

HSLs GWAS Template Instructions

Here are the guidelines for running the GWAS templates provided by the Health Sciences Library System for the All of Us Researcher Workbench. Carefully review each step in this document. Upon completion, you'll have everything you need to carry out a basic Genome Wide Association Study using the All of Us short-read Whole Genome Sequence (srWGS) data, focusing on the Hail Exome subset.

Step 1: Upload the Template to Your Workspace

[Upload the template](#) to your workspace before continuing to the next steps.

Step 2: Create Your Cohorts and Datasets

Create your final phenotype dataset for analysis. It should include the **outcome variable**, the **adjustment variables**, and the **person id** for each participant in your cohort.

As you are creating the final phenotype dataset, keep the following in mind:

1. If you select the binary outcome template, ensure the outcome is coded as 1 for positive and 0 for negative.
2. Before using the templates, review the assumptions of linear and logistic regressions.
3. Make sure you do not have missing or skipped values for all variables in the dataset.

Step 3: Save the Final Phenotype Dataset to Your Workspace

Save the final phenotype dataset to your workspace **as a tsv file** [Figure 1]. This is so it can be imported correctly in the GWAS template.

```
# import library
import subprocess
import os
# save dataframe in a csv file in the same workspace as the notebook
joined_df.to_csv('pheno.tsv', index=False, sep='\t')

# get the bucket name
my_bucket = os.getenv('WORKSPACE_BUCKET')

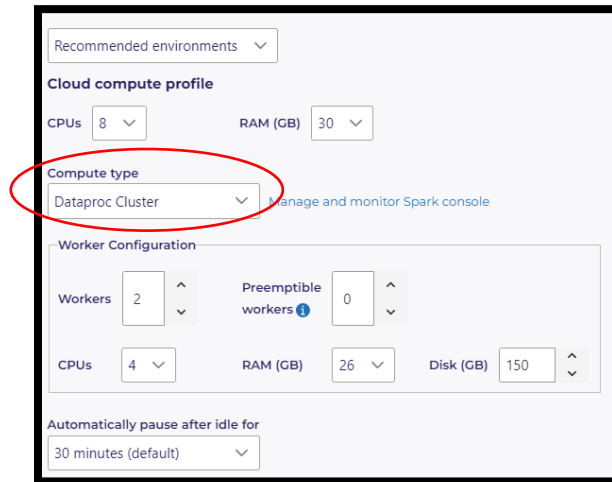
# copy csv file to the bucket
args = ["gsutil", "cp", f"./pheno.tsv", f"{my_bucket}/data/"]
output = subprocess.run(args, capture_output=True)

# print output from gsutil
output.stderr
```

Figure 1: An example of Python code for saving phenotype data to the user's workspace.

Step 4: Create Appropriate Cloud Environment for Genomic Analysis

Before opening the template in your workspace, make sure you are working in the **Dataprocc** cluster [Figure 2]. The default is the Standard VM, which will cause the template to crash.



The screenshot shows the 'Cloud compute profile' configuration in the Google Cloud console. The 'Compute type' dropdown is highlighted with a red circle and set to 'Dataprocc Cluster'. Other settings include CPUs: 8, RAM (GB): 30, Workers: 2, Preemptible workers: 0, CPUs: 4, RAM (GB): 26, and Disk (GB): 150.

Figure 2: The cloud compute profile for genomic analysis in All of Us highlighting the Dataprocc cluster as compute type.

Step 5: Edit the Appropriate Fields in the Template

Once you are in the template notebook, go to the designated cells for editing and input the correct information. Each section that can be edited is clearly marked with “### PLEASE EDIT THIS SECTION ###” so users know where changes can be made [Figure 3]. Cells that can be edited are located under the following sections: **Import Phenotype Data**, **Pulling Specific Regions of the Genome**, **MAF**, **Hardy-Weinberg Equilibrium**, **Linear Regression**, and **Stratification**.

```
##### PLEASE EDIT THIS SECTION #####  
file_name = "file_name_here.tsv"
```

Figure 3: An example of a cell that can be edited in the GWAS template, located under the **Import Phenotype Data** section.

Step 6: Run the Template!

Run all the cells. This will take a while, so you may want to run the notebook in the background.