



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

## BB-EIGHT Overview

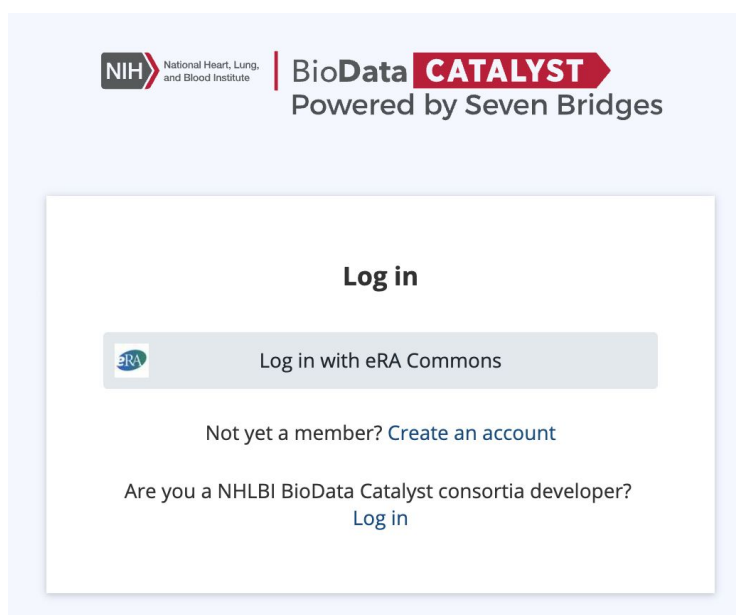
As part of the [BioData Catalyst](#) project, researchers at Harvard Medical School and Boston Children's Hospital have developed a computational tool called **BB-EIGHT**: **B**eta-**B**inomial **E**stimation **I**nterface for **G**enetic Risk across **H**uman **T**raits. BB-EIGHT addresses a ubiquitous 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 specified 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

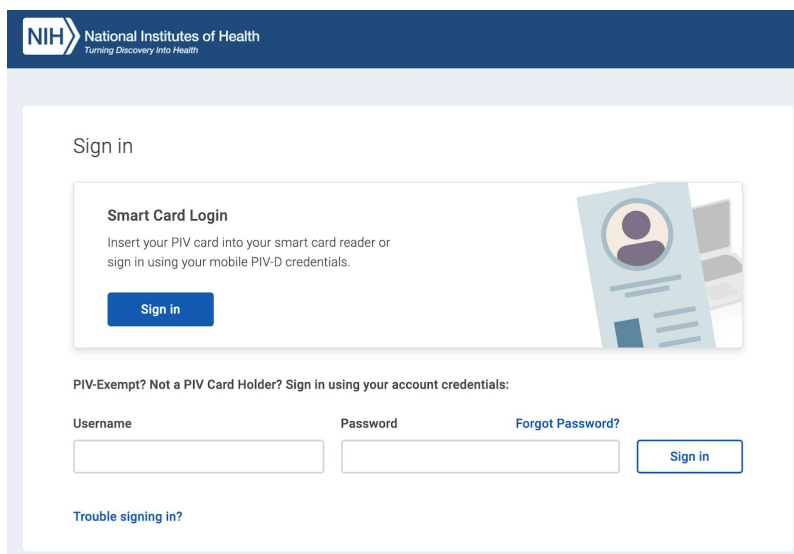
To use BB-EIGHT on BioData Catalyst:

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.



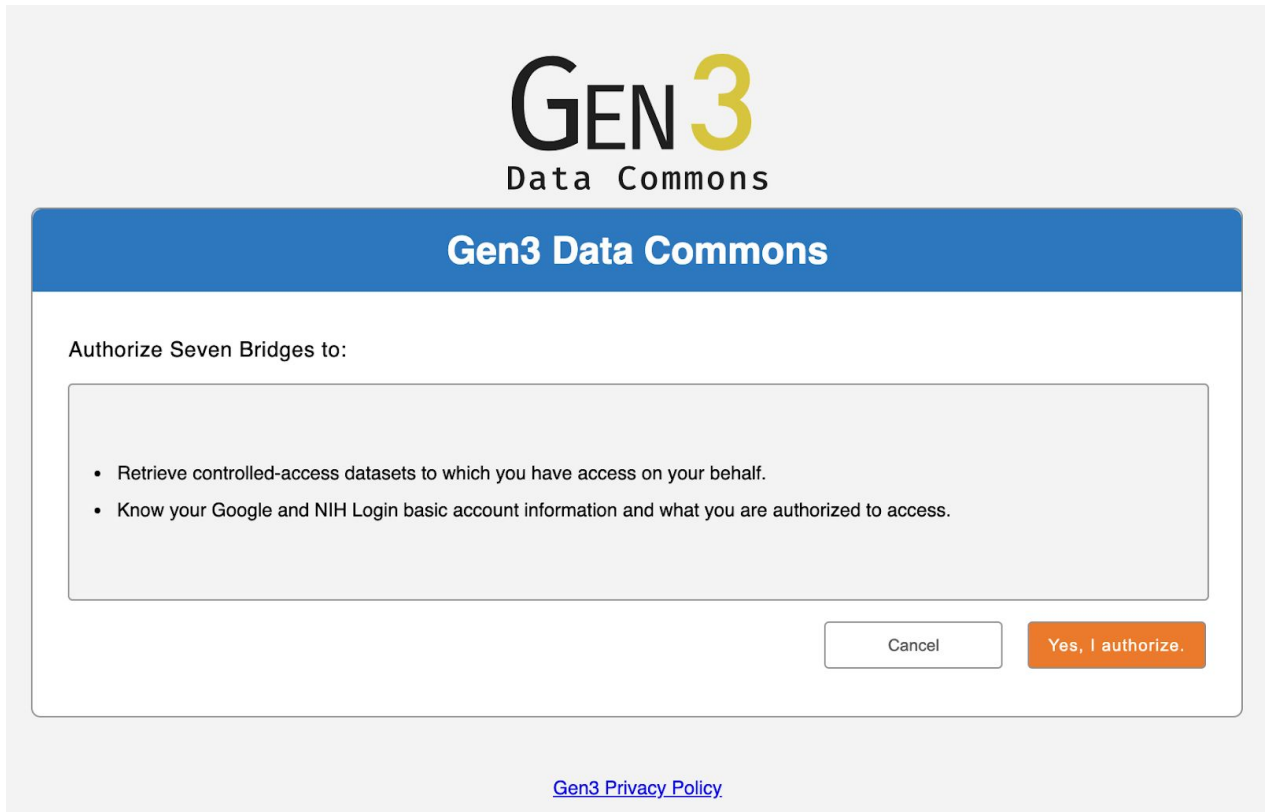
The image shows the BioData CATALYST login page. At the top, there is a header with the NIH logo (National Heart, Lung, and Blood Institute) and the BioData CATALYST logo, which includes the text "Powered by Seven Bridges". Below the header, the main heading is "Log in". Underneath, there is a button labeled "Log in with eRA Commons" with a small eRA Commons icon to its left. Below this button, there is a link that says "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 heading is "Sign in". Underneath, there is a section titled "Smart Card Login" 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 that says "PIV-Exempt? Not a PIV Card Holder? Sign in using your account credentials:". Underneath this link, there are two input fields: "Username" and "Password". To the right of the Password field is a link that says "Forgot Password?". Below the input fields is a "Sign in" button. At the bottom, there is a link that says "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 interface for Gen3 Data Commons. At the top, the logo "GEN3 Data Commons" is displayed, with "GEN3" in black and "Data Commons" in a smaller black font. Below the logo 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 main content area are two buttons: a white "Cancel" button and an orange "Yes, I authorize." button. At the bottom center of the page is a blue link labeled "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](#)

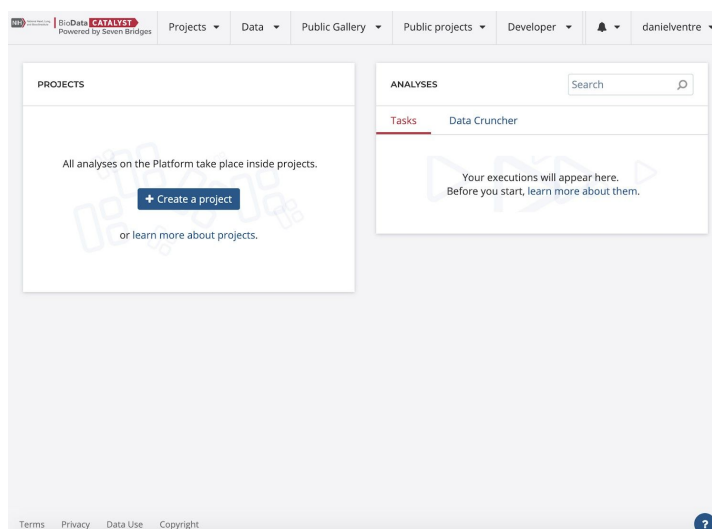
**Note:** If you do not have access to any data you will be redirected back to the login page. Please go to the [BioData Catalyst Data Access webpage](#) for help.

# 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

The image shows a 'Create new analysis' modal window. It has a close button (X) in the top right corner. There are two tabs: 'Basic information' (active) and 'Compute requirements'. Under the 'Basic information' tab, there is a text input field for 'Analysis name' containing the text 'GRAPEFRUIT'. Below this is a section for 'Environment' with two selectable options: 'JupyterLab' (selected) and 'RStudio BETA'. The 'JupyterLab' option is described as 'Web-based UI for Project Jupyter'. Below the environment selection is a section for 'Environment setup' with a dropdown menu showing 'SB Data Science - Python 3.8, R 3.6'. At the bottom right, there are 'Previous' and 'Next' buttons.

8vCPUs and 15GB RAM works well for most analyses).

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



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/bb-eight.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><li>• pic-sure-r-client</li><li>• pic-sure-r-adapter-hpds</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), bcftools (system), and the two pic-sure libraries (R). To install pic-sure for R, see: <https://github.com/hms-dbmi/pic-sure-r-client>. The remaining requirements 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

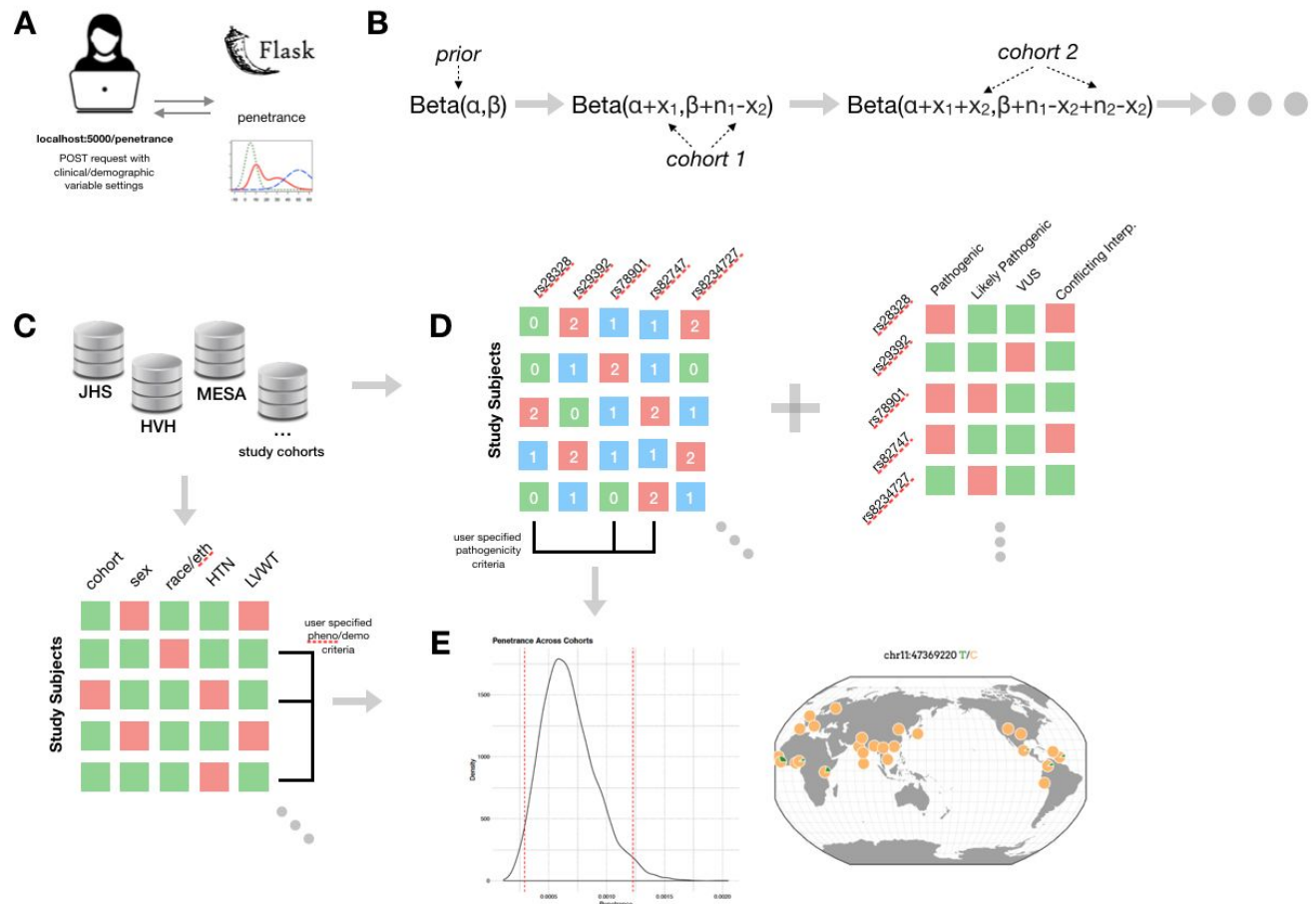
6. Launch the API:

```
$ cd bb-eight/src/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"
$ curl
"http://127.0.0.1:5000/variants/position?chr=14&pos=23882043&ref=C&alt=T"
$ curl --location --request POST 'http://127.0.0.1:5000/penetrance' \
--header 'Content-Type: application/json' \
--data-raw '{
    "vus": 1,
    "likely_pathogenic": 1,
    "pathogenic": 1,
    "lvwt_min": 8,
    "lvwt_max": 13,
    "age_min": 15,
    "age_max": 40,
    "gender": "all",
    "race": "all",
    "cohorts": ["jackson", "mesa", "hvh"],
    "htn": 1
}'
```

# How BB-EIGHT Works



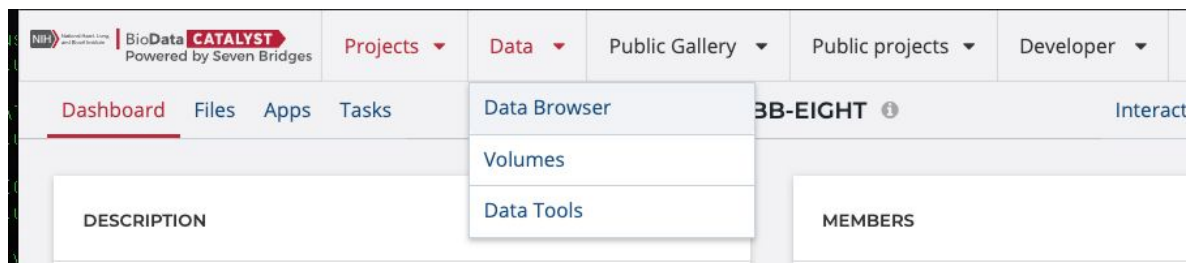
**The BB-EIGHT framework for computing penetrance.** **A** The user passes real-time queries to the BB-EIGHT Flask API, which coordinates access to multi-cohort genotypic and phenotypic data to compute penetrance subject to user-specified phenotypic, demographic, and variant criteria. **B** BB-EIGHT employs a beta-binomial approach to compute posterior distributions for each of three terms underlying penetrance and then samples from these distributions to compute overall penetrance. This approach generalizes naturally to multiple study cohorts. **C** Application of BB-EIGHT to three cardiovascular disease cohorts to study the penetrance of hypertrophic cardiomyopathy (HCM). Phenotype criteria, including age, sex, race/ethnicity, hypertension (HTN), and left ventricular wall thickness (LVWT) are extracted from the Jackson Heart Study (JHS), the Heart and Vascular Health (HVH) Cohort, and the Multi-Ethnic Study of Atherosclerosis (MESA). **D** Genomic data from the study subjects for all disease-associated variants in the eight HCM genes are extracted alongside ClinVar pathogenicity assessments. **E** Using the user's criteria, the posterior distribution for penetrance is computed and visualized along with common variants across ancestry groups.

# Adding BDC Genomic Data

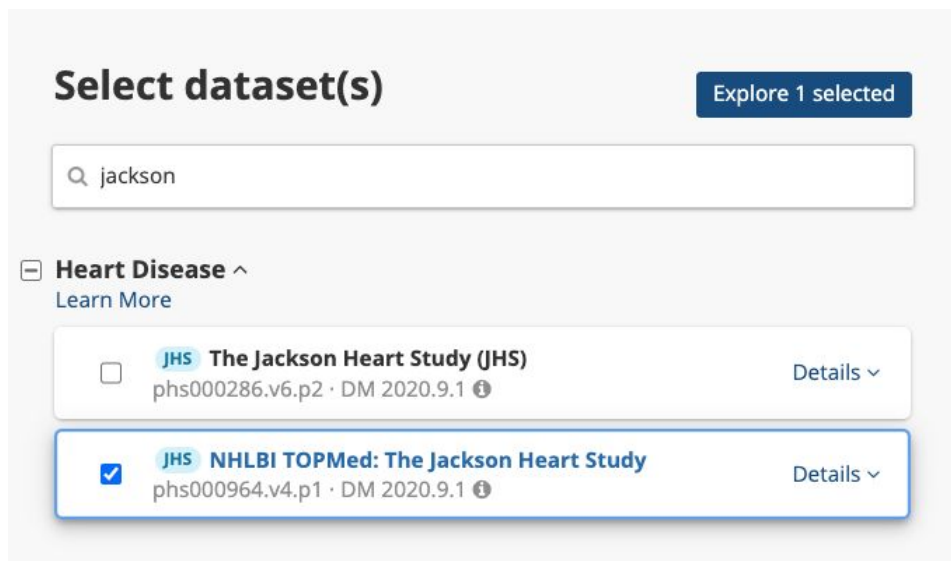
BB-EIGHT is more powerful after adding multiple cohorts of genomic and clinical data from the BDC to your project. This section walks you through importing genomic data from The Jackson Heart Study (JHS). The same steps can be taken to import data from other cohorts. The first step is to verify Authorization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





- 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 NIH BioData CATALYST interface. The top navigation bar includes links for Projects, Data, Public Gallery, Public projects, and Developer. A search bar is present with the text "Search by ID". A button labeled "Copy files to project" is visible. Below the navigation bar, a search results summary shows "File" with "3,166" results. A dropdown menu is open, showing "File", "Data Format", and "VCF". Below the summary, a table lists the details for the selected files. The table has three columns: File, Details for NWD159406.freeze5.v1.vcf.gz, and Connections. The File column lists several .vcf.gz files. The Details column shows information such as Access level, Assay Type, Assembly Name, Consent, Coverage, Data Format, and Data Type. The Connections column shows inbound and outbound connections.

File	Details for NWD159406.freeze5.v1.vcf.gz	Connections
NWD159406.freeze5.v1.vcf.gz	JHS	Inbound:
NWD163918.freeze5.v1.vcf.gz	Access level ⓘ	No inbound connections
NWD781467.freeze5.v1.vcf.gz	Assay Type ⓘ	WGS
NWD421932.freeze5.v1.vcf.gz	Assembly Name ⓘ	GRCh
NWD816276.freeze5.v1.vcf.gz	Consent ⓘ	DS- FDO- IRB
NWD348255.freeze5.v1.vcf.gz	Coverage	No outbound connections
NWD654825.freeze5.v1.vcf.gz	Data Format ⓘ	34.6
	Data Type ⓘ	VCF
		Simp

- 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. At the top, it says "Copy" with a close button. Below this, a message states "Index files will also be imported." followed by "You are about to copy 3166 file(s) to your project BB-EIGHT". Under the heading "Add tags", there is a text input field with the placeholder text "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.

The screenshot shows the BB-EIGHT web interface. The top navigation bar includes links for Projects, Data, Public Gallery, Public projects, Developer, and a user profile 'manrai'. The main header shows 'Dashboard', 'Files' (selected), 'Apps', and 'Tasks'. Below this, a red 'CONTROLLED' badge and 'BB-EIGHT' are visible. The 'Files' section shows a breadcrumb 'Files > JHS' and buttons for 'New folder', 'Add files', and a search bar. A table of files is displayed with columns: Name, Experimental strategy, Type, Size, and Sample ID. The files listed are:

Name	Experimental strategy	Type	Size	Sample ID
NWD100014.freeze5.v1.vcf.gz	WGS	VCF.GZ	2.5 GiB	NWD100014
NWD100014.freeze5.v1.vcf.gz.csi	WGS	CSI	1.6 MiB	NWD100014
NWD100597.freeze5.v1.vcf.gz	WGS	VCF.GZ	2.5 GiB	NWD100597
NWD100597.freeze5.v1.vcf.gz.csi	WGS	CSI	1.6 MiB	NWD100597
NWD100696.freeze5.v1.vcf.gz	WGS	VCF.GZ	2.5 GiB	NWD100696
NWD100696.freeze5.v1.vcf.gz.csi	WGS	CSI	1.6 MiB	NWD100696
NWD100830.freeze5.v1.vcf.gz	WGS	VCF.GZ	2.5 GiB	NWD100830
NWD100830.freeze5.v1.vcf.gz.csi	WGS	CSI	1.6 MiB	NWD100830

At the bottom, there is a 'Refresh' button and a pagination indicator 'Showing 1-100 of 6240'.

## Using the Data with BB-EIGHT

- Test that you can access the newly added genomic data 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 "^#[#;]"
```

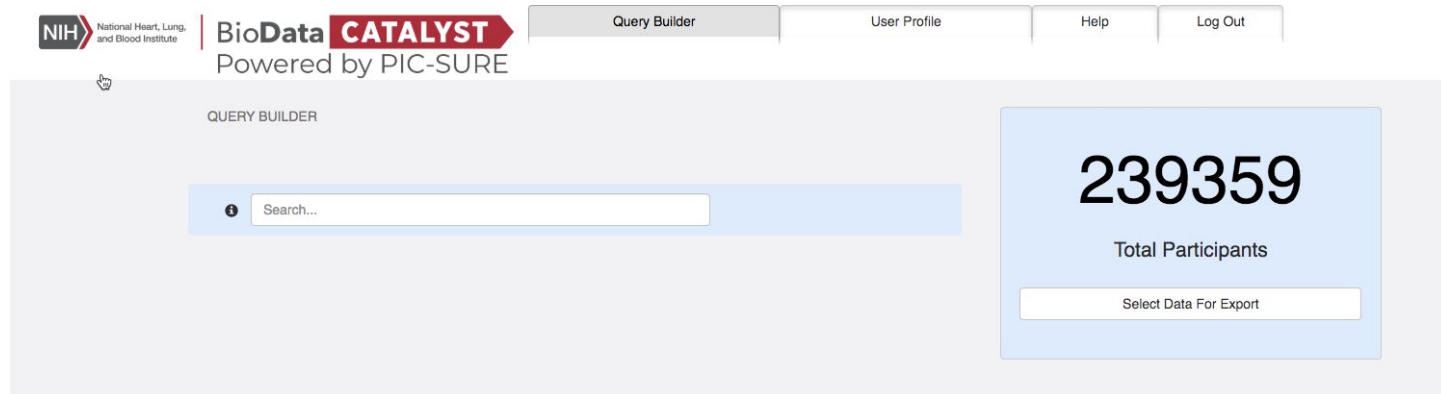
7. Repeat this process with HVH and MESA. You can then run a script from the GitHub repo to extract all relevant HCM variants needed for BB-EIGHT (warning: this process is slow if not parallelized):

```
$ cd /bb-eight/bdc-genetic-data  
$ Rscript extract_bdc_variants_<cohort_name>.R # <cohort_name> = JHS,  
MESA..etc.
```

# Using PIC-SURE for Clinical Data

## Using PIC-SURE

You can use BDC PIC-SURE to obtain clinical data for use with BB-EIGHT alongside genomic data. You can use the PIC-SURE User Interface [<https://picsure.biodatacatalyst.nhlbi.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 and identify variables of interest.



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-db  
$ Rscript build_bdc_db.R
```

This will create a database that can be used with the BB-EIGHT penetrance API.

# BB-EIGHT API Endpoints

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

## GET /home

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

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 body:

```
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

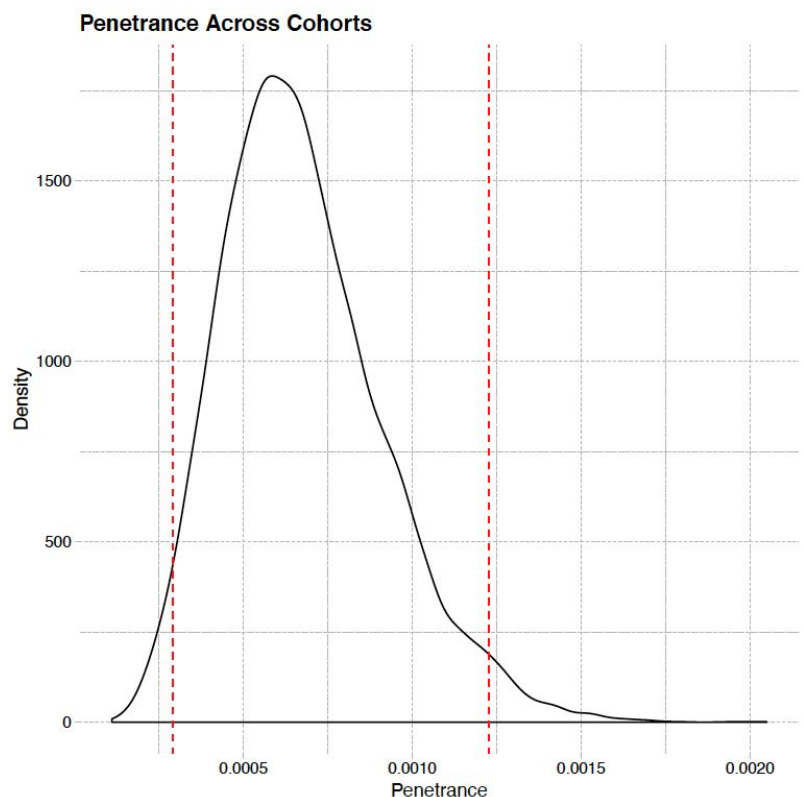
← demographic and phenotype criteria

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:

```
$ cd bb-eight/src/app
$ R
> runApp()
```

NHLBI BioData Catalyst Penetrance API

Gene and variant selection

Select gene(s) of interest

MYBPC3 MYH7 ACTC1  
MYL2 MYL3 TNNI3  
TNNT2 TPM1

Variant class

☒ Pathogenic  
☒ Uncertain significance  
☐ Likely pathogenic

Phenotype Criteria

Select cohort(s)

MESA Jackson Heart Study  
Heart and Vascular Health Study

Age

0 8 16 24 32 40 48 56 64 72 80

Gender

Male

Race/Ethnicity

American Indian, Alaskan Native, or Native American

Yes Hypertension

Left ventricular wall thickness (mm)

0 2 4 6 8 10 12 14 16 18 20

Submit

Calculate Penetrance Explore Penetrance About Method

11,933  
Number of assertions

0.43-0.73  
Penetrance Range

rs34598192  
Top variant rsID

Download Report

Show 10 entries

	rsID	Class	Gene	ClinVarRecord	Chr	Position	AF	Afr	Lat	AJ	EA	EF	ENF	SA	MostFrequent
1	rs34598192	Uncertain significance	MYH7	45302	14	23888323	0.40882	0.63693	0.24408	0.42069	0.07014	0.26932	0.36028		African
2		Uncertain significance	ACTC1	820687	15	34791307	0.13709	0.09062	0.12971	0.23125	0.17442	0.11259	0.16546	0.08649	Ashkenazi Jewish
3		Uncertain significance	ACTC1	820687	15	34791307	0.11949	0.14351	0.08842	0.05981	0.12124	0.11043	0.11691	0.13243	African
4		Uncertain significance	ACTC1	820687	15	34791307	0.03732	0.02315	0.02467	0.02958	0.01332	0.05219	0.04913	0.01848	European.Finnish
5	rs139158921	Uncertain significance	TNNI3	45543	19	55665329	0.02797	0.09668	0.00825	0	0	0	0.0002		African
6		Uncertain significance	ACTC1	820687	15	34794648	0.01146	0.03732	0.00366	0	0	0	0.0002	0	African
7	rs11570112	Pathogenic	MYBPC3	179222	11	47355475	0.00636	0.00042	0.03862	0	0.01629	0	0.00006	0.00049	Latino
8		Uncertain significance	ACTC1	820687	15	34791307	0.00523	0.00446	0.00581	0.00156	0.00341	0.002	0.00533	0.00578	European.Non.Finnish
9		Uncertain significance	ACTC1	820687	15	34794637	0.00263	0.00074	0.0019	0.00241	0	0.00105	0.00458	0.00131	European.Non.Finnish
10		Uncertain significance	ACTC1	820687	15	34791307	0.00246	0.0063	0.00094	0.00311	0.00136	0.00022	0.00092	0.00038	African

Showing 1 to 10 of 11,933 entries

Table for top variants by maximum population allele frequency (AF) for the following ancestries: African/African-American (Afr), Latino/Admixed American (Lat), Ashkenazi Jewish (AJ), East Asian (EA), European Finnish (EF), European Non-Finnish (ENF), and South Asian (SA).

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.