Notebooks QuickStart Worksheet

Welcome to the Notebooks QuuickStart Tutorial. Follow these step-by-step instructions to learn how to run an interactive analysis on a subset of data from the Data Library.

Each step is independent, with time and cost estimates for completion listed below. You will need to do the steps in order, but you don't need to do them in one setting.

Step 1: Explore data in the Data Library

(10 minutes; $0.00 to complete this step)

- Learning objectives
- Time and cost to complete
- Step-by-step instructions

Step 2: Import data to the workspace

(two minutes and $0.00 to complete this step)

- Learning objectives
- Time and cost estimates
- Step-by-step instructions

Step 3: Set up a virtual application (notebook) for analysis

(20-30 minutes and less than $1.00 to complete this step)

- Learning goals
- Time and cost estimate
- Notebooks background
- What happens when I open a notebook for the first time?
- Step-by-step instructions

Step 4: Run an interactive analysis in a notebook

(20 minutes and less than $1.00 to complete this step)

- Learning goals
- Time and cost estimates
- Step-by-step instructions

Step 1: Explore data in the Data Library

Learning objectives

In this step, you'll learn how to 1) access and explore data using the Data Explorer in the Data Library and 2) use selection criteria to define a subset (custom cohort) of participants for analysis

Time and cost to complete

This step should take 5 - 10 minutes and won't cost anything.

Background

Before running the analysis notebook, you will need to create a custom subset (cohort) of data to study. For this quickstart we will be using 1,000 Genomes data in the Data Library. You'll select exclusion criteria within the Data Explorer to generate a subset (custom cohort), and import the cohort to the workspace as a table in the Data tab.

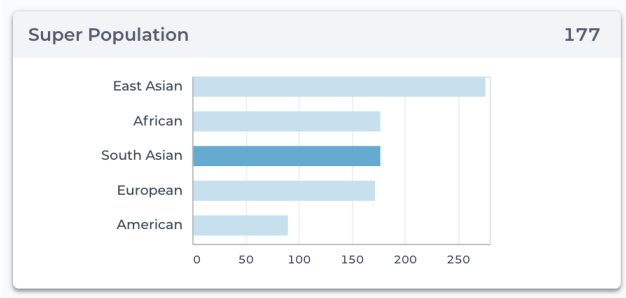
Step-by-step instructions

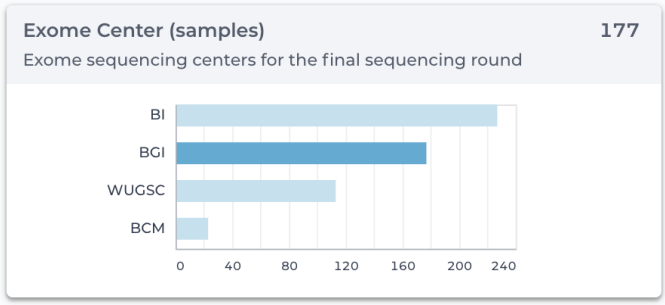
1.1. Go to the "Data Library" at https://app.terra.bio/#library/datasets

1.2. Click the button to browse the "1,000 Genomes Low Coverage" dataset. You can see there are several parameters, with bars that indicate how many participants in the dataset satisfy those parameters.

1.3. Select the exclusion criteria for your subset by clicking on one or more bars in the display panes

For example, to restrict your study to participants of South Asian descent whose exome sequencing center was BMI, you would choose those criteria in the cards following the screenshots below:





Step 2: Import data to the workspace

Learning objectives

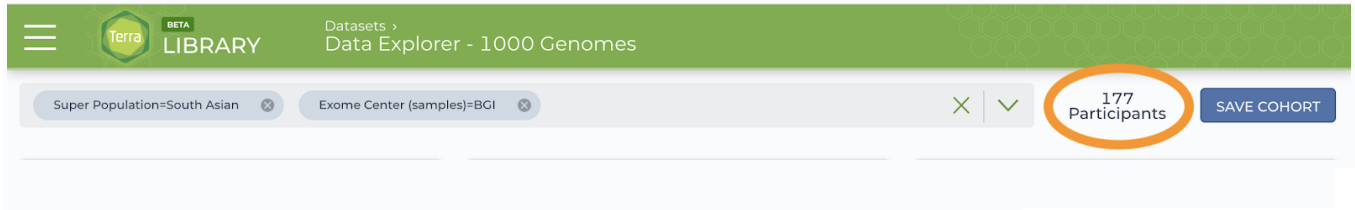
By the end of this step, you'll know how to export subsets of data from the Library to your workspace

Time and cost estimates

Expect to take about two minutes and to spend nothing to complete this step.

Step-by-step instructions

2.1. Click "Save Cohort" (blue button at top right) to save in a Terra workspace. Notice the number of participants in your cohort (circled in the screenshot below):



2.2. Remember to name your selection something you will remember easily!

2.3. Destination workspace: Choose your copy of this workspace from the dropdown menu

2.4. Click "Import"

You'll be taken to the Data tab of your workspace copy. Notice the two new tables, a "BigQuery" table and a "cohort" table.

Background on imported data tables

- The **cohort table** contains a SQL query that returns a list of IDs for those participants that satisfy the the exclusion criteria you specified in the Data Explorer
- The BigQuery table references all the data in the BigQuery dataset

In Step 4 you will join the subset IDs and the BigQuery data to get the data only for those participants in your subset. You'll bring that data into a Jupyter notebook for further analysis.

Step 3: Set up a virtual application (notebook) for analysis

Learning goals

After this step you'll understand how to run a notebook in Terra, as well as the general structure and purpose of a setup notebook.

Time and cost estimate

Running the setup notebook should take about 20 minutes (including the time to create the virtual machine or cluster) and cost less than $0.25.

Notebooks background

A notebook runs in a virtual computing environment ("application compute"). The first time you run a notebook and create your virtual environment, you will likely need to install a number of libraries and packages that are not included by default. This notebook does that step. You only need to run it once!

Notebooks in this workspace

This section uses three notebooks , which should be run in the order listed below. An optional 0_Jupyter_intro notebook - included in the Notebooks tab - gives a general intro (or refresher) on Jypyter notebooks basics.

- 1_R_environment_setup - You must first run this notebook every time you create a new notebook virtual environment ("application compute") in this workspace. It will import the libraries needed to run the next two notebooks.
- 2_BigQuery_cohort_analysis - Analyze data from a workspace cohort table in a notebook.
- 3_Access_and__plot_public_BigQuery - Explore two additional ways to access data in the cloud (unstructured data in a Google bucket and tabular data in BigQuery) in a notebook.

What happens when I open a notebook for the first time?

- When you first open a notebook in a workspace notebook, Terra creates your application compute (can be a VM or cluster). **This can take 5-10 minutes.** During this time, don't refresh the page or try to restart the notebook.
- During creation, you will see a read-only notebook copy and a note in the top of the browser that Terra is creating the virtual environment (in the orange rectangle)

- If you open any notebook again in your workspace, it won't take as long, as Terra will only need to restart the application compute, not create it.

Step-by-step instructions

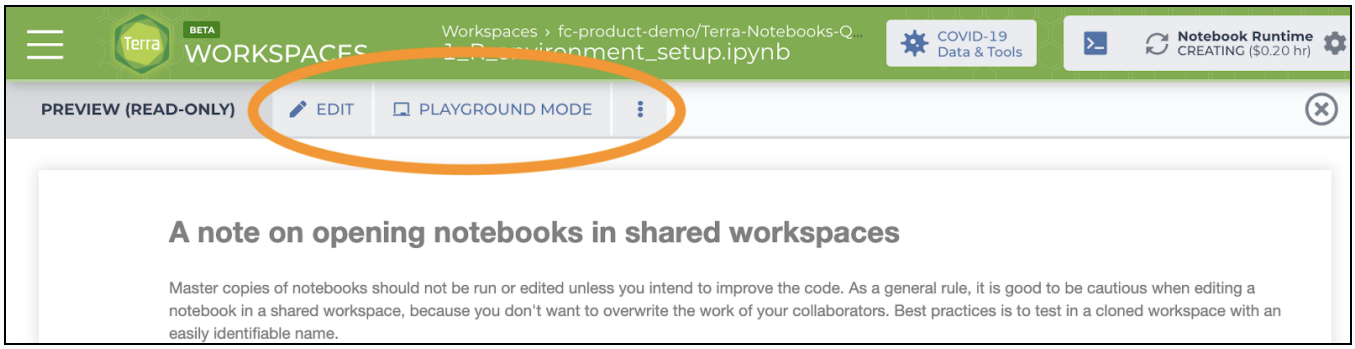
Note on running code

Notebooks include both documentation and code cells. The documentation should provide enough background to understand the code, even if you are not familiar with coding.

To learn more about notebooks in Terra, see this article.

3.1. Click on the "**Notebooks**" tab of your cloned copy of the Notebooks_Quickstart workspace and then click on the “1_R_environment_setup” notebook

3.2. Click on the "Edit" button in the Notebook interface to start a VM or cluster compute environment where you can edit and run the code in the notebook :



.3.3. Wait for the application compute to start. *This can take 5-10 minutes the first time you create a notebook virtual environment in a workspace. *

3.4. While the virtual compute environment spins up (this can take 5-10 minutes the first time you create a virtual application compute in a workspace), you can skim the read-only view to understand how it works

3.5. Once the notebook is up and running, go to the first code cell, click in the cell and then click 'Run' to execute the code in that cell. Note - you can also use the shortcut “shift” + “return” to run a cell.

36. Wait for the * to turn into a number, i.e. ['*'] --> [4], which indicates that the code in the cell has executed successfully

3.7. Click in the next cell and then click 'Run' to execute the code in the next cell, wait for execution to complete and review the results.

3.8. Continue steps 7-9 until you've executed the code in all cells from the notebook

3.9. When you've run all the code cells, close the notebook by clicking the green x in the top right.

Congratulations! You've run your first interactive notebook application in Terra!!

Step 4: Run an interactive analysis in a notebook

Learning goals

After running this notebook, you should understand how to import the cohort data you imported from the Data Library into the notebook application memory in order to run an analysis. It has some steps to verify the data, but does not go into an in-depth analysis. At this point, you could do any R or Python-based analysis interactively.

Time and cost estimates

This notebook should take about twenty minutes and cost less than $0.25 to complete.

Step-by-step instructions

4.1. Open the notebook 2_BigQuery_cohort_analysis in Edit mode following the steps above

4.2. Read the documentation and run the notebook cells in order

4.3. For an additional way of importing data into a notebook, open and run the third notebook, “3)Public_BigQuery, executing each cell in order.