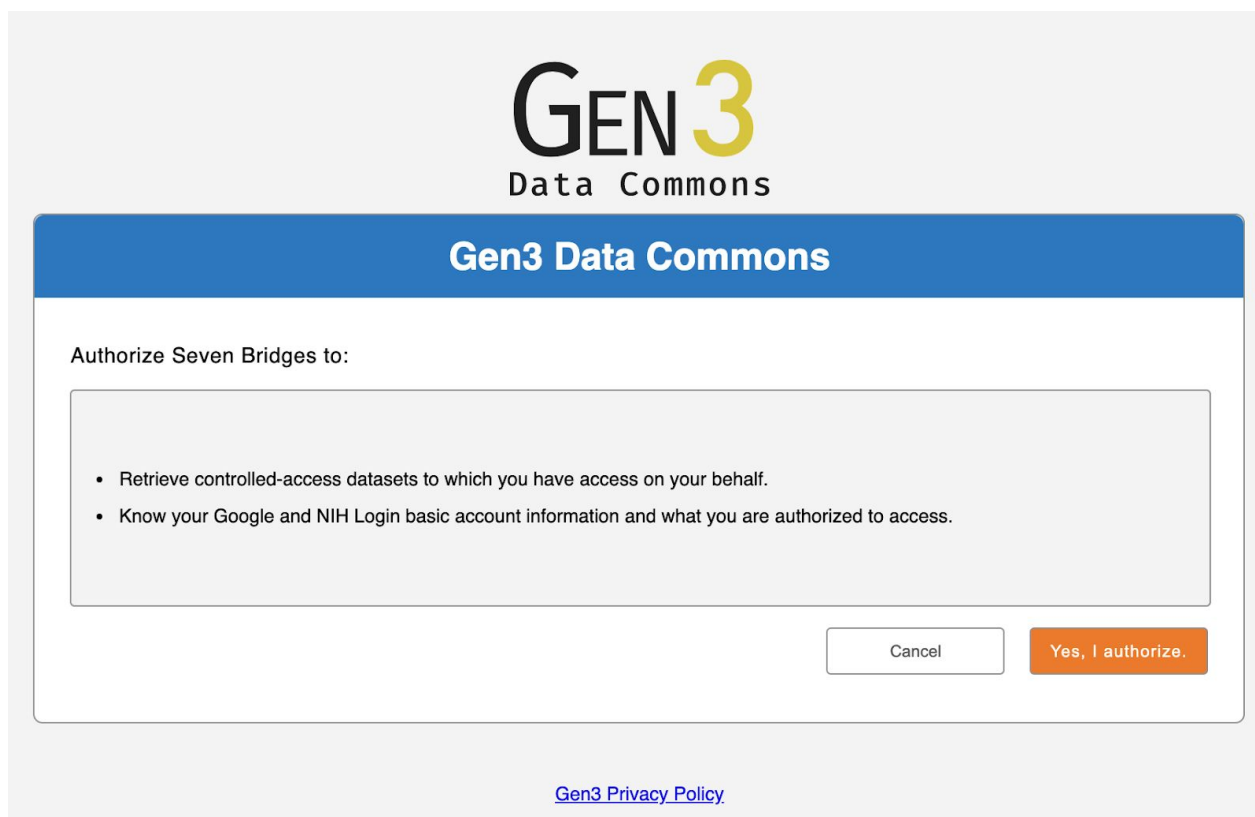
The image shows the BioData CATALYST login page. At the top, there is a header with the NIH logo (National Heart, Lung, and Blood Institute) and the BioData CATALYST logo, which includes the text "Powered by Seven Bridges". Below the header, the main content area is titled "Log in". It features a button labeled "Log in with eRA Commons" with a small eRA Commons icon to its left. Below this button, there is a link "Not yet a member? Create an account". At the bottom, there is a question "Are you a NHLBI BioData Catalyst consortia developer?" followed by a "Log in" link.

5. You will then be directed to the NIH website to log-in with your eRA commons credentials. Enter your credentials and you will be directed back to the BioData Catalyst authorization page.



The image shows the NIH Sign in page. At the top, there is a header with the NIH logo (National Institutes of Health) and the tagline "Turning Discovery Into Health". Below the header, the main content area is titled "Sign in". It features a "Smart Card Login" section with the text "Insert your PIV card into your smart card reader or sign in using your mobile PIV-D credentials." and a "Sign in" button. To the right of this section is an illustration of a smart card. Below the Smart Card Login section, there is a link "PIV-Exempt? Not a PIV Card Holder? Sign in using your account credentials:". This is followed by a form with two input fields: "Username" and "Password". To the right of the Password field is a link "Forgot Password?". Below the form is a "Sign in" button. At the bottom left, there is a link "Trouble signing in?".

6. You will be asked to authorize BioData Catalyst-SB integration to know your account information and what you are authorized to access. This process allows for the SB User Interface on the BioData Catalyst Ecosystem to know the data are authorized to access.



The image shows a web-based authorization dialog box for Gen3 Data Commons. At the top, the Gen3 logo is displayed with 'GEN' in black and '3' in yellow, followed by 'Data Commons' in black. Below this is a blue header bar with the text 'Gen3 Data Commons' in white. The main content area is white and contains the text 'Authorize Seven Bridges to:' followed by a list of two bullet points: 'Retrieve controlled-access datasets to which you have access on your behalf.' and 'Know your Google and NIH Login basic account information and what you are authorized to access.' At the bottom right of the dialog are two buttons: a white 'Cancel' button and an orange 'Yes, I authorize.' button. At the very bottom of the dialog, there is a blue link for 'Gen3 Privacy Policy'.

GEN3  
Data Commons

Gen3 Data Commons

Authorize Seven Bridges to:

- Retrieve controlled-access datasets to which you have access on your behalf.
- Know your Google and NIH Login basic account information and what you are authorized to access.

Cancel Yes, I authorize.

[Gen3 Privacy Policy](#)

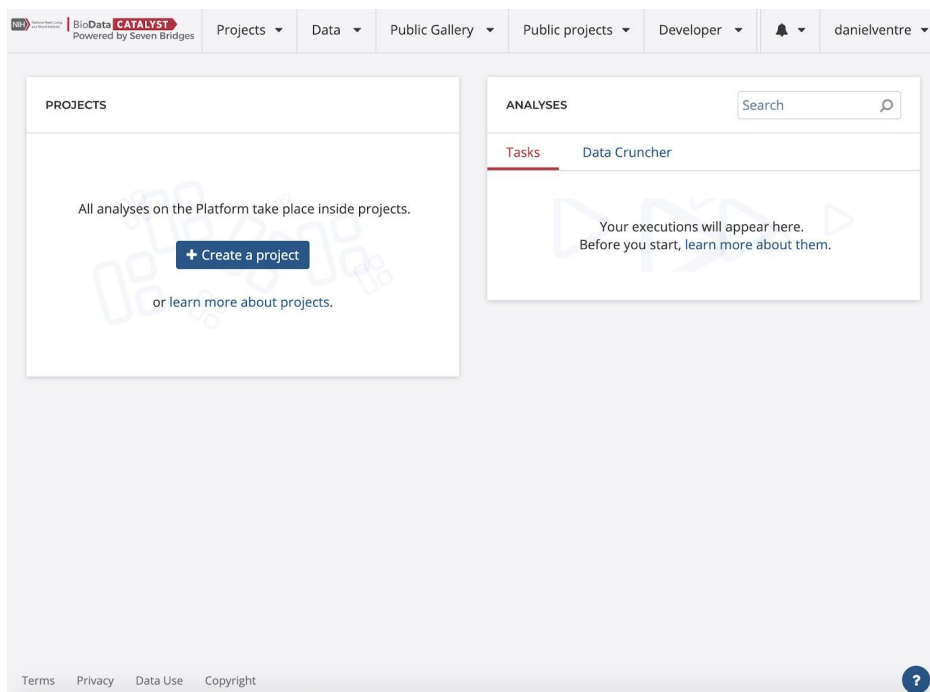
**Note:** If you do not have access to any data you will be redirected back to the login page. Please go to the [BioData Catalyst Data Access webpage](#) for help.

# Running BB-EIGHT

## Installing the API

Getting BB-EIGHT running on the NHLBI BioData Catalyst (BDC) platform should take you no more than a few minutes. The easiest way to get BB-EIGHT running on BDC is to create a controlled-access project and interactive analysis on the Seven Bridges platform.

1. First ensure that you have access to the BDC Platform powered by Seven Bridges (see Authorization and Access above). Create a controlled access project by clicking “+ Create a project” in the dashboard after logging in; this project will serve as your workspace for using the BB-EIGHT API and allow you to securely access data from the TOPMed datasets.



2. Click on interactive analysis and then “Data Cruncher” followed by “Create your first analysis.” Choose an Analysis name and select “JupyterLab (Web-based UI for Project Jupyter)” and the default environment, SB Data Science - Python 3.8, R 3.6; choose your Instance type (c4.2xlarge with 8vCPUs and 15GB RAM works well for most analyses).

The screenshot shows the 'Create new analysis' form with the 'Basic Information' tab selected. The form includes a section for 'Analysis name' with a text input field containing 'GRAPEFRUIT'. Below this is the 'Environment' section, which offers two options: 'JupyterLab' (described as 'Web-based UI for Project Jupyter') and 'RStudio BETA' (described as 'IDE for R'). The 'JupyterLab' option is currently selected. Under 'Environment setup', there is a dropdown menu showing 'SB Data Science - Python 3.8, R 3.6'. At the bottom right of the form are 'Previous' and 'Next' buttons.

3. Follow the default settings and initialize your new virtual environment. This step may take a few minutes.

**INITIALIZING** **2/3** **BB-EIGHT**

4. Once your analysis environment is ready, click “Open in editor” and then launch a new terminal process in JupyterLab. In the terminal, type the following command to retrieve the BB-EIGHT codebase:

```
$ git clone https://github.com/manrai/G-test.git
```

5. BB-EIGHT has the following requirements:

Python library requirements	<ul style="list-style-type: none"><li>• flask</li><li>• numpy</li><li>• pandas</li><li>• sqlite3</li><li>• json</li><li>• subprocess</li></ul>
R library requirements	<ul style="list-style-type: none"><li>• tidyverse</li><li>• RSQLite</li><li>• ggthemes</li><li>• optparse</li><li>• stringi</li></ul>
System requirements	<ul style="list-style-type: none"><li>• bcftools</li></ul>

Most of these libraries should be installed by default in your BDC environment, with the likely exceptions of flask (Python), ggthemes (R), optparse (R), and bcftools (system). These can be installed easily:

```
$ pip3 install flask

$ R
> install.packages("ggthemes")
> install.packages("optparse")

$ sudo apt-get update -y
$ sudo apt-get install -y bcftools
```

## Launching BB-EIGHT

6. Launch the API:

```
$ cd penetrance-api  
$ python3 bb-eight.py
```

7. Test the API:

```
$ curl http://127.0.0.1:5000/variants/random  
$ curl http://127.0.0.1:5000/variants/rsid/rs45548631
```

## How BB-EIGHT Works

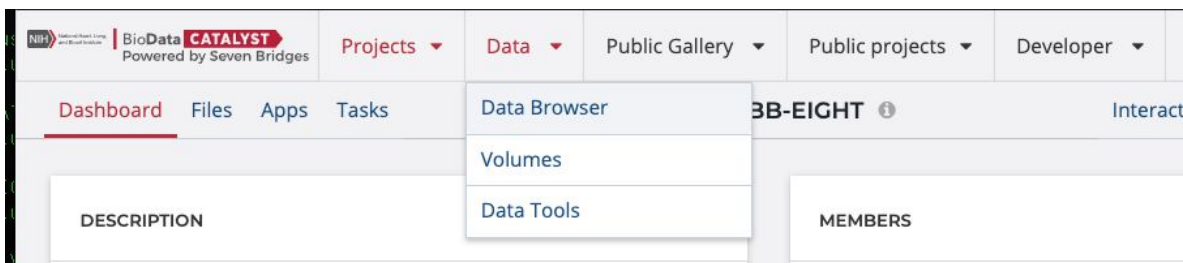
[schematic figure]

# Adding BDC Genomic Data

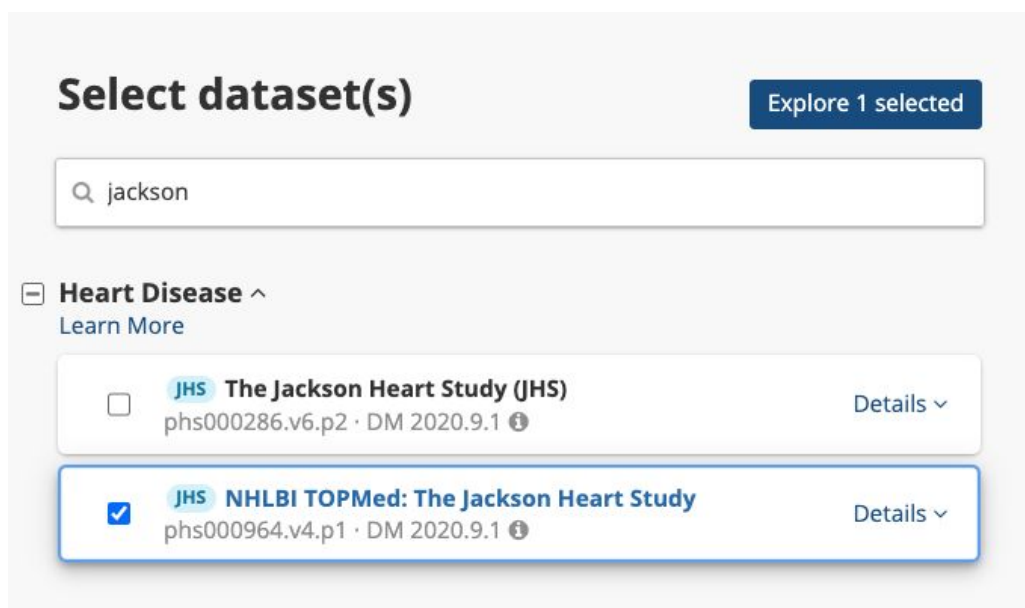
BB-EIGHT is more powerful after adding multiple cohorts of genomic and clinical data from the BDC to your project. This section walks you through importing genetic data from The Jackson Heart Study (JHS). The same steps can be taken to import data from other cohorts. The first step is to verify Authorization and Access (above).

## Importing the Data

1. Navigate to Data > Data Browser in the header navigation bar.



2. Query “Jackson” in the search bar and select the check box for “JHS NHLBI TOPMed: The Jackson Heart Study” (phs000964.v4.p1). Then click on “Explore 1 selected.”



3. Click File > Search for all “VCF” Data Format files which yields 3166 files. Click “Copy files to project” and select the controlled project “BB-EIGHT.”

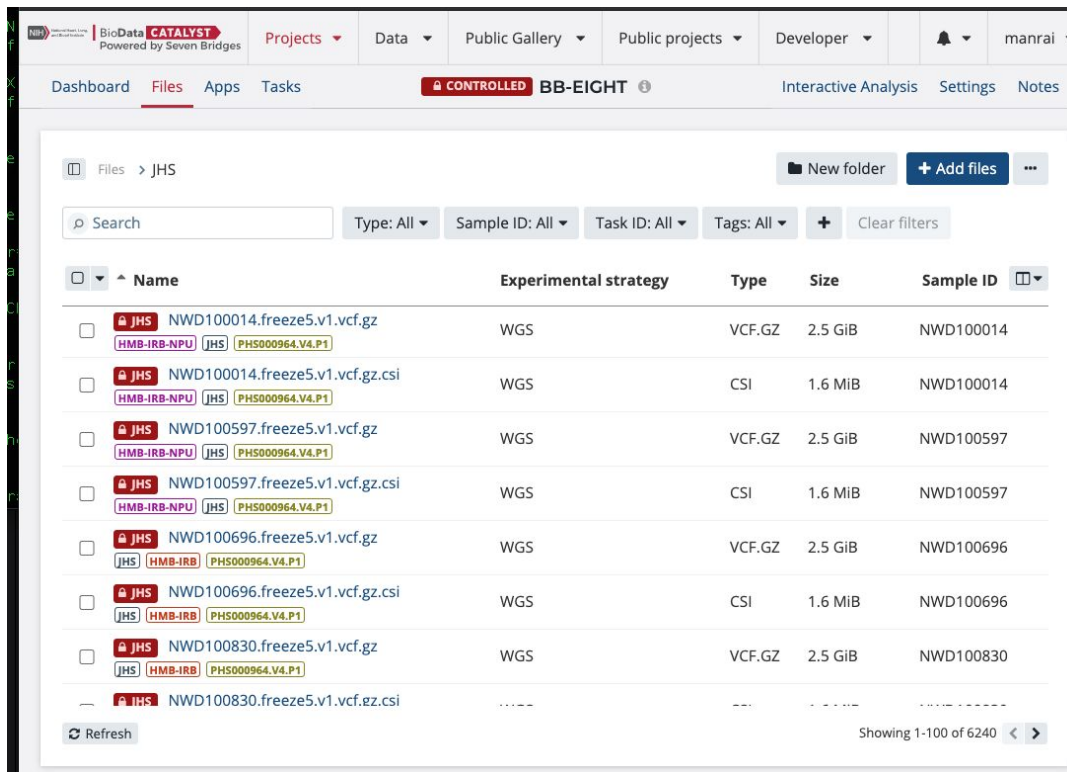
The screenshot shows the BioData CATALYST interface. At the top, there's a navigation bar with tabs for Projects, Data, Public Gallery, Public projects, and Developer. Below this, a search bar shows 'New query' and 'Copy files to project'. A modal window titled 'File' is open, showing 'Data Format VCF'. Below the modal, a filter bar shows 'File' with a count of '3,166'. The main content area is divided into three sections: 'File' (a list of VCF files), 'Details for NWD159406.freeze5.v1.vcf.gz' (showing metadata like Access level, Assay Type, Assembly Name, Consent, Coverage, Data Format, and Data Type), and 'Connections' (showing inbound and outbound connections). The footer contains links for Privacy Policy, Data Sharing Policy, Freedom of Information Act (FOIA), Accessibility, and U.S. Department of Health & Human Services.

4. Click “Copy selected files” in the confirmation window (this will also copy over index files); add a helpful tag like “JHS”.

The screenshot shows a 'Copy' confirmation window. It has a title bar with 'Copy' and a close button. Inside, there's a blue information box that says 'Index files will also be imported.' Below this, it says 'You are about to copy 3166 file(s) to your project BB-EIGHT'. There's a section titled 'Add tags' with a text input field containing the placeholder 'Add multiple tags by separating them by a comma, enter or tab k'. At the bottom, there are two buttons: 'Cancel' and 'Copy selected files'.



- Verify that you can see the newly added files by navigating to “Files” for the Project BB-EIGHT. Create a folder called “JHS” and move all the JHS files to this folder.



## Using the Data with BB-EIGHT

- Test that you can access the files by opening your interactive analysis, starting a new terminal, and typing the following commands to see the first 100 files in the JHS directory and then running bcftools to look at one of the VCF files in JHS at chr11 and position 47332517 (you should see a single output row for this locus):

```
$ ls -lU --block-size=M | head -100
$ bcftools view -r chr11:47332517
../project-files/JHS/NWD100014.freeze5.v1.vcf.gz | grep "^[^#;]"
```

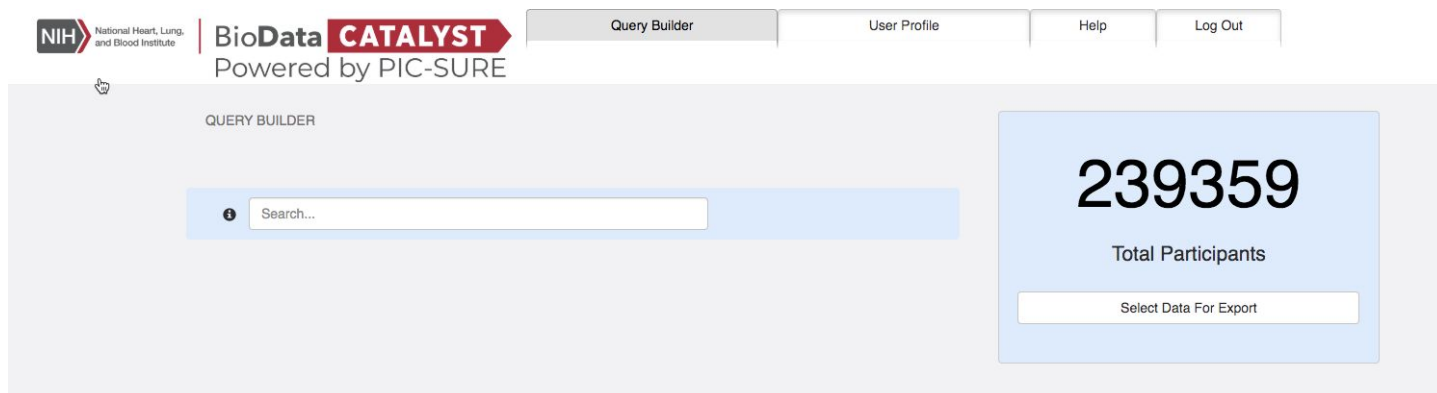
- Repeat this process with HVH and CARDIA. You can then run a script from the GitHub repo to extract all relevant HCM variants needed for BB-EIGHT, and then create a SQLite3 database from the extracted variants:

```
$ cd /bb-eight/bdc-genetic-data
$ bash extract_bdc_variants.sh
$ python create_variant_db.py
```

# Using PIC-SURE for Clinical Data

## Using PIC-SURE

You can use BDC PIC-SURE to obtain clinical data for use with BB-EIGHT. You can use the PIC-SURE User Interface [<https://picsure.biodatacatalyst.nihbi.nih.gov/>] to browse available data. After authorization, you will see the total number of participants available based on the data you can access. Additionally, you can also navigate to the list of studies you are authorized to access.



For a detailed tutorial of using PIC-SURE on BDC, see the [PIC-SURE User Guide](#).

## Using PIC-SURE with BB-EIGHT

In addition to the user interface and manual data download, BDC PIC SURE includes programmatic ways to access the BDC clinical data so that they can be used alongside genomic data. This is the recommended way to use PIC-SURE and clinical data with BB-EIGHT. To access the relevant phenotypes for HCM for BB-EIGHT, you can issue the following commands in a terminal which leverages the PIC-SURE High-Performance Data Store (HPDS, [learn more](#)).

```
$ cd bdc-pic-sure
$ Rscript extract_bdc_clinical.R
```

You can then create a database of clinical variables from the extracted data:

```
$ python create_clinical_db.py
```

# BB-EIGHT API Endpoints

BB-EIGHT supports a number of endpoints to query the data and understand penetrance across cohorts.

## GET /home

Sample Address: <http://127.0.0.1:5000/>

Sample Output:



## Welcome to the NHLBI BioData Catalyst BB-EIGHT API

This API allows investigators to specify phenotype and genotype criteria dynamically to obtain penetrance estimates for genetic variants across populations.

For more information please visit: [BB-EIGHT codebase](#)

## GET /variants/rsid

Sample Address: <http://127.0.0.1:5000/variants/rsid/rs45548631>

Sample Output:

```
[
  - {
    Alternate: "T",
    Chromosome: 14,
    Frequency: 0.0046902523398204546,
    Position: 23882043,
    Reference: "C",
    rsID: "rs45548631"
  }
]
```

## GET /variants/position

Sample Address: <http://127.0.0.1:5000/variants/position?chr=14&pos=23882043&ref=C&alt=T>

Sample Output:

```
[
  - {
    Alternate: "T",
    Chromosome: 14,
    Frequency: 0.0046902523398204546,
    Position: 23882043,
    Reference: "C",
    rsID: "rs45548631"
  }
]
```

## GET /variants/random

Sample Address: <http://127.0.0.1:5000/variants/random>

Sample Output:

```
{
  Alternate: "C",
  Chromosome: 1,
  Frequency: 0.00009901340217122247,
  Position: 201330507,
  Reference: "CAAG",
  rsID: "rs397516480"
}
```

## GET /variants/all

Sample Address: <http://127.0.0.1:5000/variants/all>

Sample Output:

```
-
[ ...

  • -
    { ...
      ◦ Alternate: "G",
      ◦ Chromosome: 14,
      ◦ Frequency: 0.000031849162367029746,
      ◦ Position: 23881997,
      ◦ Reference: "T",
      ◦ rsID: "rs1431875543"
    },
  • -
    { ...
      ◦ Alternate: "T",
      ◦ Chromosome: 14,
      ◦ Frequency: 0.00003184307731499172,
      ◦ Position: 23882001,
      ◦ Reference: "C",
      ◦ rsID: "rs1308662448"
    },
  • -
    { ...
      ◦ Alternate: "G",
      ◦ Chromosome: 14,
      ◦ Frequency: 0.000021241795356543534,
      ◦ Position: 23882012,
      ◦ Reference: "A",
      ◦ rsID: "rs780861108"
    },
  • -
    { ...
      ◦ Alternate: "A",
      ◦ Chromosome: 14,
      ◦ Frequency: 0.00004382190776683558,
      ◦ Position: 23882014,
      ◦ Reference: "G",
      ◦ rsID: "rs747614764"
    },
  • -
    { ...
      ◦ Alternate: "A",
      ◦ Chromosome: 14,
      ◦ Frequency: 0.000003983778055756957,
      ◦ Position: 23882015,
```

## GET /clinvar/position

Sample Address: <http://127.0.0.1:5000/clinvar/position?chr=14&pos=23882043>

Sample Output:

```
[
  - {
    Alternate: "T",
    Chromosome: 14,
    Pathogenic: 0,
    Position: 23882043,
    Reference: "C"
  }
]
```

## POST /penetrance

Sample Address: <http://127.0.0.1:5000/penetrance>

POST input:

```
1  {
2    "vus": 0,
3    "likely_pathogenic": 1,
4    "pathogenic": 1,
5    "lwt_min": 8,
6    "lwt_max": 13,
7    "age_min": 15,
8    "age_max": 40,
9    "gender": "all",
10   "race": "all",
11   "cohorts": ["jackson", "cardia", "framingham"],
12   "htn": 1
13 }
```

← which variant types to include

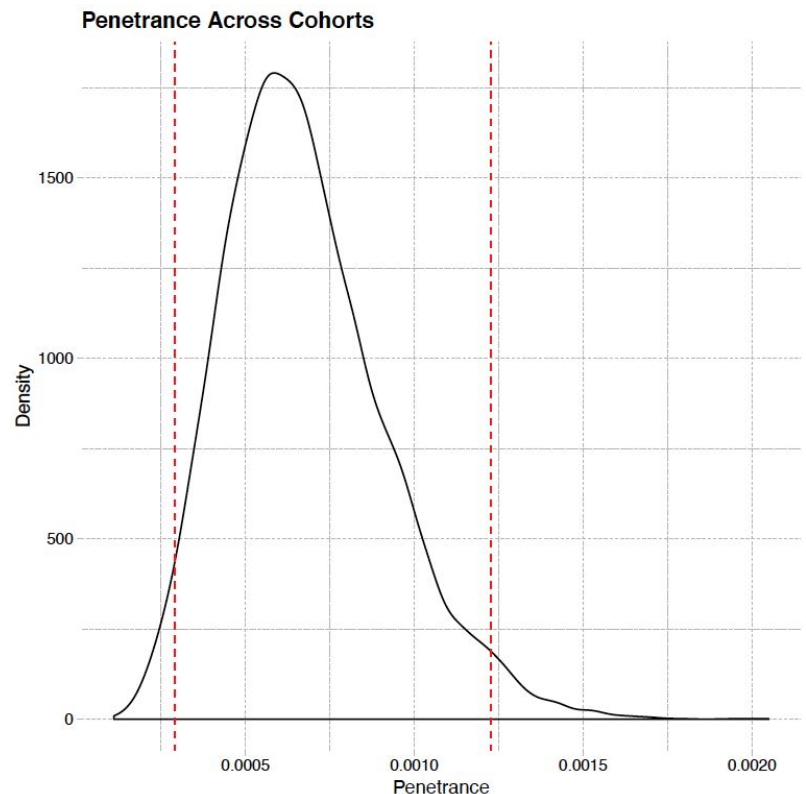
← demographic and phenotype criteria

Sample Output:

```
1  {
2    "Low": 0.0003,
3    "MAP": 0.0006,
4    "High": 0.0012
5  }
```

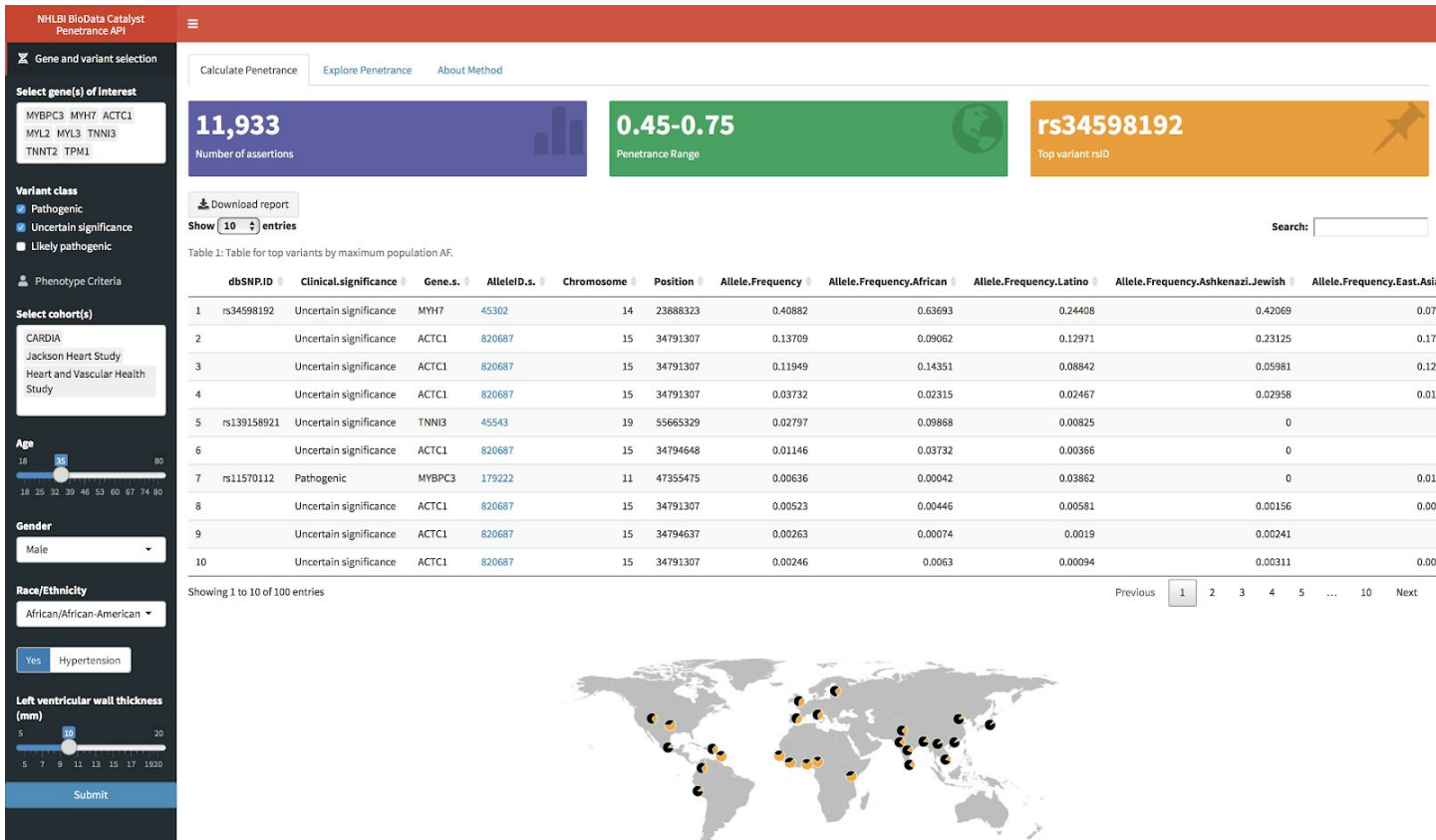
**Top:** JSON output of MAP, low, and high credible interval estimates for penetrance.

**Right:** Posterior distribution of penetrance saved to user's directory with filename corresponding to user parameters.



# BB-EIGHT App

BB-EIGHT is available as an R/Shiny app in the /app directory. After running:



This screenshot of the NHLBI BioData Catalyst BB-EIGHT app shows results for 11,933 variant assertions retrieved from ClinVar for the genes *MYBPC3*, *MYH7*, *ACTC1*, *MYL2*, *MYL3*, *TNNI3*, *TNNT2*, and *TPM1* for variant classes P/VUS.