This workspace demonstrates how to find a TOPMed dataset to use in a single variant, mixed-models GWAS from start to finish using multiple platforms in the BioData Catalyst project.

Data disclaimer

Because this workspace uses controlled access data, it has been registered under an Authorization Domain that limits its access to only researchers with the appropriate approvals. When you copy this workspace for your own use, all copies will also be protected under the same Authorization Domain and cannot be shared with external users. **How you use products from this workspace tutorial are held accountable to your own IRB approval and data use restrictions.**

This tutorial is based on using the TOPMed Amish dataset. If you do not have permissions to use this dataset, the base of this tutorial can be adapted for use with another dataset. However, you will need to carefully consider how to update this analysis for the dataset and how this may affect the scientific question you are asking.

A brief outline of this tutorial is as follows:

Part 1: Navigate the BioData Catalyst multi-platform environment Walk through a series of steps to learn how to search and export data from Gen3 and worfklows from Dockstore into a Terra workspace. Each of these cloud-based platforms easily operate with one another for fast and secure research. The workspace we have created here can be cloned for you to walk through exactly as suggested as a tutorial, or you can use the basics you learn here to perform your own analysis.

Part 2: Explore TOPMed data in an interative Jupyter notebook In this Terra workspace, you can find a series of interactive notebooks to explore TOPMed data. The first, **1-GWAS-preliminary-analysis**, will lead you through a series of steps to explore the phenotypic and genotypic data. However, fist this notebook will use a series of functions found in the **terra*data*util** companion notebook. This companion notebook is available to researchers to manipulate graph-structured TOPMed data from Gen3 (which you will learn more about below). You don't need to open or edit this notebook for this tutorial, but it is available for you to use and edit for your own use case. Once the graph structured data is consolidated, the python-based notebook examines genetic relatedness using the HAIL genomic data analysis tool.

Part 3: Perform mixed-model association tests using workflows Next, perform mixed models genetic association tests (run as a series of batch workflows using GCP Compute engine). For details on the four workflows and what they do, scroll down to **Perform mixed model association test workflows**. The workflows are publicly available in Dockstore in this collection.

Mixed models require two steps within the GENESIS package in R:
1) Fitting a null model assuming that each genetic variant has no effect on phenotype and 2) Testing each genetic variant for association with the outcome, using the fitted null model.

Part 4: Interactive GWAS summarization in notebook Finally, we provide an optional notebook **2-GWAS-summarization** for further summarization of the GWAS results. This notebook provides visible code, editable by the user, that aggregates variants into individual loci grouped by lead variant and generates customizable visualizations.

Helpful resources to master this tutorial

- If you have never used Terra before, peruse the Terra knowledge base. You can also physically attend or watch a recording of one of the Broad Institute Introduction to Terra workshops.
- Controlling cloud costs
- Intro to Jupyter notebooks in Terra
- Intro to Hail using a Terra workspace
- GWAS tutorial using open data from the 1000 Genomes Project

Notes on data in this workspace

This workspace shows a user how to conduct a GWAS analysis on TOPMed data hosted by the BioData Catalyst Project. The user will explore phenotypes and their associated genotype files (specifically, cohort-level VCF files).

TOPMed Metadata

Some types of metadata will always be present: GUID, Case ID, Project Name, Number of Samples, Study, Gender, Age at Index, Race, Ethnicity, Number of Aliquots, SNP Array Files, Unaligned Read Files, Aligned Read Files, Germline Variation Files.

Other metadata depend on the analysis plan submitted when applying for TOPMed access. Examples include BMI, Years Smoked, years smoked greater than 89, hypertension, hypertension medications, diastolic blood pressure, systolic blood pressure, etc.

The TOPMed Data Coordinating Center (DCC) is currently harmonizing select phenotypes across TOPMed, which will also be deposited into the TOPMed accessions. The progress of phenotype metadata harmonization

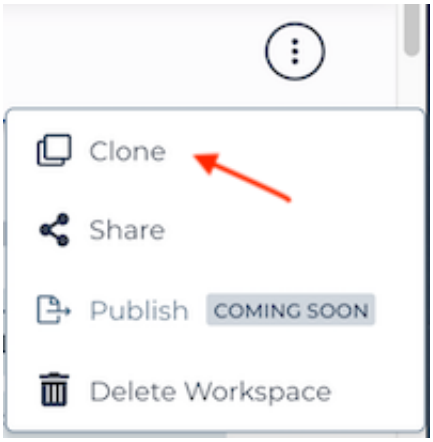
can be tracked here.

Part 1: Navigate the BioData Catalyst multi-platform environment

DataSTAGE encompasses multiple platforms that allow researchers to discover, access, store, and compute large sets of data generated from biomedical and behavioral research. Data is secure in the cloud, where you can scale analyses easily and cost-efficiently.

Use Terra to create your own workspace for computing in the cloud

1. At the top right corner of this workspace. Use your mouse to click the circle with three dots. This will open a window where you can click "clone". Cloning creates a copy of this workspace that you own and can use to work through this tutorial, or use a a template for your own analyses.

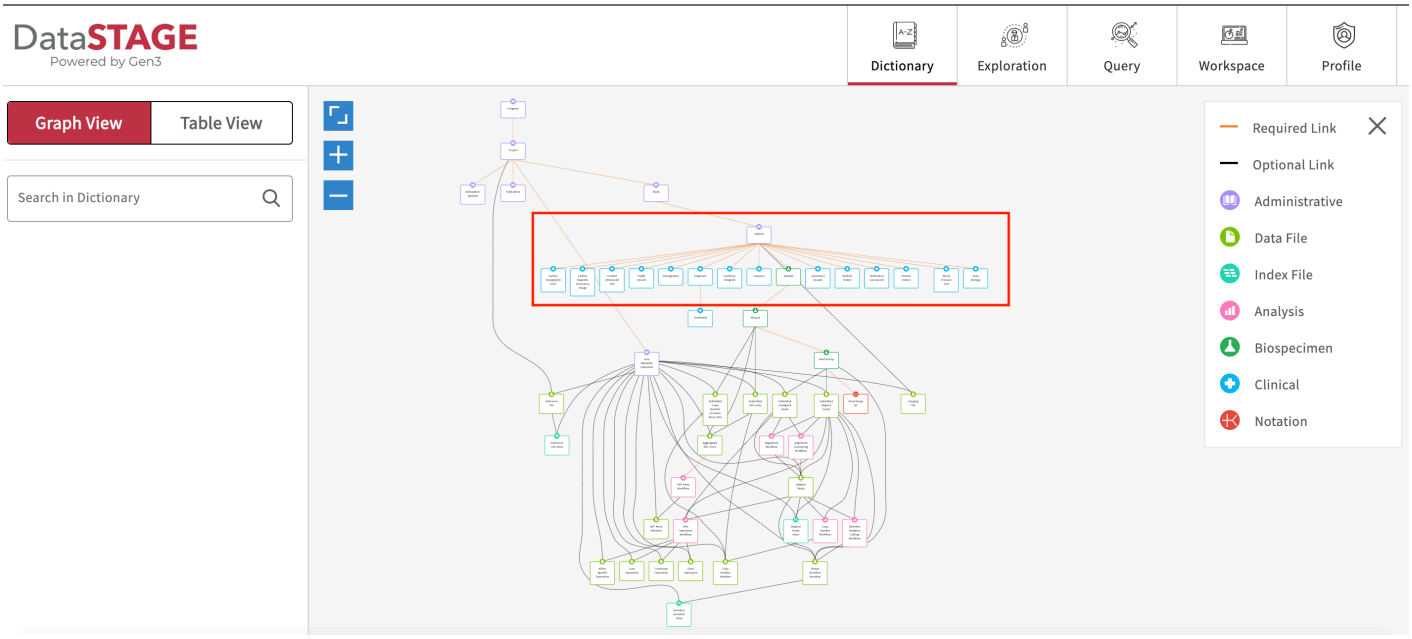


1. After you hit clone, a pop-up window appears asking you to fill out some information.
2. Workspace name: Enter a name a name that is meaningful for your records.
3. Billing Project: Select the billing projects available to you. If you are a new user, you can use the $300 of free credits offered.
4. Authorization Domain: This workspace has an "Inherited Group" that secures that any copy of this workspace continues to be protected to only authorized users.

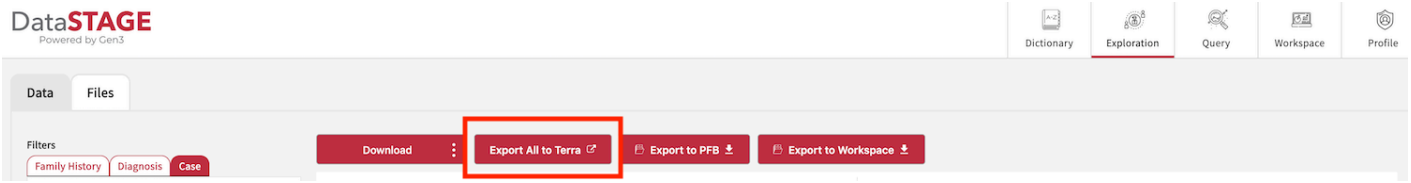
Find TOPMed genotype and harmonized phenotype data in Gen3

1. Log into Gen3 through the NIH portal using your eRA Commons username and password.
2. Navigate to the Gen3 data dictionary and review what metadata files are available and how they are linked to one another in the graph structure.

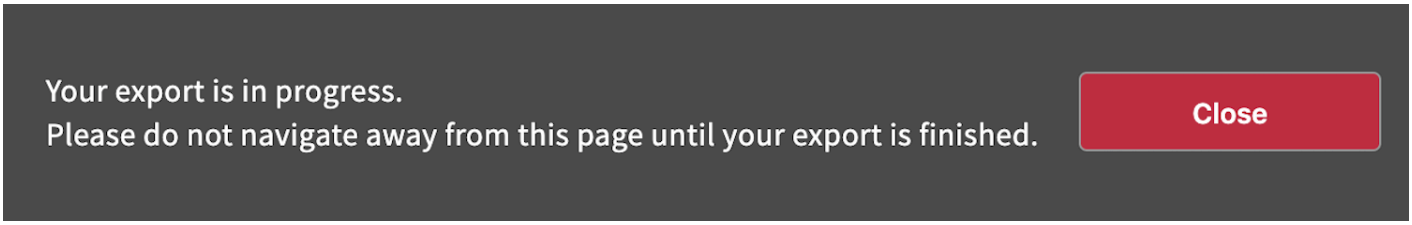
3. On the left hand side of the graphic, click on the "subject" box. The subject refers to a collection of all data related to a specific subject in the context of a specific experiment.



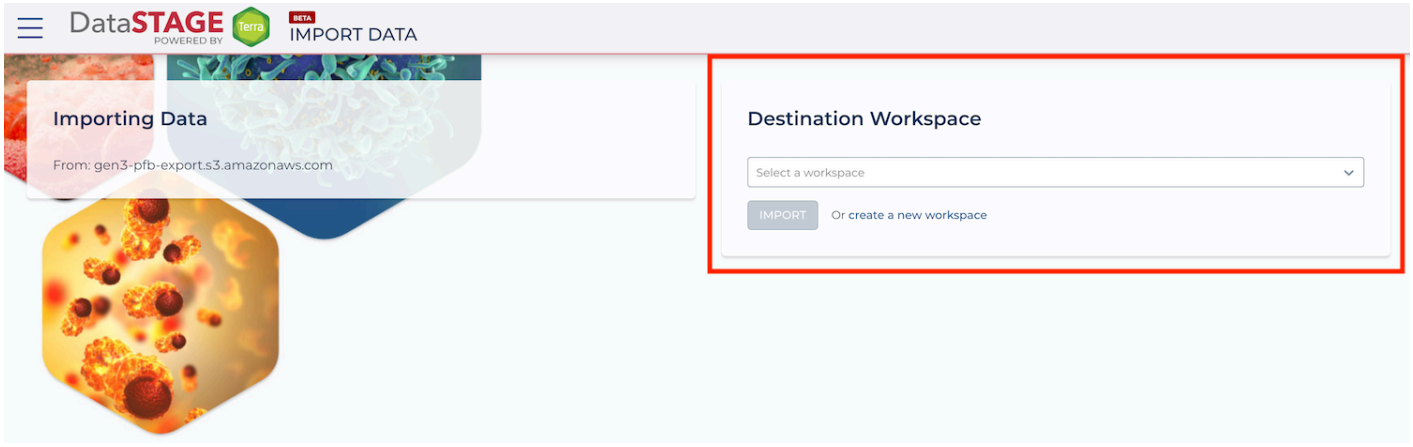
4. Click on one of the blue boxes below "subject" that is linked by a single line (for example, "demographic"). Each of these blue boxes represents clinical data collected for individuals in a study. The example "demographic" refers to the characterization of the patient by means of segementing the population (e.g., characterization by age, sex, or race). You can click through other clinical trait data to see what metadata are available.
5. Navigate to the Gen3 Exploration view to see what datasets you are currently authorized to access. On the left hand side, you can apply filters to narrow your search results.
6. For this tutorial, navigate to the "subject" tab, under "ProjectID" click "show more" and select the "topmed-Amish_HMB-IRB-MDS" project.
7. Once selected, click the red button "Export all to Terra".



8. You should see an "Export in Progress" banner appear. This process generally takes from 30 seconds to several minutes, depending on the size of the files.



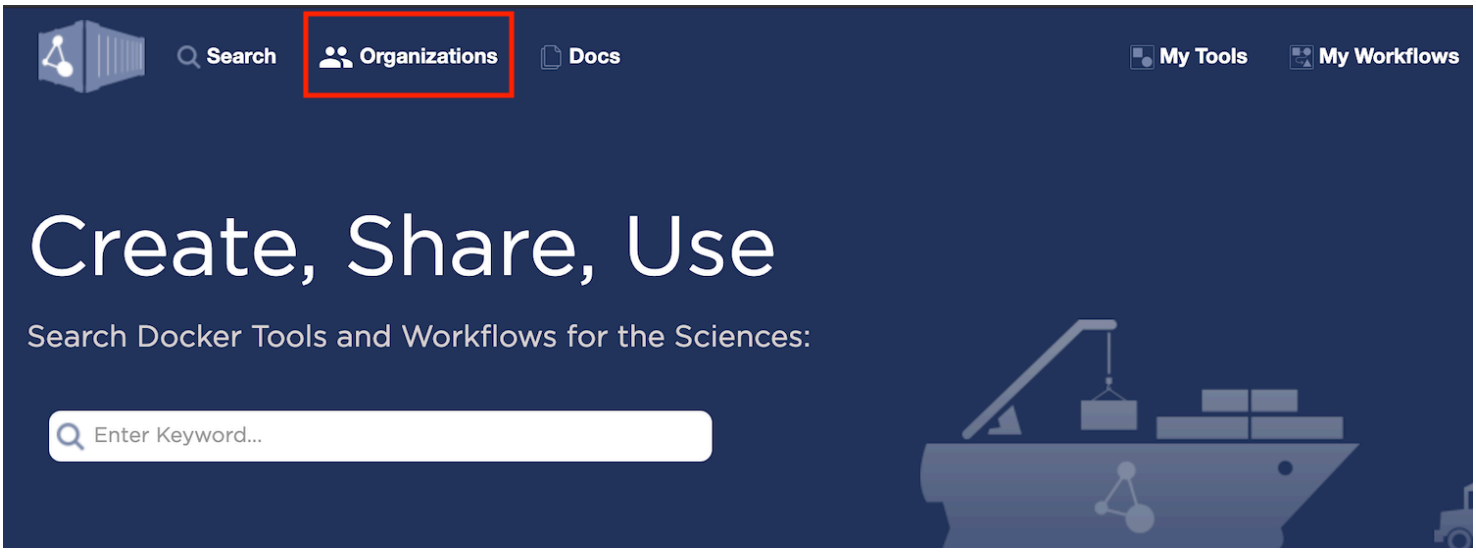
9. When the data export is ready, a new Terra window appears asking you to select a destination workspace. Select the copy of this workspace that you previously made.



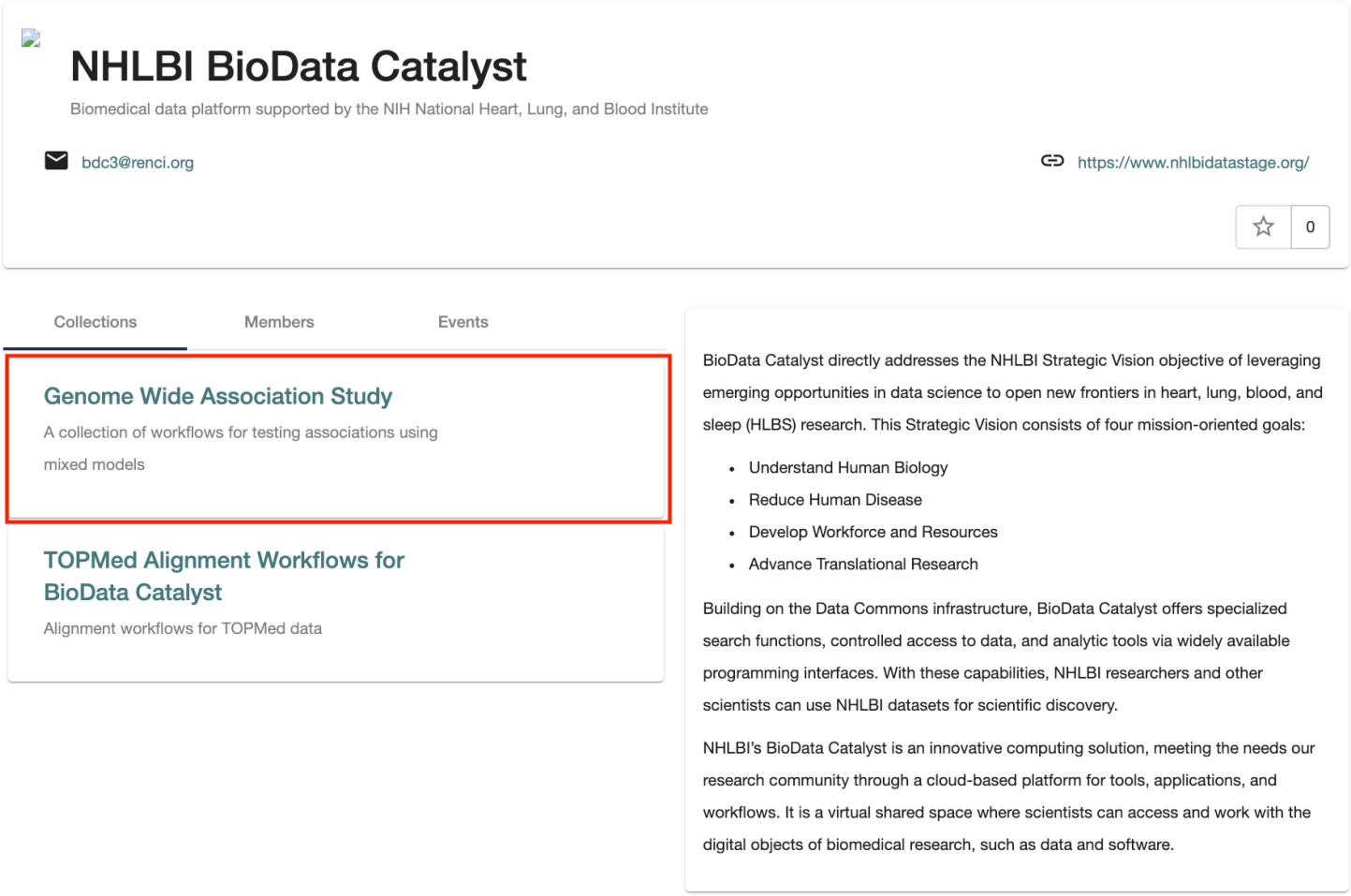
Find reproducible workflows in [Dockstore]

For this tutorial, we have already loaded Dockstore workflows and configured their inputs. However, we describe how you can easily import workflows into a Terra workspace as well.

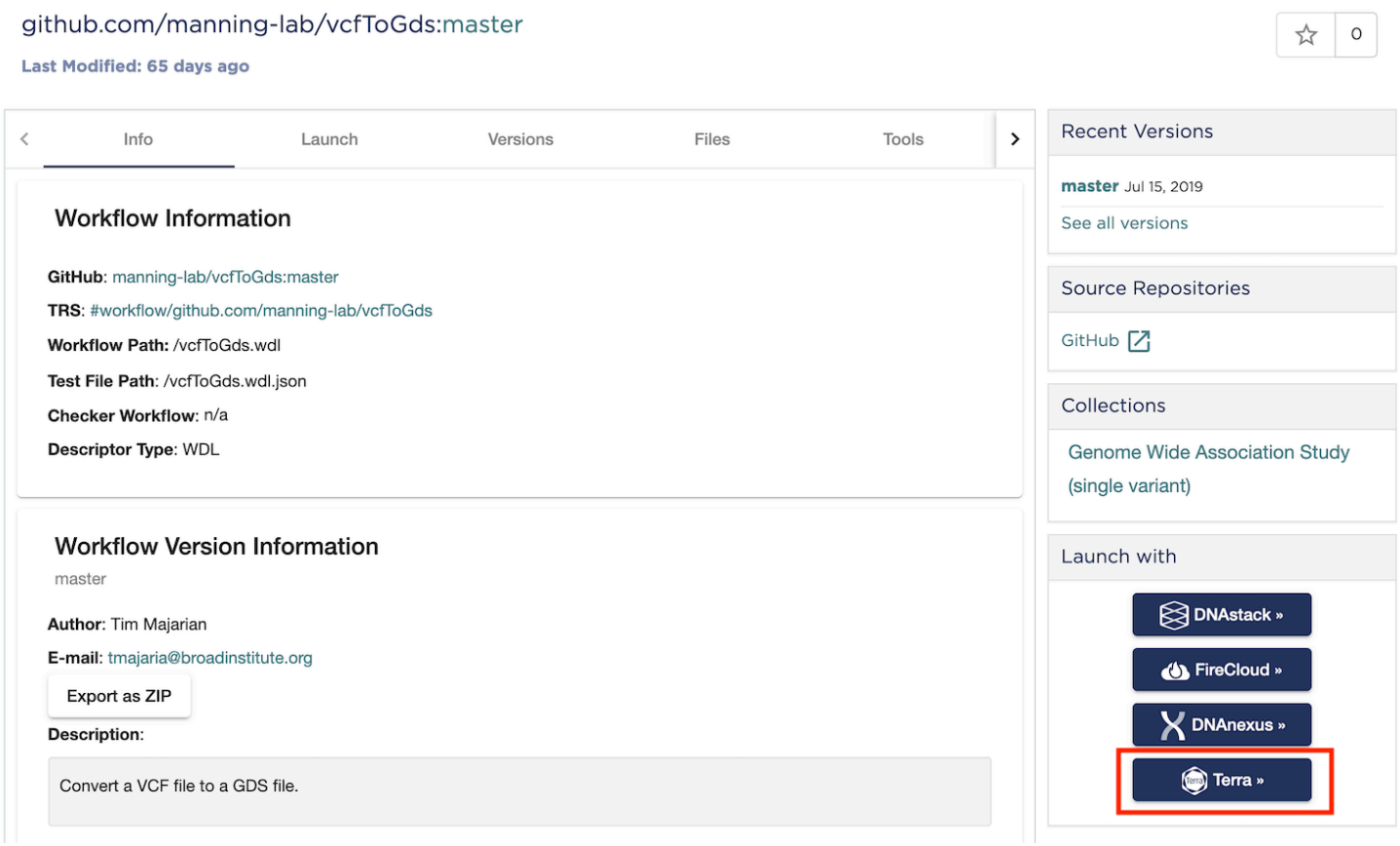
1. In your Terra workspace, go to the workflows tab.
2. Select "Find a Workflow".
3. At the bottom left, select the Dockstore icon under "Find additional workflows"..
4. This will open the search feature in Dockstore where you can browse easily reproducible workflows that have been written in workflow languages with their associated Docker images and json parameter files.
5. We will find workflows for DataSTAGE projects by clicking the "Organizations" button along the top banner.



6. Scroll down and click on the DataSTAGE organization. This will display collections of workflows, including the one we use in this tutorial under the collection "Genome Wide Association Study".



7. Four workflows will appear. Each one can be exported to Terra by clicking on a single workflow to view its contents, then clicking "Launch with Terra" at the lower right.



8. This will generate a new page asking you to "Select the workspace destination". 9. Navigate to the workspace you chose and select the workflows tab to check that each workflow was imported successfully.

Conduct a common variant GWAS in Terra

Part 2: Explore TOPMed data in a Jupyter Notebook

The **1-GWAS-preliminary-analysis** notebook explores the phenotype data by performing the following steps:

1. Use functions that were created to easily reformat TOPMed data in the Gen3 grap format into a single entity that can be loaded as a dataframe in your notebook.
2. Subset the dataframe to include only our traits of interest and remove any individuals that lack data for each specified trait.
3. Visualize phenotype distributions in a series of plots.
4. Filter genomic data to common variants.
5. Perform a principal component analysis (PCA) to assess if population stratification is detected in your cohort or not. If it is, it should be accounted for in association testing.
6. Generate a genetic relatedness matrix (GRM) for downstream use in association testing.

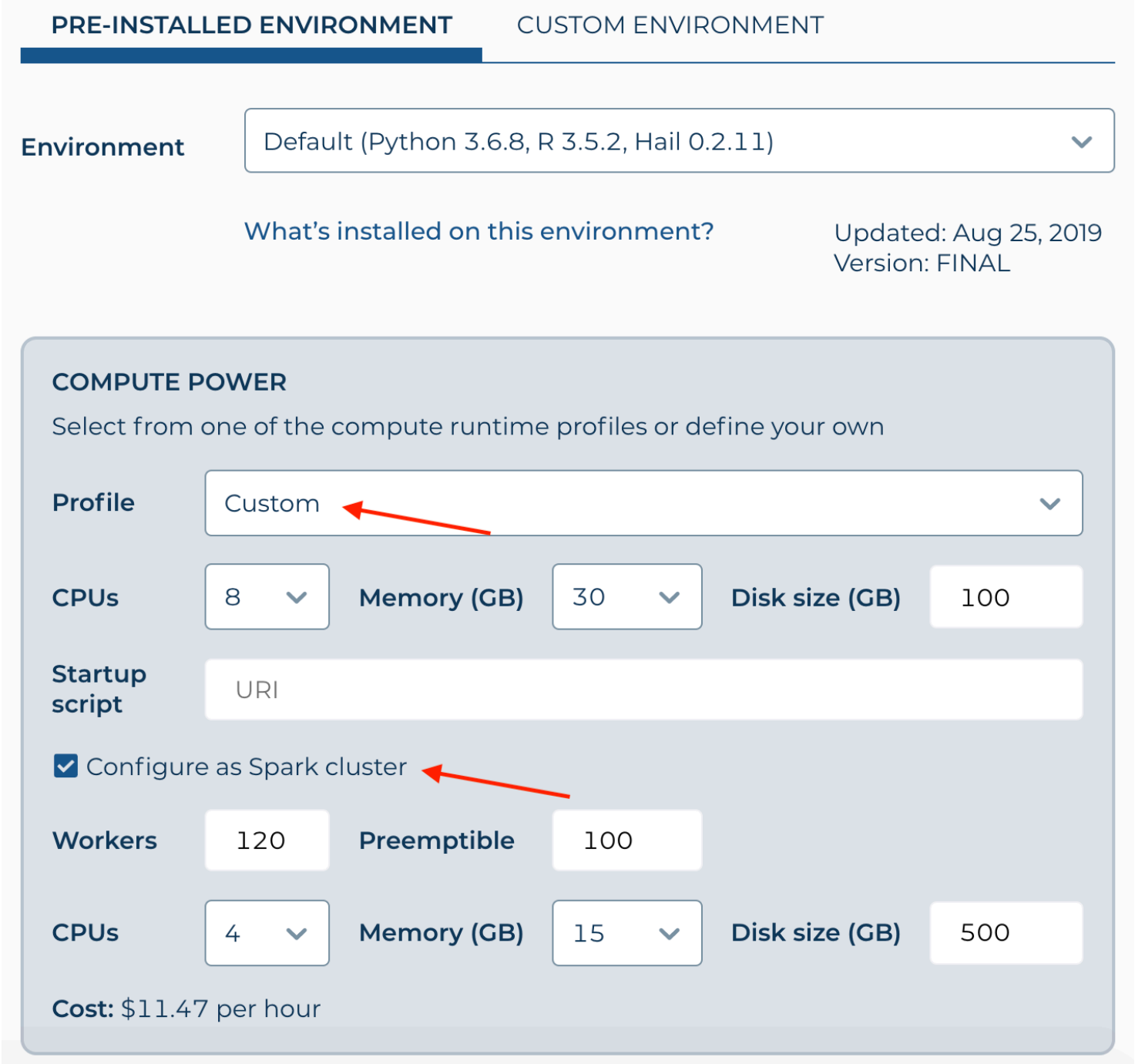
Genetic analyses in this notebook utilize the Hail software. Hail is a framework for distributed computing with a focus on genetics. Particularly relevant for whole genome sequence (WGS) analysis, Hail allows for efficient, nearly boundless computing (in terms of variant and sample size).

To facilitate the downstream analysis, this notebook is set up to save output data to the workspace bucket and then write the associated metadata and derived genetic data to the data model using the FireCloud Service Selector (FISS) package.

For more information on using Terra's data table see this article on "Linking data in a Google bucket to the workspace data table".

Time and cost estimate

You are able to adjust the runtime configuration to fit your computational needs in the Jupyter notebook. We recommend selecting the default environment and selecting the custom profile to use and configure the spark cluster for parallel processing. Using the suggested profile below, running this notebook on this dataset takes about 60 minutes and $1.37 to compute.



When working in a notebook that may have compute times over 30 minutes, learn more about Terra's auto-pause feature and how to adjust auto-pause for your needs. Please carefully consider how adjusting auto-pause can remove protections that help you from accidentally accumulating cloud costs that you did not need.

Part 3: Perform mixed-model association tests using workflows

Below, we describe the four workflows in this workspace and their cost estimates for running on the sample set we create in this tutorial. We have already configured input and output parameters for each workflow in the workflows tab. If you clone the workflows in this tutorial to other Terra workspaces, their parameters will come along.

1-vcfToGds

This workflow converts genotype files from Variant Call Format (VCF) to Genomic Data Structure (GDS), the input format required by the R package GENESIS.

Time and cost estimates

Sample Set Name	Sample Size	# Variants	Time	Cost $
Amish-systolicbp	1,052 samples	6,429,788	6m	$2.26

Inputs: * VCF genotype file (or chunks of VCF files)

Outputs: * GDS genotype file

2-genesis_nullmodel

- This workflow generates a mixed model under the hypothesis of no variant effect using the GENESIS software.

Time and cost estimates

Sample Set Name	Sample Size	Time	Cost
Amish-systolicbp	1,052 samples	4m	$0.02

Inputs: * GDS genotype file * Genetic Relatedness Matrix * Trait outcome name * Trait outcome type * CSV file of covariate traits * Sample ID list

Outputs: * A null model as an RData file

3-genesis_tests

This workflow generates per-variant association statistics with our mock phenotype, using the null model.

Time and cost estimates

Sample Set Name	Sample Size	# Variants	Time	Cost
Amish-systolicbp	1,052 samples	6,429,788	8m	$0.55

Inputs: * GDS genotype file * Null model as an RData file

Outputs: * Compressed csv file(s) containing raw results

4-summaryCSV

This workflow combines multiple summary statistics files produced above and generates quantile-quantile (QQ) and Manhattan (MH) plots (common descriptive figures of the GWAS results). The workflow outputs the plots (as a pdf) and two CSV files: one with all results combined into a single table and one a table of top results with p-value less than the specified threshold.

Time and cost estimates

Sample Set Name	Sample Size	# Variants	Time	Cost
Amish-systolicbp	1,052 samples	6,429,788	17m	$0.06

Inputs: * Compressed csv file(s) containing raw results

Outputs: * CSV file containing all associations * CSV file containing top associations * PNG file of QQ and Manhattan plots

---

Part 4: Interactive GWAS summarization in a notebook

The final workflow (4-summaryCSV) outputs two csv files with association results (the top candidates and all associations) as well as a png containing summary figures. The **2-GWAS-summarization** notebook is another resource for viewing results of your analyses and the notebook feature allows you to edit for specific use cases.

Bring your own data

Both the notebook and workflow can be adapted to other genetic datasets. The steps for adapting these tools to another dataset are outlined below:

Update the data tables Learn more about uploading data to Terra here. You can use functions available from the terra*data*util companion notebook to consolidate new data tables you generate.

Update the notebook Accommodating other datasets may require modifying many parts of this notebook. Inherently, the notebook is an interactive analysis where decisions are made as you go. It is not recommended that the notebook be applied to another dataset without careful thought.

Run an additional workflow You can search Dockstore for available workflows and export them to Terra following this method.

Authors, contact information, and funding

This workspace is a product of the Manning Lab and NHLBI's BioData Catalyst, in collaboration with the Computational Genomics Platform at UCSC Genomics Institute and the Data Sciences Platform at The Broad Institute. Contributing authors include:

- Tim Majarian (Manning Lab)
- Alisa Manning (Manning Lab)
- Beth Sheets (UC Santa Cruz Genomics Institute)
- Michael Baumann (UC Santa Cruz Genomics Institute)

Workspace change log

Date	Change	Author
December 3, 2019	Gen3 updates	Beth
November 22, 2019	Updates from Alisa	Beth
October 22, 2019	User experience edits from Beri	Beth