# ColabADMIXTOOLS

## Quickstart Guide (2nd Edition)

**Written by Florio**

**Revised January 13th, 2025**

# Table of Contents

# What is ADMIXTOOLS?

ADMIXTOOLS is a powerful software suite designed for population genetics research. It provides tools to analyze genetic data and uncover ancestry, admixture, and population structure patterns in humans and other species. These tools have been widely used in academic studies and are now increasingly accessible to hobbyists interested in exploring ancestry using genetic datasets.

ADMIXTOOLS offers two main versions:

**ADMIXTOOLS 1**:

- The original version was developed to perform tests such as f-statistics (e.g., f3, f4, D-statistics) and qpAdm modeling. It is written in C and is highly efficient for standard analyses of admixture and phylogeny.

**ADMIXTOOLS 2**:

- The updated version, written in R, is more user-friendly and includes improved workflows for hobbyists and researchers alike. It integrates modern computational tools, making processing and visualizing genetic data easier.

ADMIXTOOLS is most commonly used to:

- Detect gene flow (admixture) between populations.
- Test hypotheses about relationships among populations.
- Build and refine population phylogenies.
- Analyze ancient DNA data to explore human prehistory.

ADMIXTOOLS is especially suited for those working with genome-wide data, such as SNP arrays or whole-genome sequences. By leveraging its statistical methods, users can uncover fascinating stories about genetic history and population interactions.

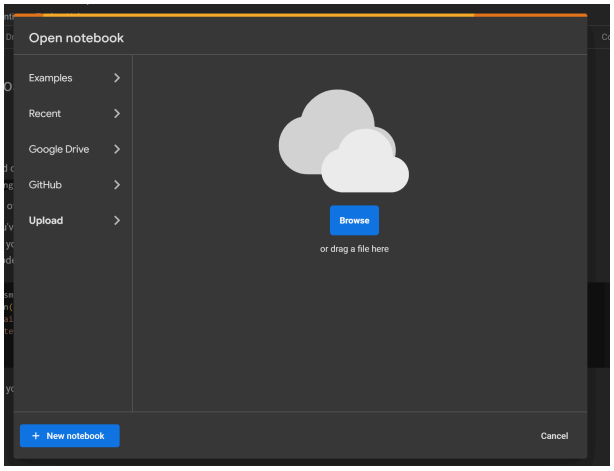
## Useful Links and Documentation:

- Ancient Admixture in Human History | Genetics | Oxford Academic
- Supplemental Material for Harney et al., 2020
- Assessing the performance of qpAdm: a statistical tool for studying population admixture | Genetics | Oxford Academic
- ADMIXTOOLS 2 Tutorial
- qpAdm best practices and common pitfalls | Indo-European.eu
- Testing times: disentangling admixture histories in recent and complex demographies using ancient DNA | Genetics | Oxford Academic

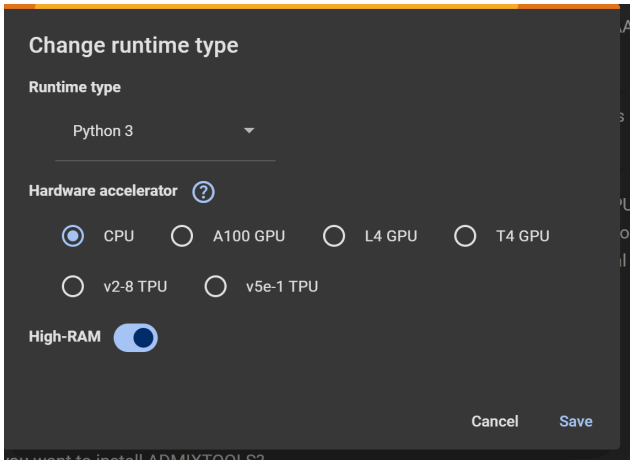
## Getting Started

Go to Colab. You will need a Google account to access this tool. Preferably at least 2 GB of Drive space if you want to merge, store your merged files, and return to modeling.

Upload the .ipynb (interactive python notebook) to Google Colab from your device file explorer.

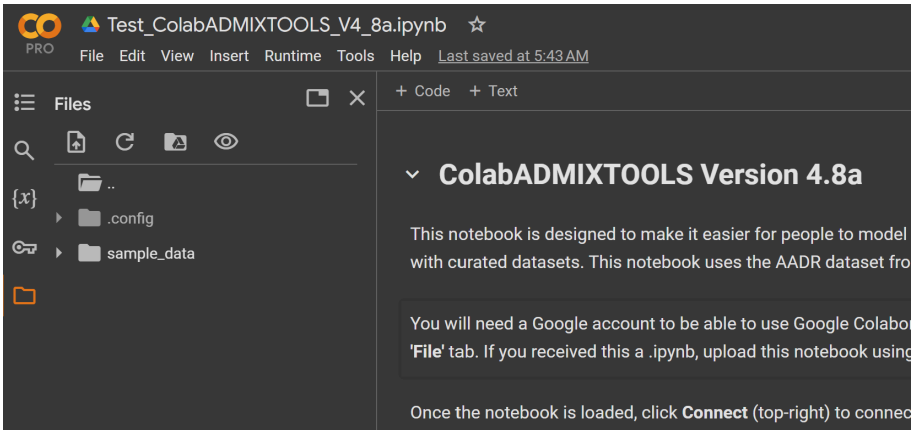



Go to Connect and change the runtime type, make sure you are connected to CPU and Python 3. There is an R environment option but we can bypass this later with a python package called rpy2. A standard free Colab runtime gives you a 12 GB CPU. I have ColabPro so I have access to 25 and 50 GB CPUs.




For this guide, I will be using Version 4.8a of my ColabADMIXTOOLS notebook. I cannot guarantee the same performance for older versions that may be circulating.

I like having the Colab File Explorer open on the left at all times to track my files. By default, our Working Directory will be the **/content/** folder. All of our intermediate and final files will be saved there. Colab runs on a Linux environment so it is best to at least understand **Linux filesystem** and filepath naming conventions going forward.

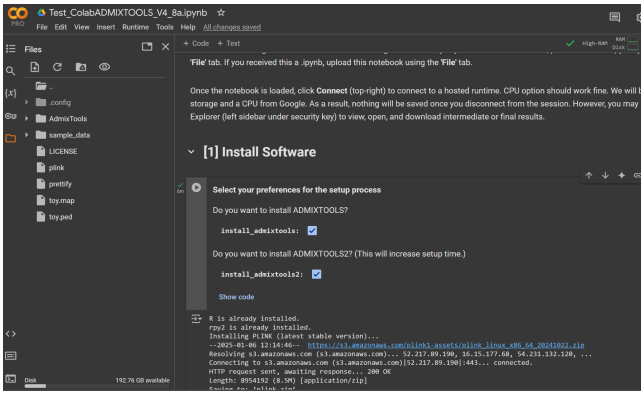


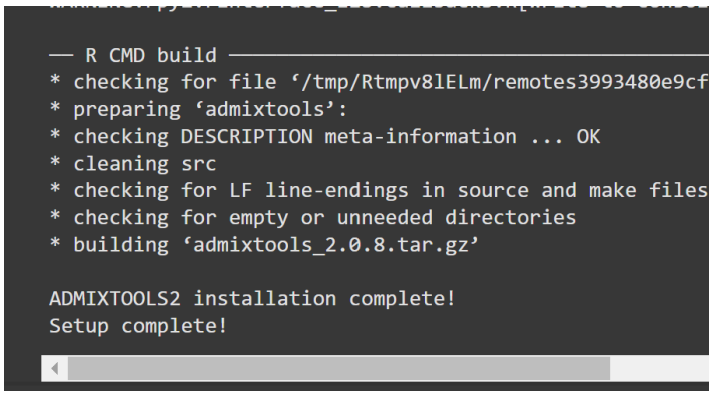

## [1] Install Software

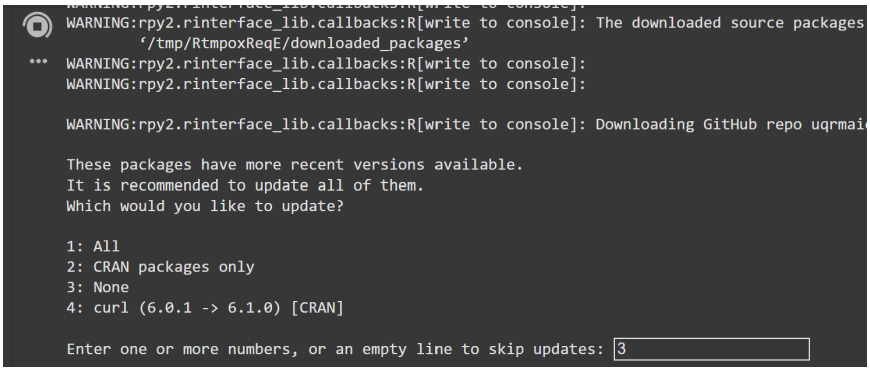
**Time:** Estimated 7 minutes and 32 seconds (both AT1 and AT2)

Go ahead and Click on the Play Button for the first Code Cell. First, we will ensure rpy2, PLINK, and R are ready to go in the environment. Then we will install both AT1 and AT2. If for whatever reason you wish not to have one or the other, you may uncheck the box. If you are curious about

the underlying code, feel free to click on Show Code. Double-click outside the code to hide the code.




If you get a prompt to update something, you can enter the number 3 and press enter to skip dependency updates for now. Performance should be the same regardless of what option you pick.

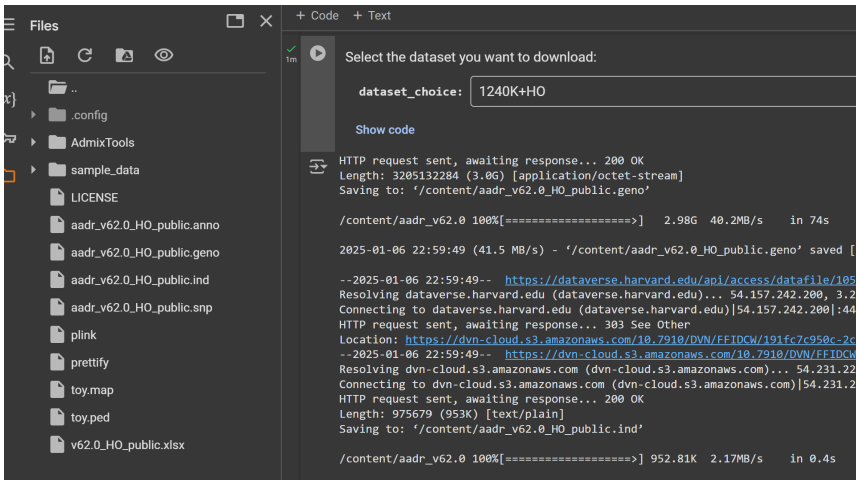
## [2] Your DNA File Upload

This will be useful later on for those interested in modeling themselves but for now, let us skip it.

## [3] Dataset Selection

**Time:** 45 seconds for 1240K+HO AADR, 1240K 1 min 18 seconds.

We will assume you know the difference between the two datasets. If you want to know more about the AADR, read **this paper**. For simplicity, I will download the 1240K+HO as it has less coverage and thus quicker analysis steps.




# [4] Merging Yourself with the AADR

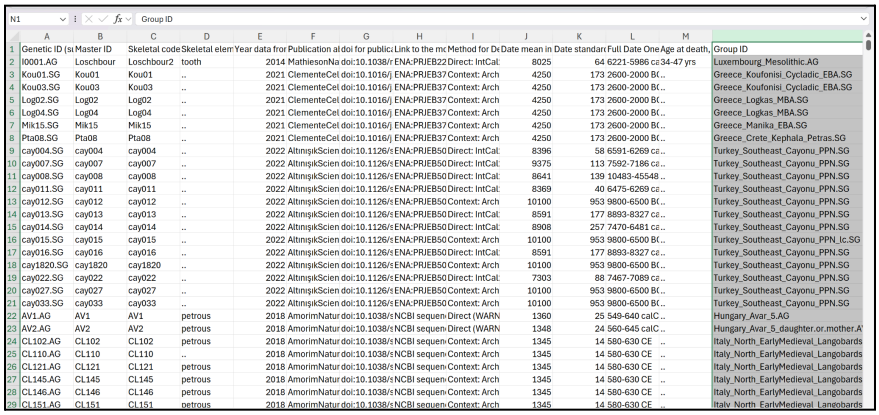
**Time:** Merging with HO about 5-8 minutes. Merging with 1240K about 6-16 minutes.

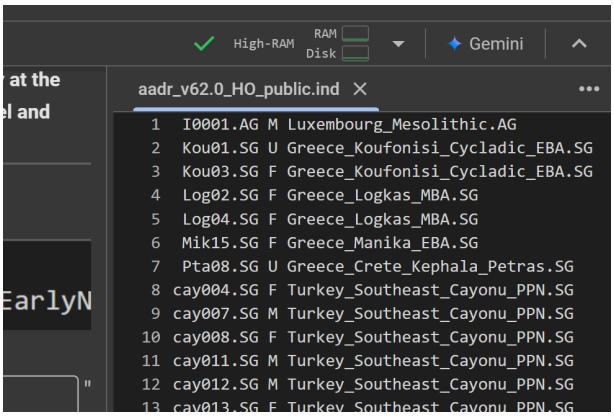
**Merging Overlap Reference:** AncestryDNA overlaps with 1240K: about 390K snps. With HO: 130K snps. MyHeritage overlaps with 1240K: about 177K snps. With HO: about 60K snps. 23andMe overlaps with 1240K: about 140K snps. With HO: about 50K snps.

This will be useful later on if you want to model yourself, for now, we will skip this.

# [5] AT2 qpAdm Prep and Running

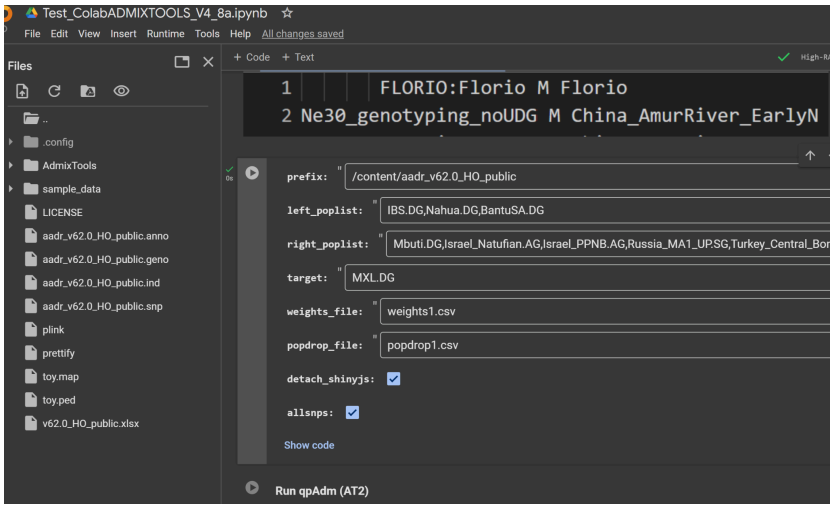
On the left is a *.xlsx* file corresponding to the HO dataset samples. Double-click or right-click to download it to your local device and open it in Excel or Sheets. For population selection, we will be looking at the **Group ID** column. Feel free to browse through the rest of the samples and information stated there (e.g. locality, country, haplogroups, **Sample ID**, associated paper, coverage, etc). Alternatively, you can double-click on the *.ind* file and it should open on the right, with the Group ID of the population being the last column.






For this guide, I will be using my recently downloaded AADR HO dataset, so I will point to the prefix of that dataset, which we can manually type in or right-click on any of its files to copy the file path (either .geno, .ind, .snp file). So the file path should be **/content/aadr_v62.0_HO_public** as it is the prefix without any file extensions.

I have chosen **MXL.DG** as the Target population we wish to make a random model. I will explore modeling them with **IBS.DG,Nahua.DG,BantuSA.DG** as my Left/Source populations. I used: **Mbuti.DG,Israel_Natufian.AG,Israel_PPNB.AG,Russia_MA1_UP.SG,Turkey_Central_Boncuklu_PPN.AG,Turkey_Central_Pinarbasi_Epipaleolithic.AG,Morocco_Iberomaurusian.AG,Serbia_IronGates_Mesolithic.AG,Luxembourg_Mesolithic.DG,Russia_YuzhniyOleniyOstrov_Mesolithic.AG** as my Right/Reference populations. Please read the Documentation for details.




You may choose to rename the weights and popdrop files if you will be doing repeated runs and choose not to overwrite existing results. Moreover, leave **allsnps** and **detach_shinyjs** checkboxes checked unless you know what they do. For now, click Play on that code cell to save those

parameters, then let us click Play on the next code cell to Run AT2 qpAdm. We should get a bunch of information. At the bottom, there will be a table along with a popdrop table.

```
# A tibble: 3 × 5
  target left       weight      se     z
  <chr>  <chr>       <dbl>   <dbl> <dbl>
1 MXL.DG IBS.DG     0.430  0.0110   39.1
2 MXL.DG Nahua.DG   0.534  0.0123   43.6
3 MXL.DG BantuSA.DG 0.0364 0.00347  10.5
```

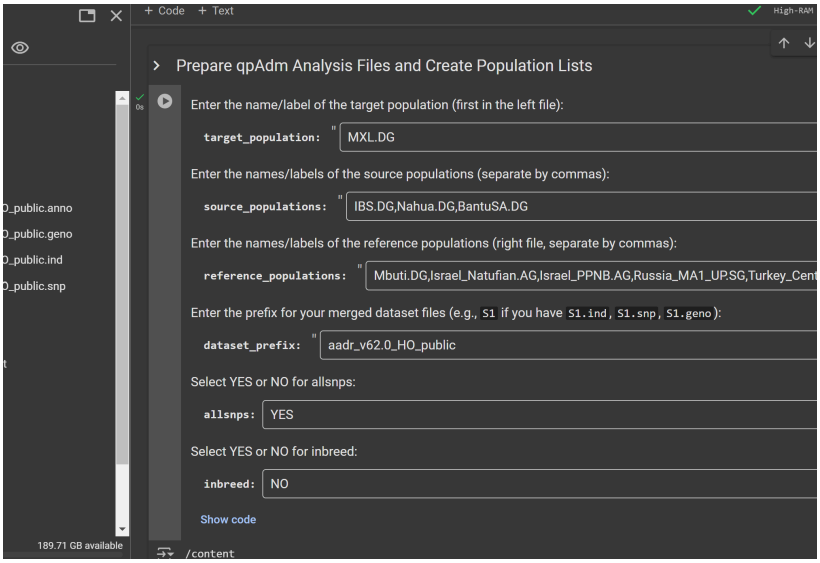
In short, this random model says MXL.DG can be modeled as having 43.0% IBS.DG with a SE of 1.1%, 53.4% Nahua.DG with a SE of 1.2%, and 3.6% BantuSA.DG with a SE of 0.35%. This makes sense considering Mexicans in Los Angeles are primarily of Iberian and Amerindian mixed ancestry with minor Sub-Saharan African admixture (*on average*). Now let's look at the popdrop table under it.

```
# A tibble: 7 × 14
  pat      wt   dof   chisq        p f4rank IBS.DG Nahua.DG BantuSA.DG feasible
  <chr> <dbl> <dbl>   <dbl>    <dbl>  <dbl>  <dbl>    <dbl>      <dbl> <lgl>
1 000       0     7    25.8 5.53e- 4      2  0.430    0.534     0.0364 TRUE
2 001       1     8   110.  4.00e-20      1  0.380    0.620    NA      TRUE
3 010       1     8  2425.  0             1  0.898   NA         0.102  TRUE
4 100       1     8   478.  3.82e-98      1 NA        1.01     -0.0127 FALSE
5 011       2     9  3982.  0             0  1       NA        NA      TRUE
6 101       2     9   483.  2.15e-98      0 NA        1        NA      TRUE
7 110       2     9 16542.  0             0 NA       NA         1      TRUE
```

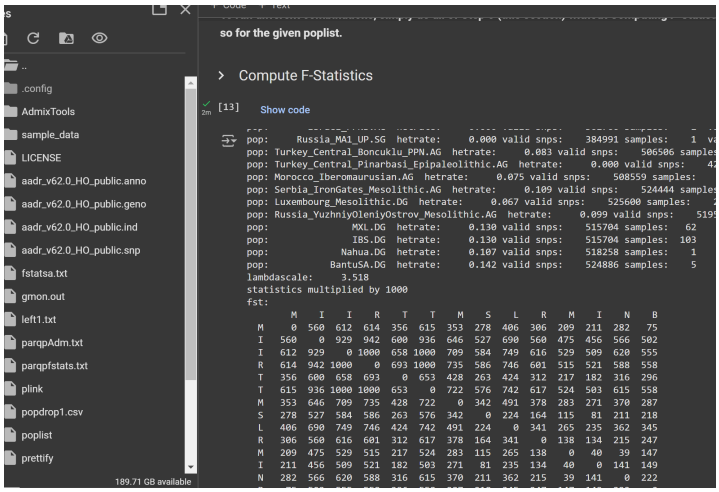
The pat column tells you which Left/Source population is being used (with a 0) and what happens if it is not being used (with a 1). This table is great for diagnosing sample selection. As we can see, our initial model is feasible but the p-value of 5.53e-4 is relatively low. For reference, a p-value close to and above 0.05 is enough to reject the assumption that no admixture occurred. This is counter-intuitive, I know. In our scenario, the higher the p-value the better. Please read on Model/Sample Rotation within the Documentation links above for more information on interpreting and troubleshooting a model.

If you get a negative coefficient we can automatically reject the model. Since we have positive coefficients and it is somewhat feasible, we can go ahead and use AT1 to cross-reference our model.

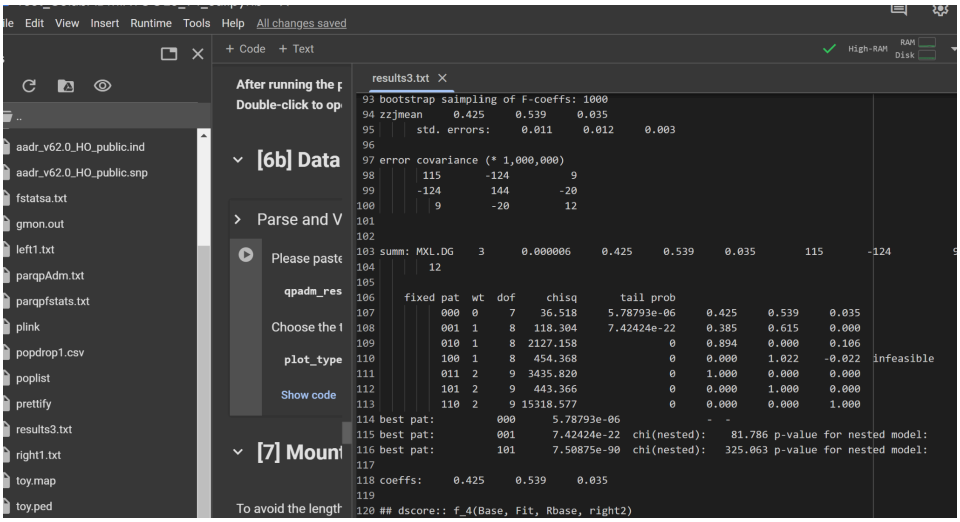
# [6] AT1 qpAdm Prep and Running




The parameter format is the same as before. Note pointing to the dataset prefix does not require */content/* in this step. Leave **allsnps** and **inbreed** params the same for now. Let's click through these code cells. It is good practice to **Compute Fstats** for our population list so that we can remove and swap around Target, Left, and Right sources without having to re-compute their Fstats.

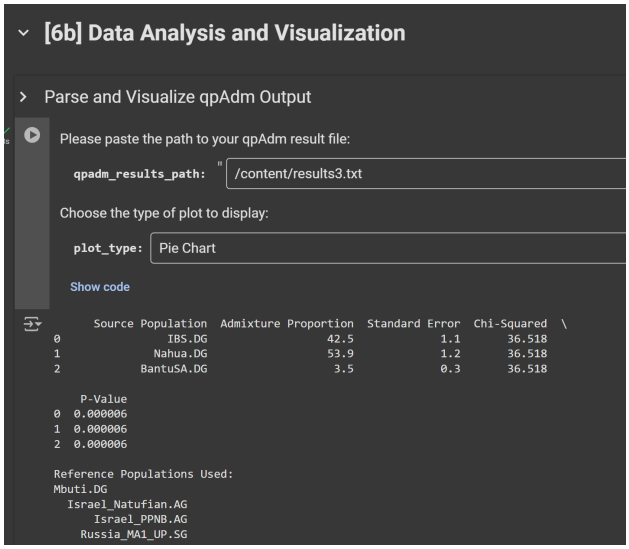


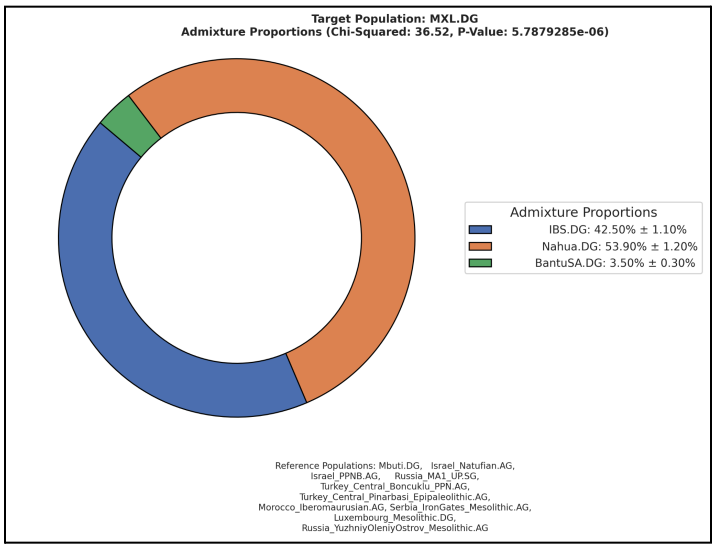

Now let's run the AT1 qpAdm. Feel free to change the results output name if you do not want to overwrite previous runs. We should be able to see our weights/coefficients and popdrop table for our model in the results *.txt* file.




# [6b] Data Analysis and Visualization

It can be troublesome to read through the results output file sometimes. In this section, we can point to our results file path and extract the necessary information we need for visualization (assuming non-negative percentages). For now, let's do a simple Pie Chart.




The weights/coefficients/percentages corroborate our AT2 model but the p-value dropped a bit. Technically it is feasible but we could improve this model with an improved Right list. For the instructive purposes of this guide, this random model will suffice.

# Modeling Yourself with the AADR

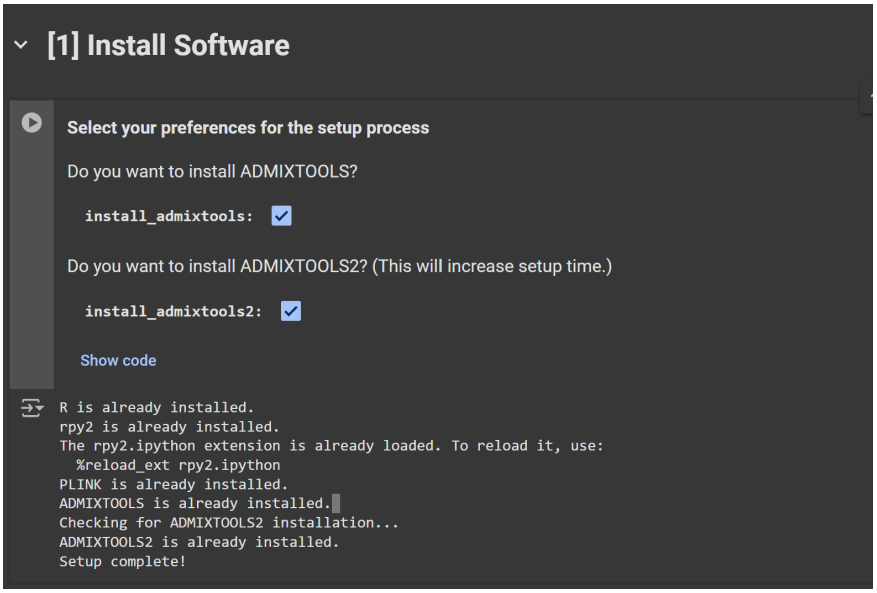
Now that we understand the workflow of the notebook and the process. We can choose to model ourselves with the AADR. However, the overlap will vary. Generally, AncestryDNA files have the most overlap, followed by MyHeritage and 23andMe. It is possible to impute your raw file using the **DNAGenics imputation service** and increase your coverage; however, this statistical imputation and downstream models may not reflect true admixture.

Moreover, if you have tested with multiple kits you can combine them to maximize coverage with the AADR using DNAGenics DNA Kit Studio (see above).

If you have done WGS, then converting to a workable format is possible by using **WGS Extract**, then converting that vcf to eigenstrat format (see **here** to go down that rabbit hole). I have never done WGS so cannot attest to the performance of this workflow.

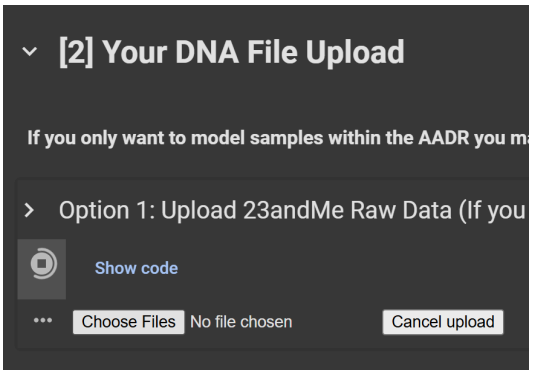
# [1] Install Software

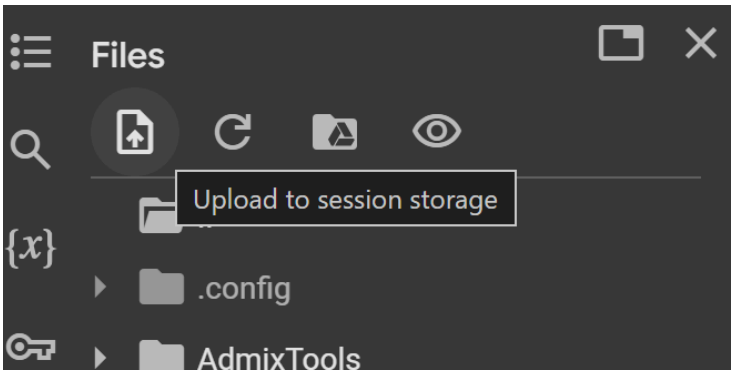
First, let's start at the top of the notebook by double-checking that everything is installed and loaded into the environment correctly. If you are ever disconnected or receive an unexpected error, double-checking **Step 1** should be the first troubleshooting step you should perform after checking for typos and correct file paths.



# [2] Your DNA File Upload

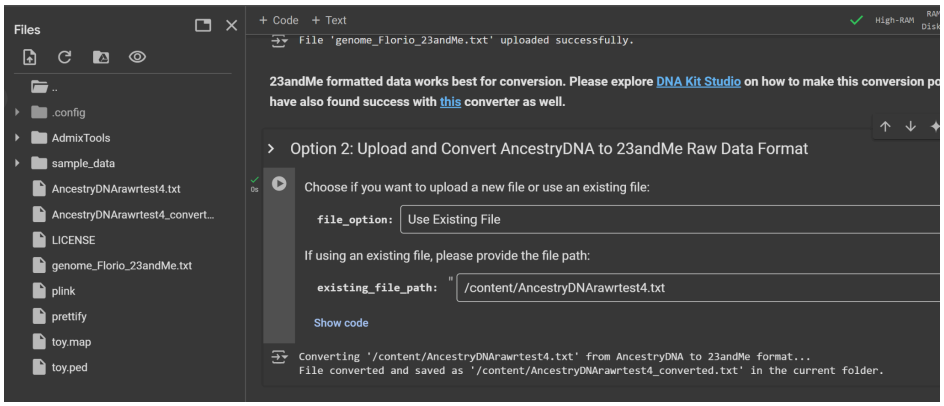
Now let's get our raw file into the Colab environment. **Option 1** is straightforward for those with raw data already in 23andMe format. We simply Play the code cell, Choose Files, and then select our 23andMe formatted file using our local device file explorer. Alternatively, you can upload to this session manually (for any file from your local device) using the Colab File Explorer upload button.



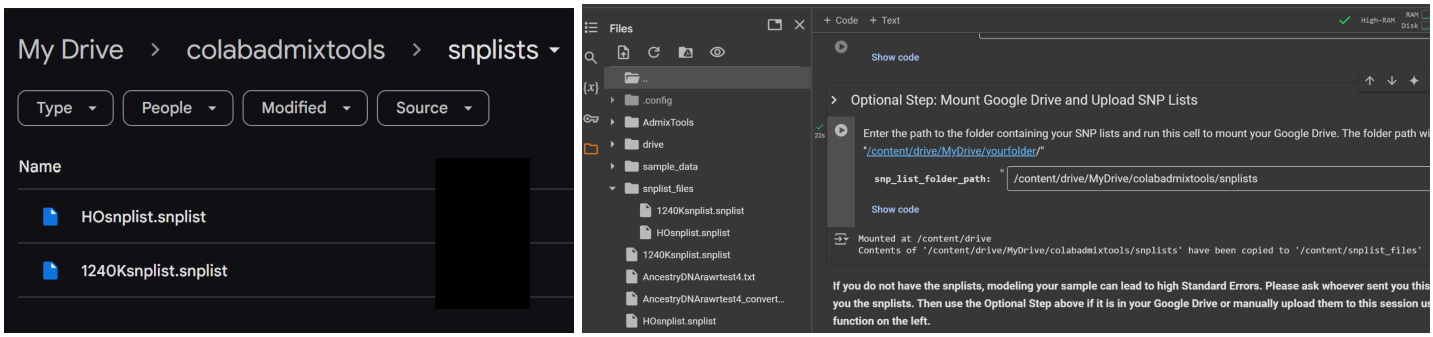

I mention **DNA Kit Studio** and **another converter** for other formatted files from other companies. Feel free to try those out if none of my converter scripts work for you.

If you have an AncestryDNA formatted file. You can go ahead and perform **Option 2**. In this example, I have manually uploaded an AncestryDNA file and pointed to the file path of the file for conversion. The 23andMe formatted file has *_converted* and added to the end.

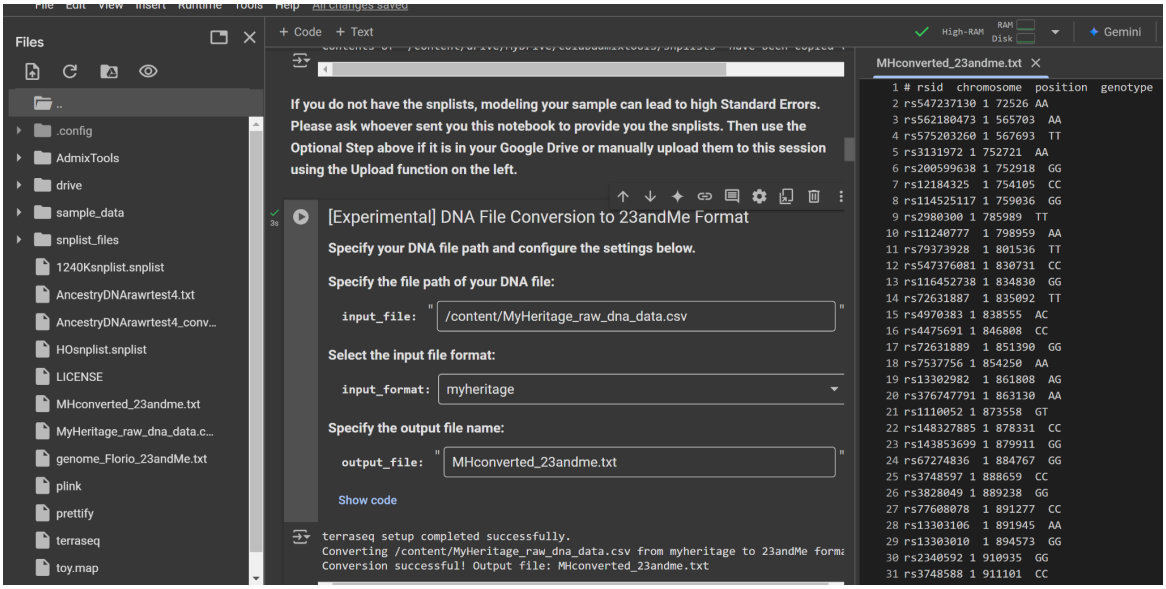



Let's skip **Option 3** for now. It will be useful later. Let's instead focus on uploading the 1240K and HO snplists into our Colab session. You can do so manually from your local device or using the **Optional Step** if they are in your *colabadmixtools* folder in your Google Drive. Do so manually if you do not wish to Mount your Drive account, they will go to the **/content/** folder. If using the

script, it will be saved within **/content/** as a new folder called **/snplist_files/**. Either way is fine, it is recommended you trim your raw file to the appropriate snplist for best performance.




For those with MyHeritage or FTDNA formatted files, I have an experimental script that utilizes **terraseq** (courtesy of X) that may work for you if DNA Kit Studio does not. However, I cannot guarantee it will always work as expected. In this example, I manually uploaded an MH .csv file to my working directory (*/content/*) and pointed to it. Selected myheritage as the input format and chose **MHconverted_23andme.txt** as my desired output name for it.

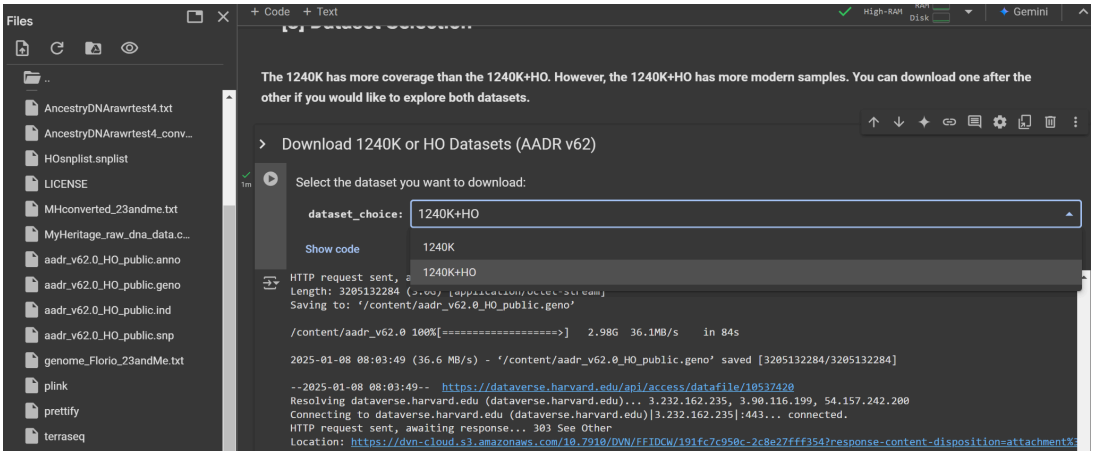



Alright, now we have our raw file in 23andMe format.

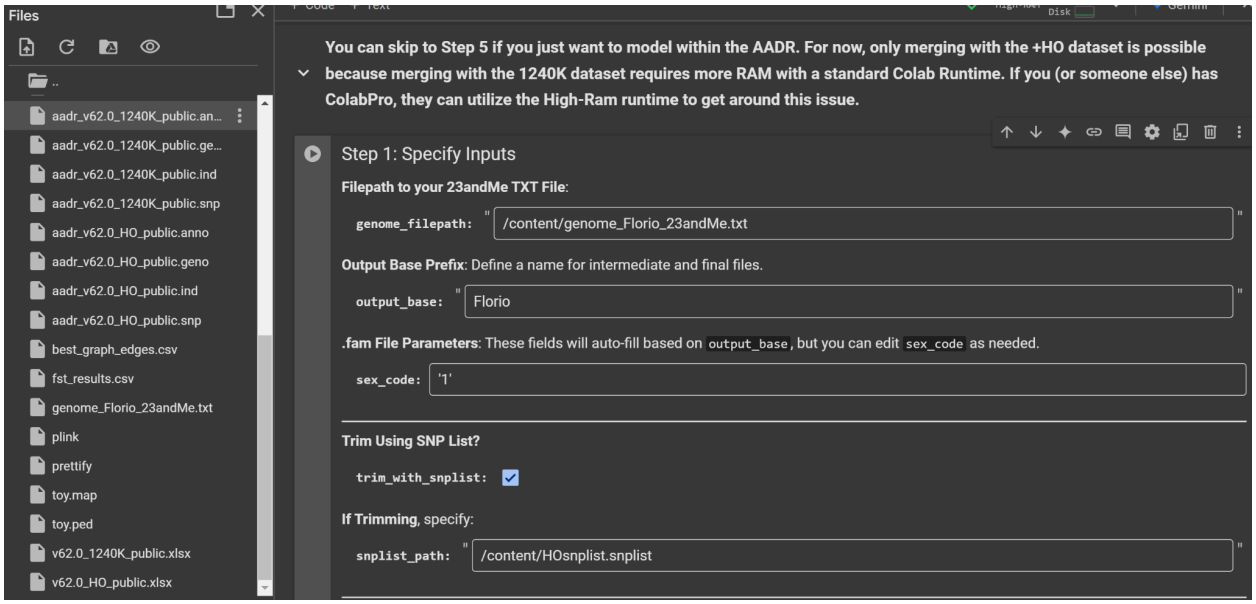
## [3] Dataset Selection

For this example, I will merge with the **1240+HO** dataset (or just **HO** for short). Unfortunately, in a free standard Colab Runtime, we have limited dedicated RAM so merging with 1240K is only possible with ColabPro HighRAM runtimes at the moment. If you trust someone to merge you

with the 1240K on their ColabPro runtime, I recommend removing all personal identifiers in your raw file before sending it over. Normally, I request the person to send me their raw file as a zipped file and I save it to my **colabadmixtools** folder in my Drive. Then I used **Step 2 Option 3**, point to that zipped file, and uncompress it into the Colab session. From there I perform all merging and use **Step 7** to compress their .geno, .ind, and .snp merged files into a single zipped folder in my Drive. From there I use an anonymous filesharing service like [file.io](file.io) to share it with them to upload to their Colab sessions and model themselves. Otherwise, you can pay for ColabPro to resolve RAM runtime issues.



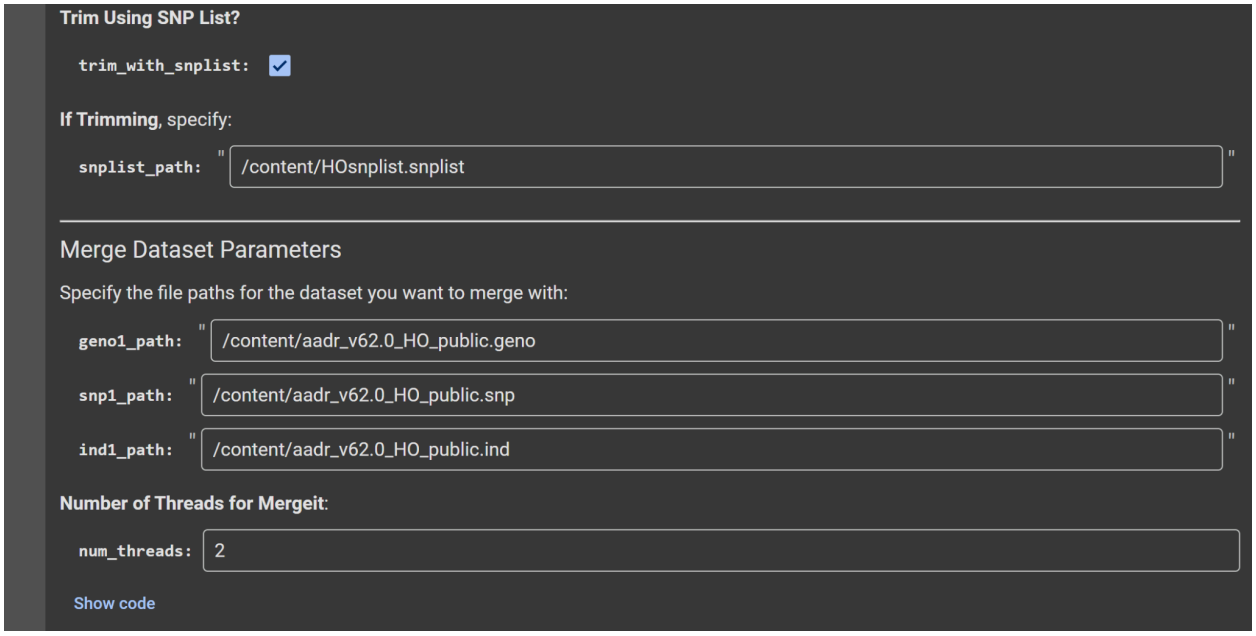

# [4] Merging Yourself with the AADR

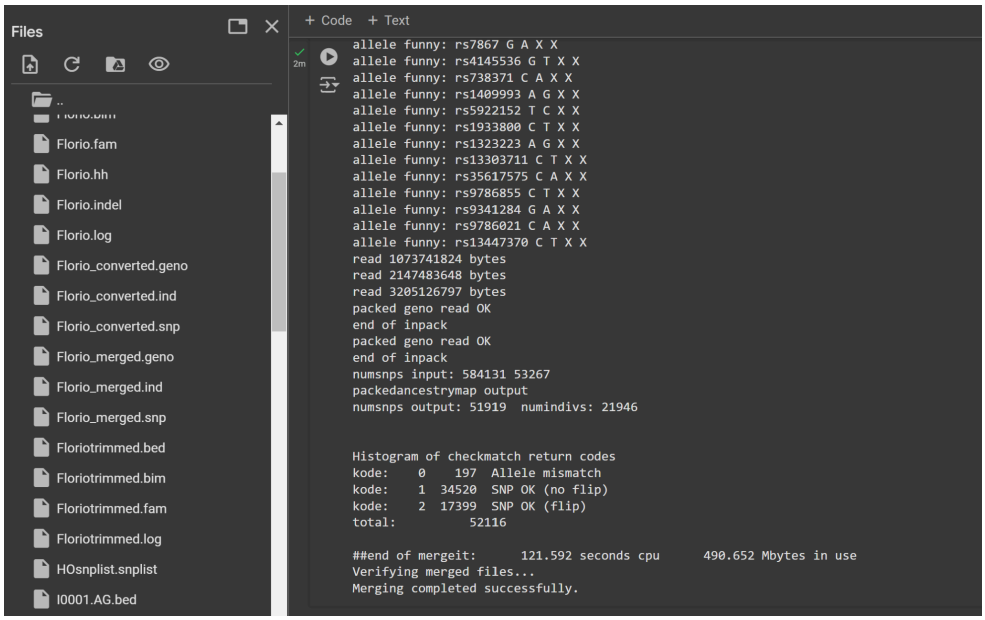


**Time:** Merging with HO about 5-8 minutes. Merging with 1240K about 6-16 minutes.

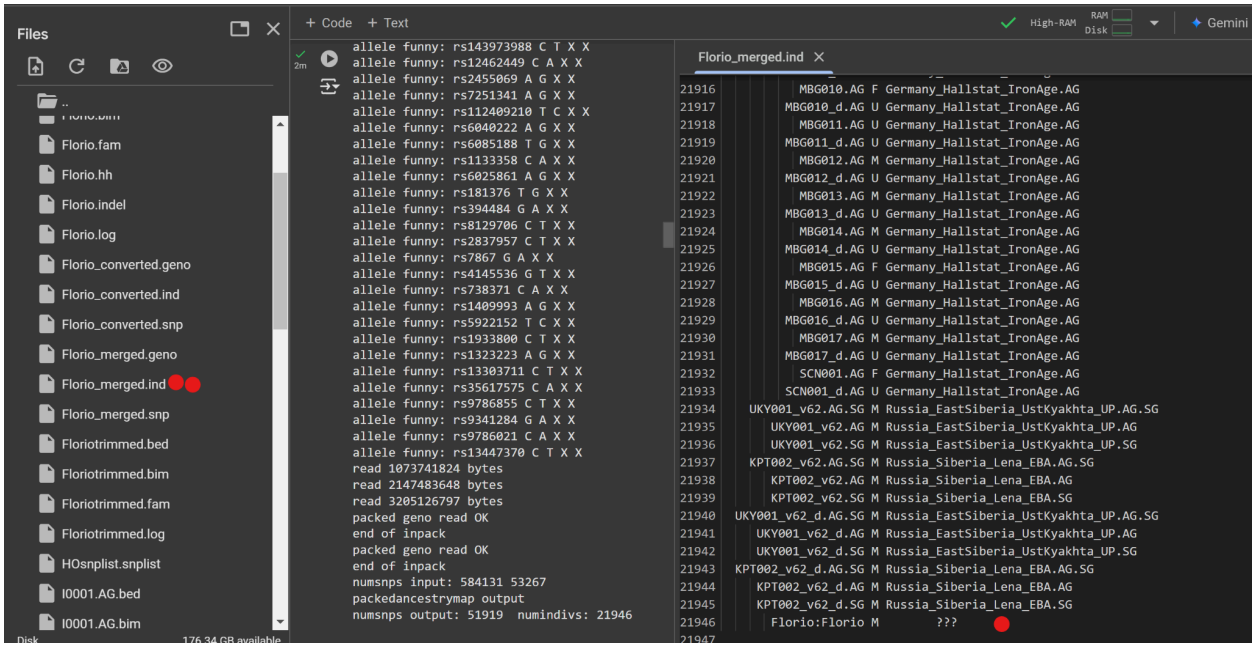
**Merging Overlap Reference:** AncestryDNA overlaps with 1240K: about 390K snps. With HO: 130K snps. MyHeritage overlaps with 1240K: about 177K snps. With HO: about 60K snps. 23andMe overlaps with 1240K: about 140K snps. With HO: about 50K snps.

Alright, here I have pointed to the filepath for my uploaded 23andMe formatted raw file. Remember to upload the correct snplist to trim to and point to the correct .geno, .snp, and .ind files for the dataset you intend to merge with. Sex code 1 for Male, and 2 for Female, 0 to omit. Leave the **num_threads** parameter alone for now.

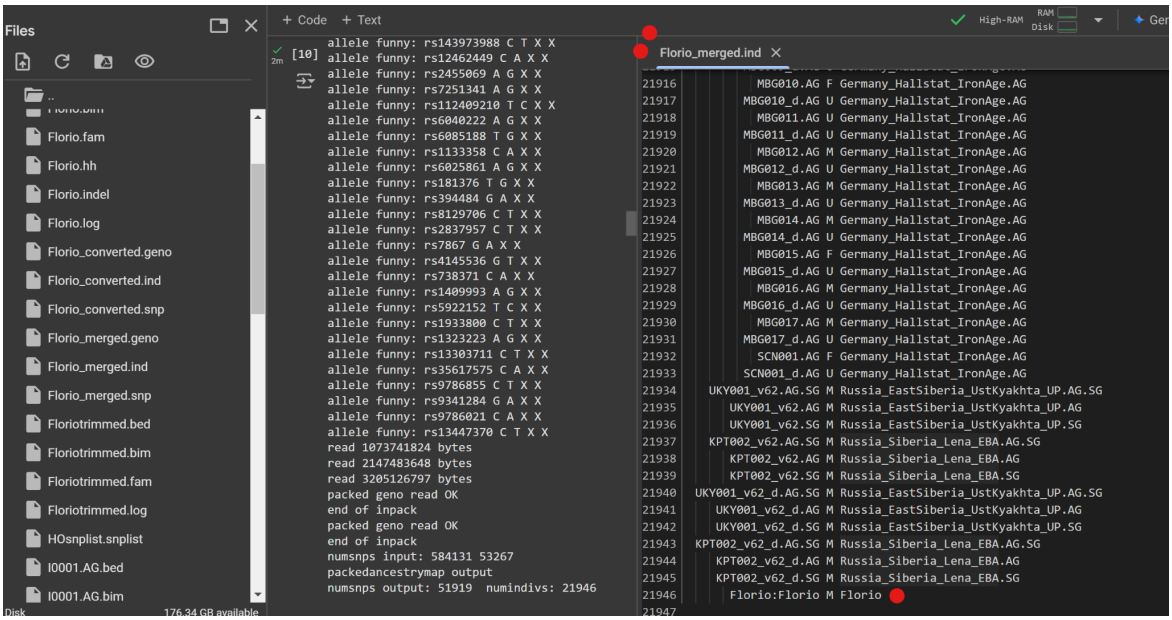





We can see the intermediate converted (23andMe to PLINK (PACKEDPED) format) files, the trimmed files (to the HO snplist), and the merged files (in PACKEDANCESTRYMAP format). The overlap of 51K snps is not great but you can improve this by merging with a MyHeritage and/or an AncestryDNA raw file.

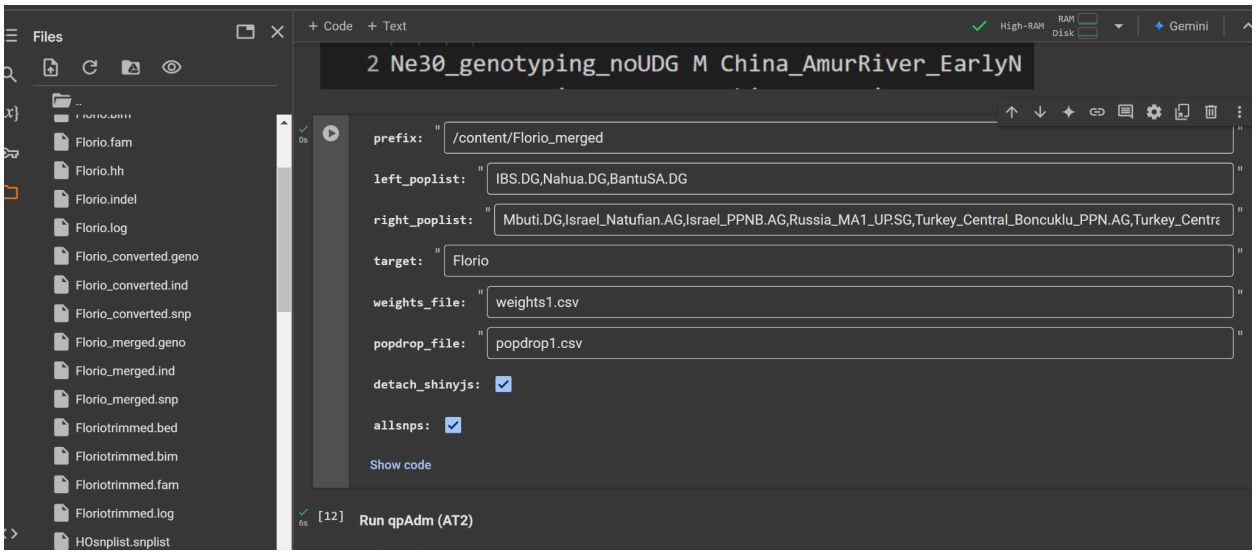



We should now open our *_merged .ind* file by double-clicking on it in the File Explorer to the left and then scrolling down where we will find our sample identifier. We can rename the **???** and backspace (leaving one space between your sex identifier) to our desired sample name.




Once we have done that you can either wait a few seconds for the .ind file name asterisk (meaning it is being edited) to disappear or you can double-click on the tab name to autosave it.

## [5] AT2 qpAdm Prep and Running



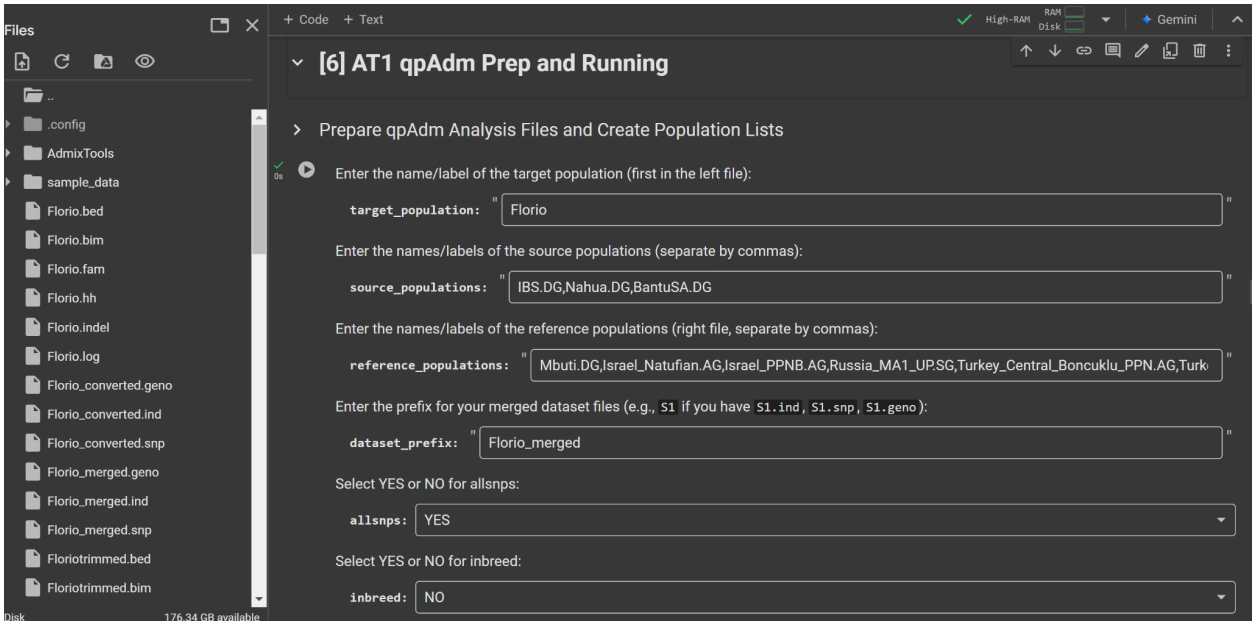

Now we can run a random model to see if it works out. Note that I am pointing to the filepath prefix to my merged dataset now. I am Salvadoran so these Left/Source pops I have entered will suffice but you may want to look through the *.xlsx* or *.anno* file to choose appropriate samples for you that will give you non-negative percentages. This time I am trying with this Right/Reference

list:
**Mbuti.DG,Israel_Natufian.AG,Israel_PPNB.AG,Russia_MA1_UP.SG,Turkey_Central_Boncuklu_PPN.AG,Turkey_Central_Pinarbasi_Epipaleolithic.AG,Morocco_Iberomaurusian.AG,Serbia_IronGates_Mesolithic.AG,Brazil_MG_C_LapaDoSanto_EH_HG_9600BP.AG.** Please read the Documentation above regarding Right/Reference population selection.

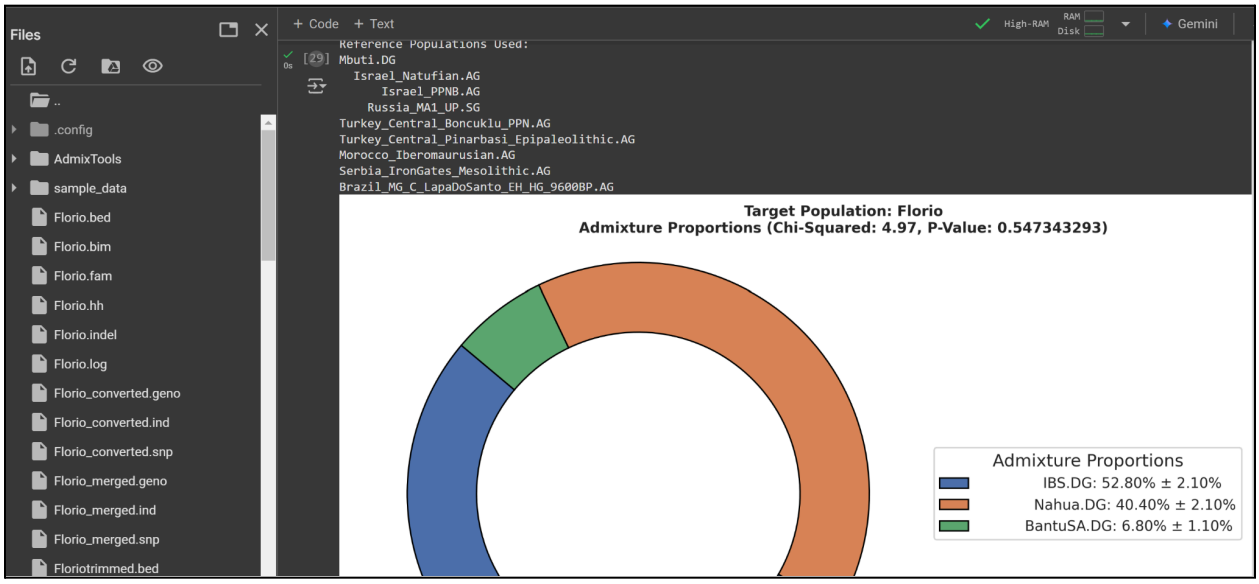
```
# A tibble: 3 x 5
  target left       weight     se     z
  <chr>  <chr>       <dbl>  <dbl> <dbl>
1 Florio IBS.DG     0.532  0.0217 24.5
2 Florio Nahua.DG   0.400  0.0210 19.1
3 Florio BantuSA.DG 0.0677 0.0113  5.98
# A tibble: 7 x 14
  pat      wt   dof   chisq         p f4rank IBS.DG Nahua.DG BantuSA.DG feasible
  <chr> <dbl> <dbl>   <dbl>     <dbl>  <dbl>  <dbl>    <dbl>      <dbl> <lgl>
1 000       0     6    5.17 5.22e-  1      2  0.532    0.400     0.0677 TRUE
2 001       1     7   67.0  6.08e- 12      1  0.571    0.429    NA      TRUE
3 010       1     7  695.   9.03e-146      1  0.897   NA         0.103  TRUE
4 100       1     7  808.   2.97e-170      1 NA        0.876     0.124  TRUE
5 011       2     8 1524.   4.94e-324      0  1       NA        NA      TRUE
6 101       2     8  943.   3.17e-198      0 NA        1        NA      TRUE
7 110       2     8 6645.   0              0 NA       NA         1      TRUE
```

For reference, I am ~55% European+Western Asian & North African, ~38% Indigenous American, ~6% Sub-Saharan African, and ~3% Unassigned on 23andMe. We could manually rotate and try different Right populations but for this instructive tutorial, this should suffice.

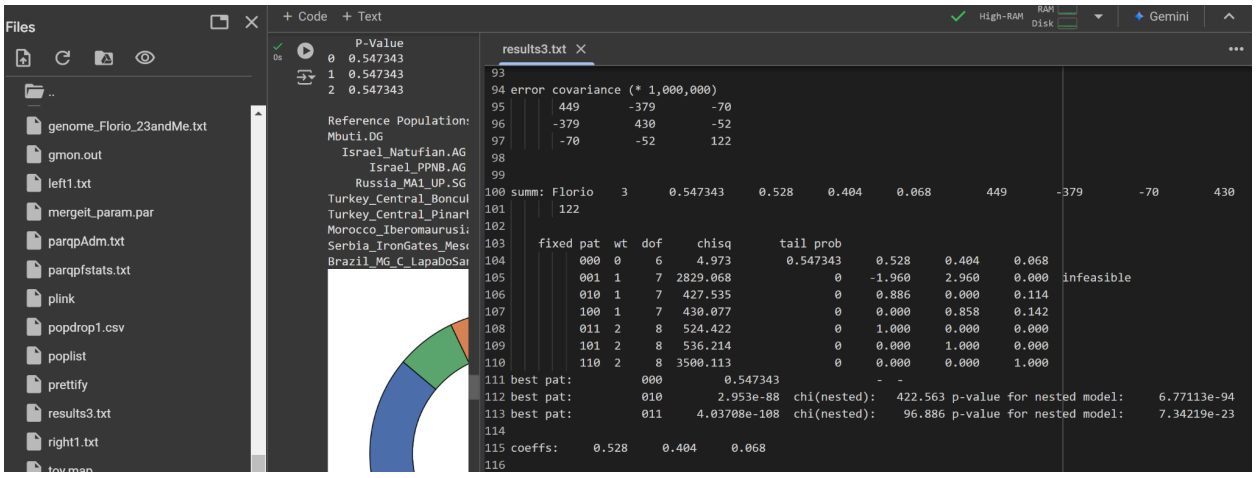
## [6] AT1 qpAdm Prep and Running



Now let's cross-reference with AT1 qpAdm and visualize those results. The parameters should be the same, you can leave allsnps and inbreed alone for now unless you know what they do. Moreover, notice that when I point to the dataset prefix in this code cell I am not including */content/* before it.

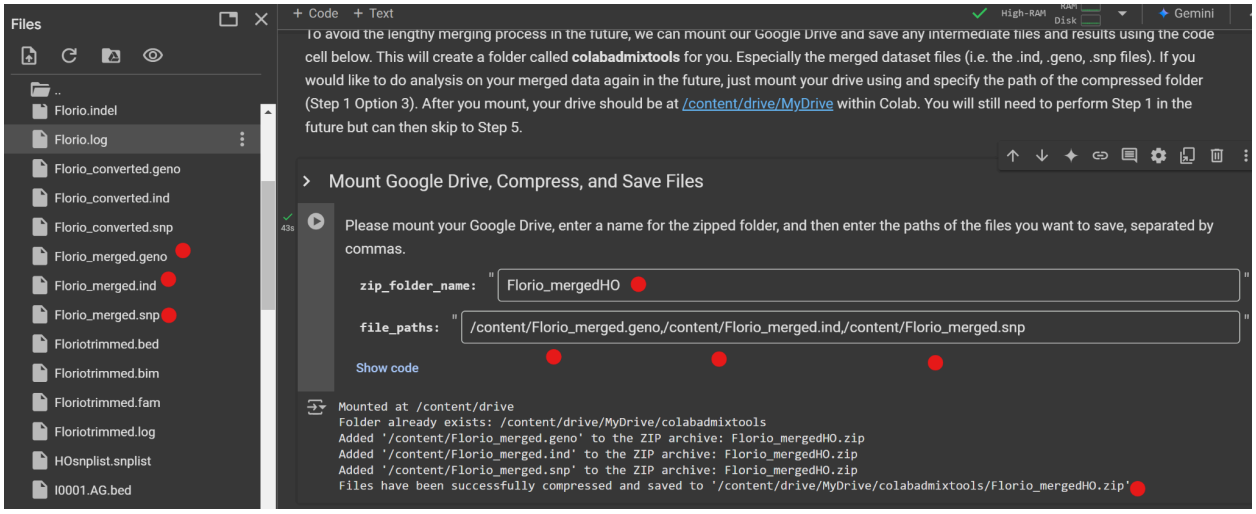



Everything seems in working order. We can save the image or Copy and Paste it somewhere else by right-clicking on it. Please read the Documentation above regarding model interpretation, especially the Harney et al. Supplementary document.

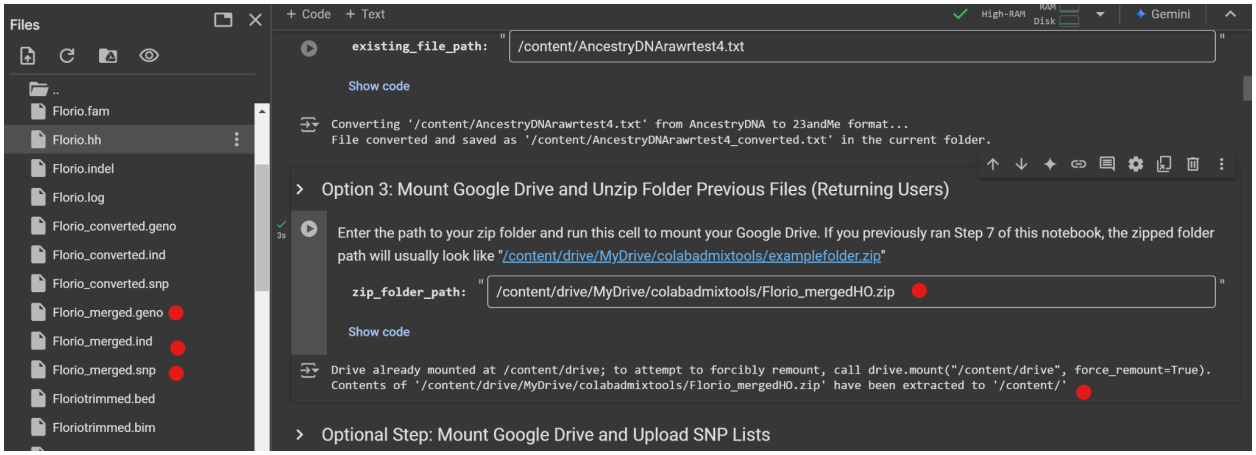



We can also open up the results output file and scroll to the summary table to get the popdrop table and diagnose if other combinations might be better. Feel free to rename the results file if you do not wish to overwrite it with subsequent runs.

# [7] Mounting Drive and Saving Compressed Files



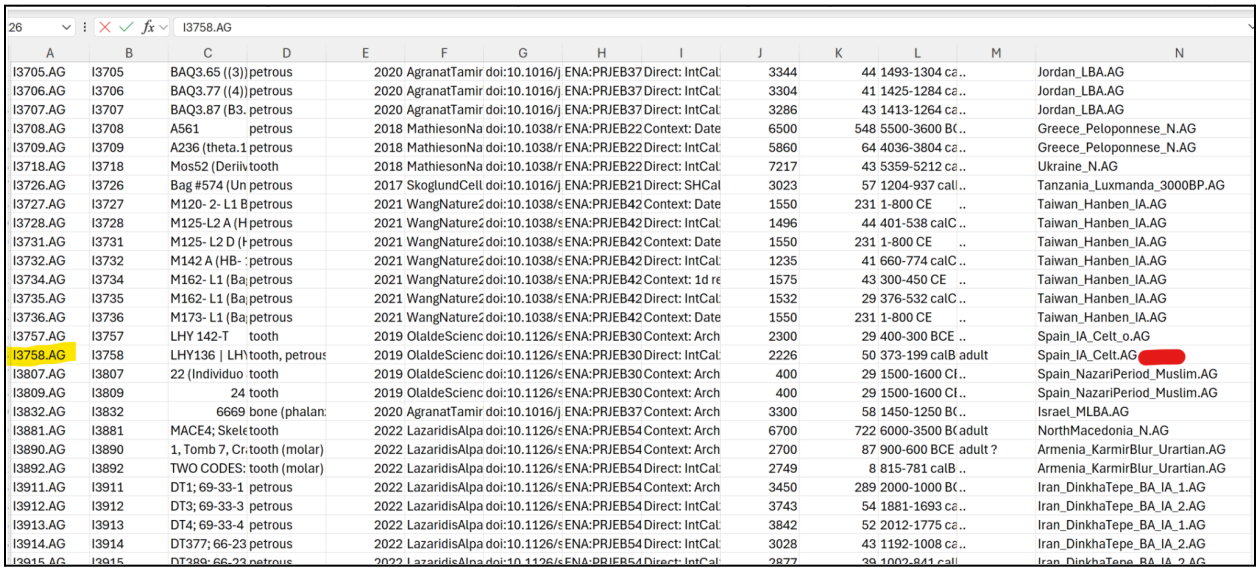

Now that we have merged ourselves, we can Mount our Drive and save our 3 dataset files into a compressed ZIP folder in a folder called **colabadmixtools**, which will be created for you if it does not already exist in your Drive. I have chosen to name it *Florio_mergedHO* as I merged with HO but you can name it however you'd like. You can right-click on the *.geno, .ind*, and *.snp* files to copy their filepaths and paste them (separated by commas) into the code cell form field before running.




This way when we close the browser or disconnect for whatever reason, we can open up this notebook again and enter the filepath to our compressed folder which should be in a similar path to the one shown in the image above (replacing *Florio_mergedHO* with whatever you had named it). From there we can do **Step 1** to get the software we need and go straight to modeling.

# Utilities

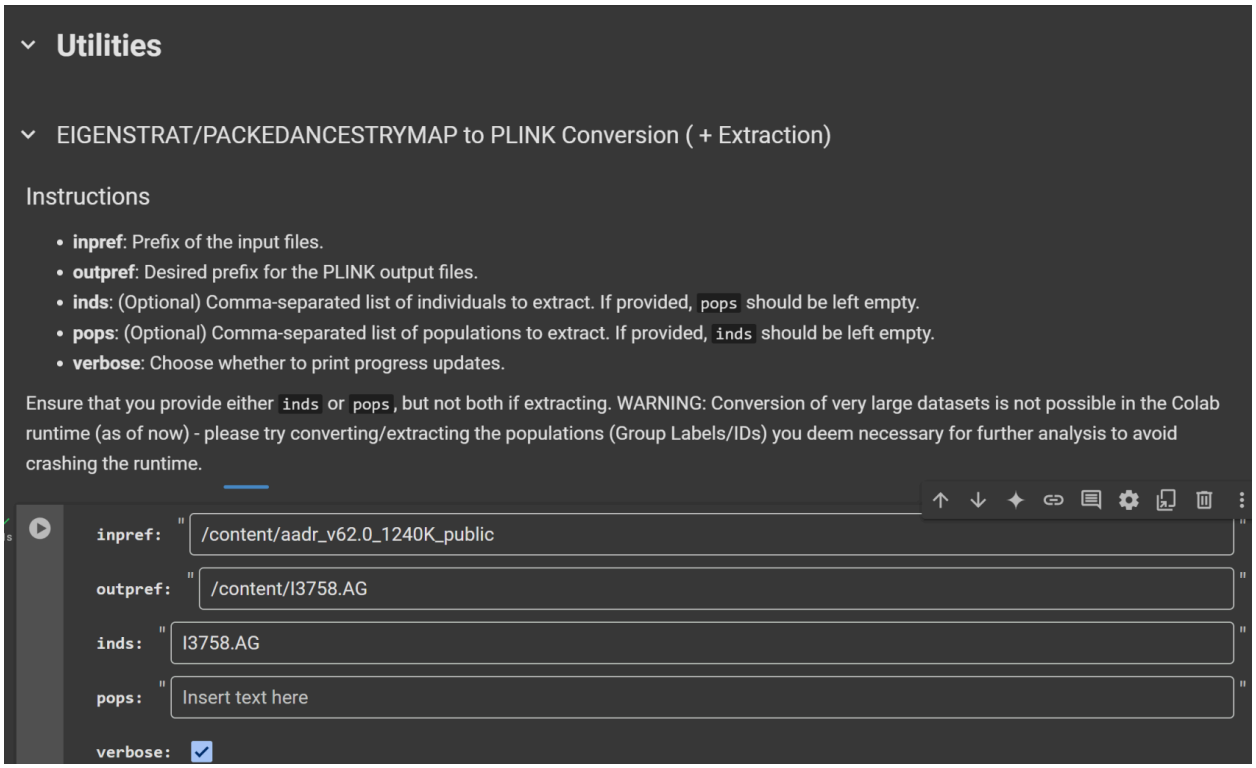
## Extracting a Sample and Converting to 23andMe Format



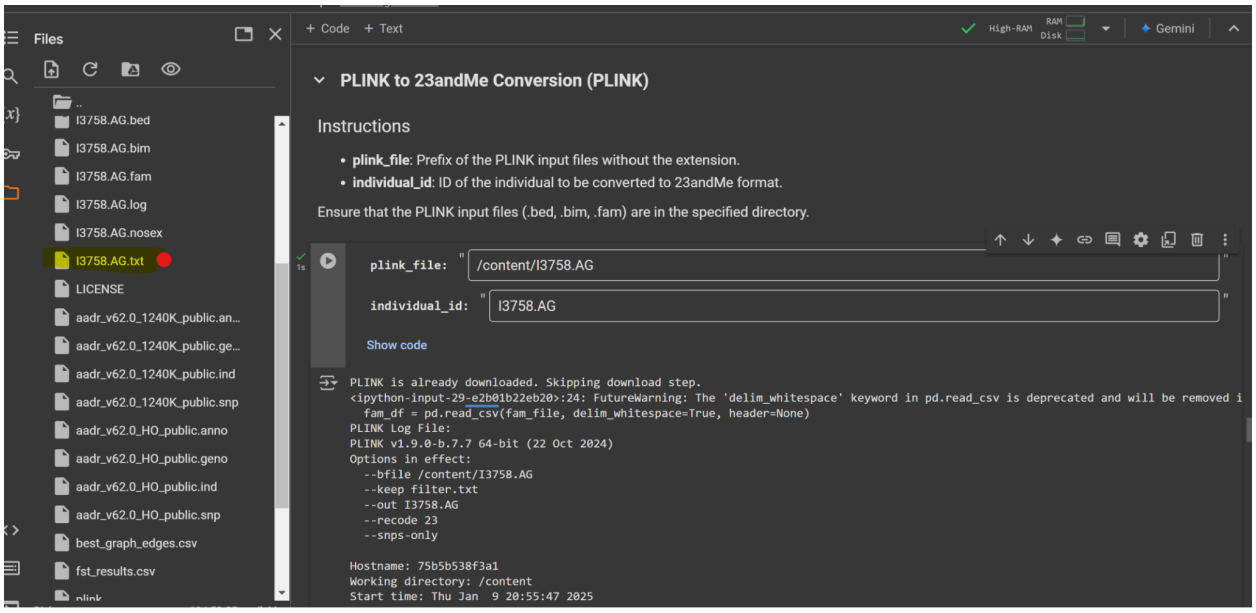
26 | I3758.AG

| A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I3705.AG | I3705 | BAQ3.65 ((3)) | petrous | 2020 | AgranatTamir | doi:10.1016/j | ENA:PRJEB37 | Direct: IntCal | 3344 | 44 | 1493-1304 ca | .. | Jordan_LBA.AG |
| I3706.AG | I3706 | BAQ3.77 ((4)) | petrous | 2020 | AgranatTamir | doi:10.1016/j | ENA:PRJEB37 | Direct: IntCal | 3304 | 41 | 1425-1284 ca | .. | Jordan_LBA.AG |
| I3707.AG | I3707 | BAQ3.87 (B3. | petrous | 2020 | AgranatTamir | doi:10.1016/j | ENA:PRJEB37 | Direct: IntCal | 3286 | 43 | 1413-1264 ca | .. | Jordan_LBA.AG |
| I3708.AG | I3708 | A561 | petrous | 2018 | MathiesonNa | doi:10.1038/r | ENA:PRJEB22 | Context: Date | 6500 | 548 | 5500-3600 B( | .. | Greece_Peloponnese_N.AG |
| I3709.AG | I3709 | A236 (theta.1 | petrous | 2018 | MathiesonNa | doi:10.1038/r | ENA:PRJEB22 | Direct: IntCal | 5860 | 64 | 4036-3804 ca | .. | Greece_Peloponnese_N.AG |
| I3718.AG | I3718 | Mos52 (Deriv | tooth | 2018 | MathiesonNa | doi:10.1038/r | ENA:PRJEB22 | Direct: IntCal | 7217 | 43 | 5359-5212 ca | .. | Ukraine_N.AG |
| I3726.AG | I3726 | Bag #574 (Un | petrous | 2017 | SkoglundCell | doi:10.1016/j | ENA:PRJEB21 | Direct: SHCal | 3023 | 57 | 1204-937 cal | .. | Tanzania_Luxmanda_3000BP.AG |
| I3727.AG | I3727 | M120- 2- L1 B | petrous | 2021 | WangNature2 | doi:10.1038/s | ENA:PRJEB42 | Context: Date | 1550 | 231 | 1-800 CE | .. | Taiwan_Hanben_IA.AG |
| I3728.AG | I3728 | M125-L2 A (H | petrous | 2021 | WangNature2 | doi:10.1038/s | ENA:PRJEB42 | Direct: IntCal | 1496 | 44 | 401-538 calC | .. | Taiwan_Hanben_IA.AG |
| I3731.AG | I3731 | M125- L2 D (H | petrous | 2021 | WangNature2 | doi:10.1038/s | ENA:PRJEB42 | Context: Date | 1550 | 231 | 1-800 CE | .. | Taiwan_Hanben_IA.AG |
| I3732.AG | I3732 | M142 A (HB- : | petrous | 2021 | WangNature2 | doi:10.1038/s | ENA:PRJEB42 | Direct: IntCal | 1235 | 41 | 660-774 calC | .. | Taiwan_Hanben_IA.AG |
| I3734.AG | I3734 | M162- L1 (Ba | petrous | 2021 | WangNature2 | doi:10.1038/s | ENA:PRJEB42 | Context: 1d re | 1575 | 43 | 300-450 CE | .. | Taiwan_Hanben_IA.AG |
| I3735.AG | I3735 | M162- L1 (Ba | petrous | 2021 | WangNature2 | doi:10.1038/s | ENA:PRJEB42 | Direct: IntCal | 1532 | 29 | 376-532 calC | .. | Taiwan_Hanben_IA.AG |
| I3736.AG | I3736 | M173- L1 (Ba | petrous | 2021 | WangNature2 | doi:10.1038/s | ENA:PRJEB42 | Context: Date | 1550 | 231 | 1-800 CE | .. | Taiwan_Hanben_IA.AG |
| I3757.AG | I3757 | LHY 142-T | tooth | 2019 | OlaldeScienc | doi:10.1126/s | ENA:PRJEB30 | Context: Arch | 2300 | 29 | 400-300 BCE | .. | Spain_IA_Celt_o.AG |
| I3758.AG | I3758 | LHY136 \| LH\ | tooth, petrous | 2019 | OlaldeScienc | doi:10.1126/s | ENA:PRJEB30 | Direct: IntCal | 2226 | 50 | 373-199 calB | adult | Spain_IA_Celt.AG |
| I3807.AG | I3807 | 22 (Individuo | tooth | 2019 | OlaldeScienc | doi:10.1126/s | ENA:PRJEB30 | Context: Arch | 400 | 29 | 1500-1600 CI | .. | Spain_NazariPeriod_Muslim.AG |
| I3809.AG | I3809 | 24 | tooth | 2019 | OlaldeScienc | doi:10.1126/s | ENA:PRJEB30 | Context: Arch | 400 | 29 | 1500-1600 CI | .. | Spain_NazariPeriod_Muslim.AG |
| I3832.AG | I3832 | 6669 | bone (phalan | 2020 | AgranatTamir | doi:10.1016/j | ENA:PRJEB37 | Context: Arch | 3300 | 58 | 1450-1250 B( | .. | Israel_MLBA.AG |
| I3881.AG | I3881 | MACE4; Skele | tooth | 2022 | LazaridisAlpa | doi:10.1126/s | ENA:PRJEB54 | Context: Arch | 6700 | 722 | 6000-3500 B( | adult | NorthMacedonia_N.AG |
| I3890.AG | I3890 | 1, Tomb 7, Cr | tooth (molar) | 2022 | LazaridisAlpa | doi:10.1126/s | ENA:PRJEB54 | Context: Arch | 2700 | 87 | 900-600 BCE | adult ? | Armenia_KarmirBlur_Urartian.AG |
| I3892.AG | I3892 | TWO CODES: | tooth (molar) | 2022 | LazaridisAlpa | doi:10.1126/s | ENA:PRJEB54 | Direct: IntCal | 2749 | 8 | 815-781 calB | .. | Armenia_KarmirBlur_Urartian.AG |
| I3911.AG | I3911 | DT1; 69-33-1 | petrous | 2022 | LazaridisAlpa | doi:10.1126/s | ENA:PRJEB54 | Context: Arch | 3450 | 289 | 2000-1000 B( | .. | Iran_DinkhaTepe_BA_IA_1.AG |
| I3912.AG | I3912 | DT3; 69-33-3 | petrous | 2022 | LazaridisAlpa | doi:10.1126/s | ENA:PRJEB54 | Direct: IntCal | 3743 | 54 | 1881-1693 ca | .. | Iran_DinkhaTepe_BA_IA_2.AG |
| I3913.AG | I3913 | DT4; 69-33-4 | petrous | 2022 | LazaridisAlpa | doi:10.1126/s | ENA:PRJEB54 | Direct: IntCal | 3842 | 52 | 2012-1775 ca | .. | Iran_DinkhaTepe_BA_IA_1.AG |
| I3914.AG | I3914 | DT377; 66-23 | petrous | 2022 | LazaridisAlpa | doi:10.1126/s | ENA:PRJEB54 | Direct: IntCal | 3028 | 43 | 1192-1008 ca | .. | Iran_DinkhaTepe_BA_IA_2.AG |
| I3915.AG | I3915 | DT389; 66-23 | petrous | 2022 | LazaridisAlpa | doi:10.1126/s | ENA:PRJEB54 | Direct: IntCal | 2877 | 39 | 1002-841 cal | .. | Iran_DinkhaTepe_BA_IA_2.AG |

When we download the HO or 1240K datasets in Colab, it is good practice to use Colab File Explorer to download a local copy of the *.xlsx* sheet locally on our device to search for samples more easily. The first column will be the **Sample ID** (or the individual ID) and a few columns down will be the **Group Label ID** (or the population ID). In this example, I want to extract a single individual sample from the 1240K AADR and convert it to PACKEDPED (PLINK format).

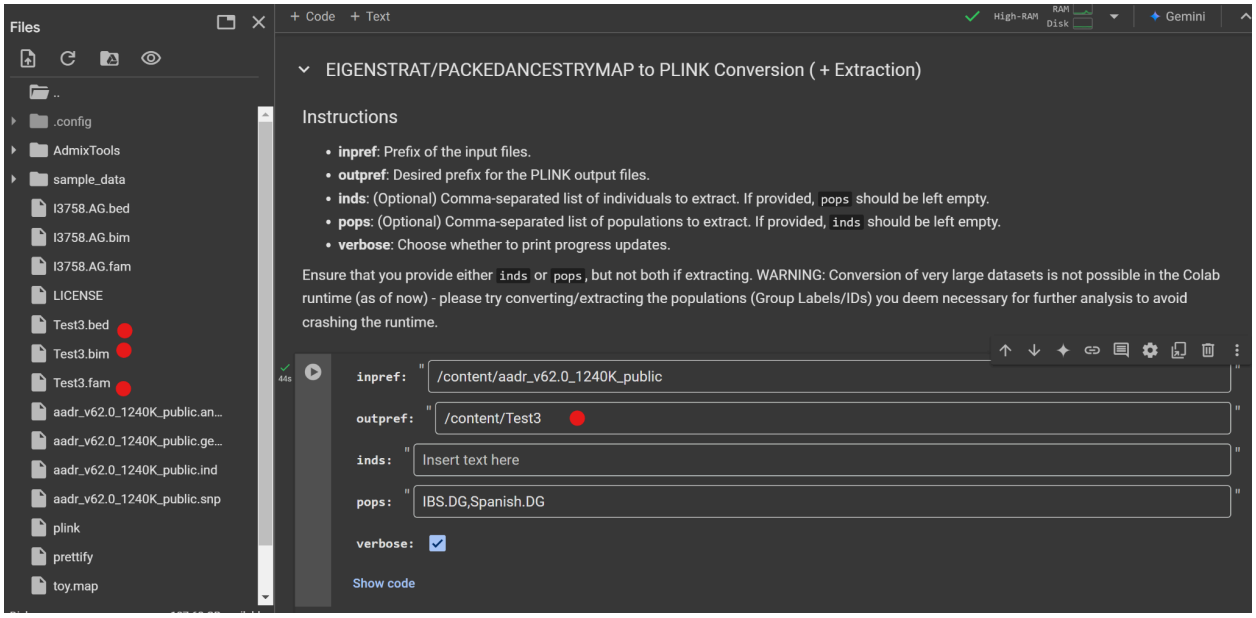



We point to the dataset prefix we wish to extract from. We enter the desired prefix/name for these PLINK files (include */content/* as this is the destination). We enter the individual/sample ID according to the xlsx sheet. We can also extract multiple samples at once, just separate them by a single comma. However, extracting too many at once can lead to a RAM runtime error. Alternatively, we can also extract group populations entirely (made up of multiple samples). Note that you can only extract individuals or groups at a time, not both at the same time. So for this example, I am extracting one individual sample and leaving pops blank.



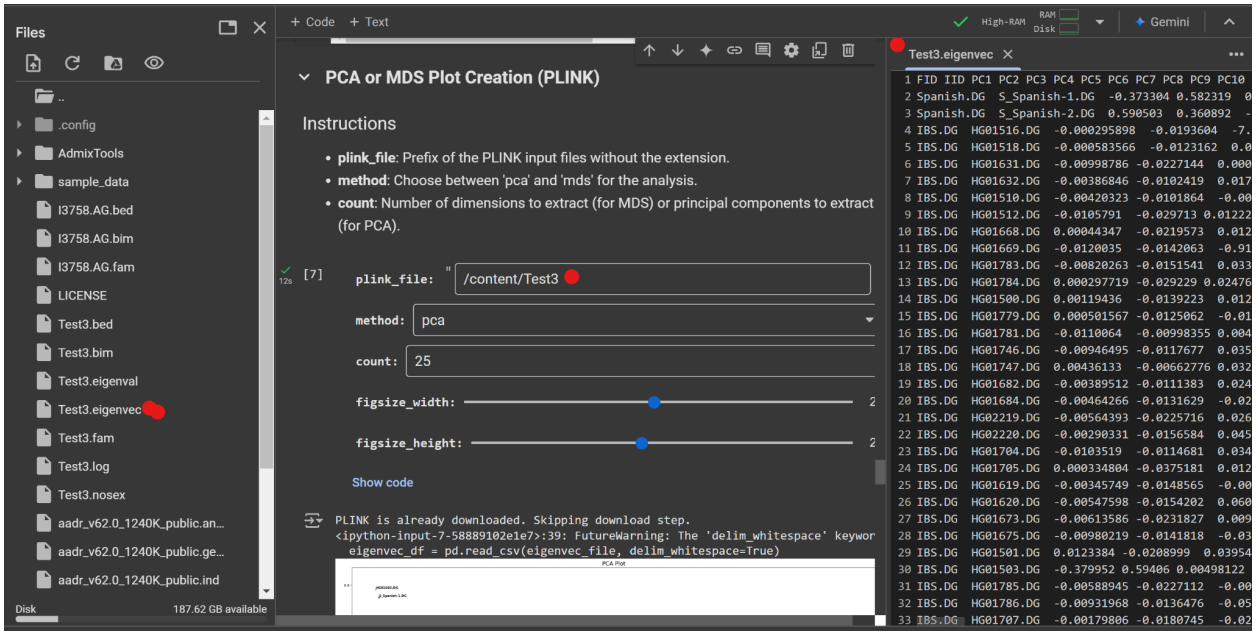

After playing that code cell we can see that the sample was converted successfully and a *.txt* file should appear in the Colab File Explorer which we can right-click and Download locally to use.

## Extracting Population Groups and Plotting PCA

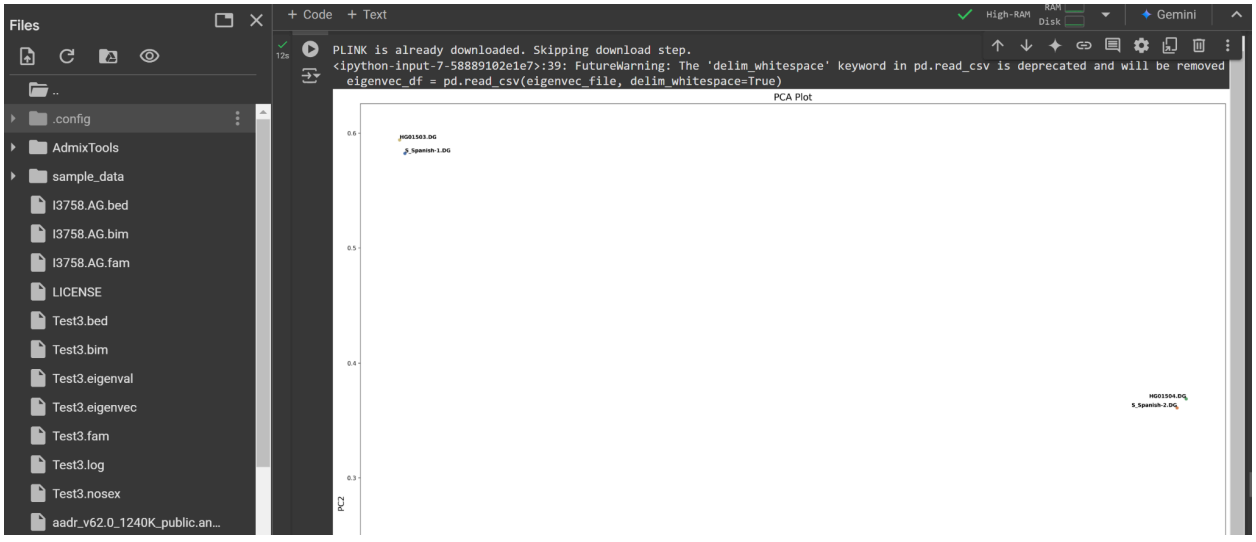



We can use the same code cell we used before to extract multiple samples using Group/Population Label IDs using the xlsx sheets. In this example, I will be extracting the 102

IBS.DG samples along with the 2 Spanish.DG samples from the 1240K AADR and performing a PCA on all of them. I will name this PLINK dataset *Test3* to not overwrite my previous PLINK files.



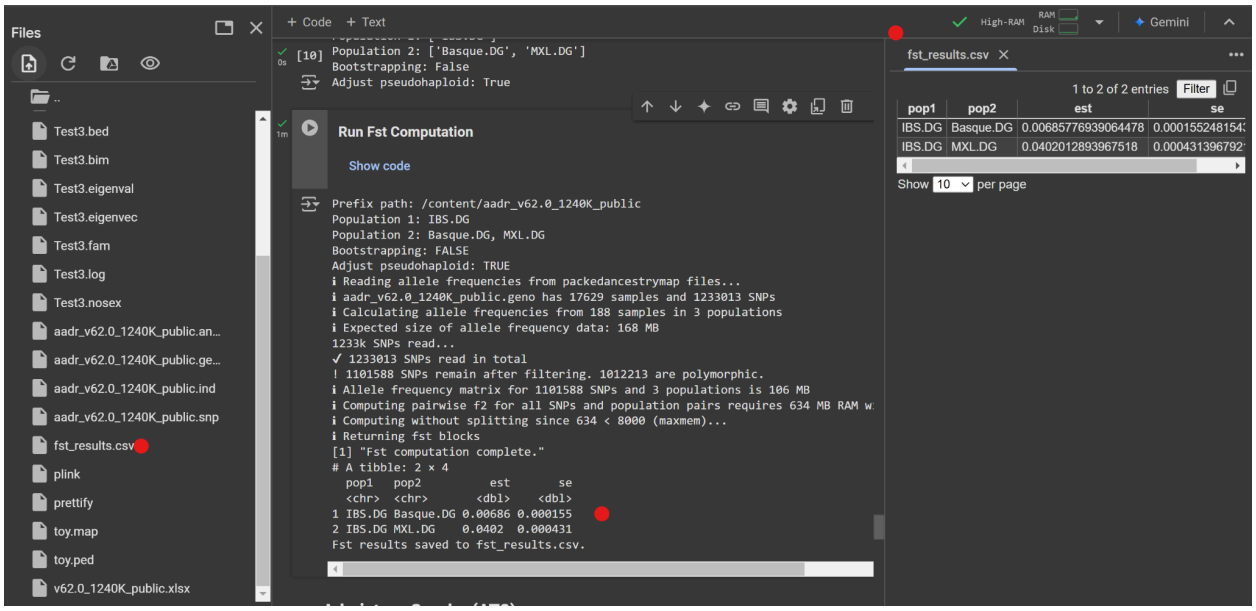

Now we can scroll down to this code cell and point to the filepath prefix for our new dataset. We can choose the PCA method and instruct it to analyze using 25 principal components from the data. We can adjust plot dimensions if the initial graph is difficult to visualize. Moreover, once we run this code cell we will also get an *.eigenvec* file which we can open by double-clicking on it in the Colab File Explorer and then seeing these values for each sample in a separate window.




Sometimes the sample markers overlap heavily, so we can adjust the plot dimensions and re-run the code cell to improve visibility. Alternatively, you can plot yourself using the coordinates.

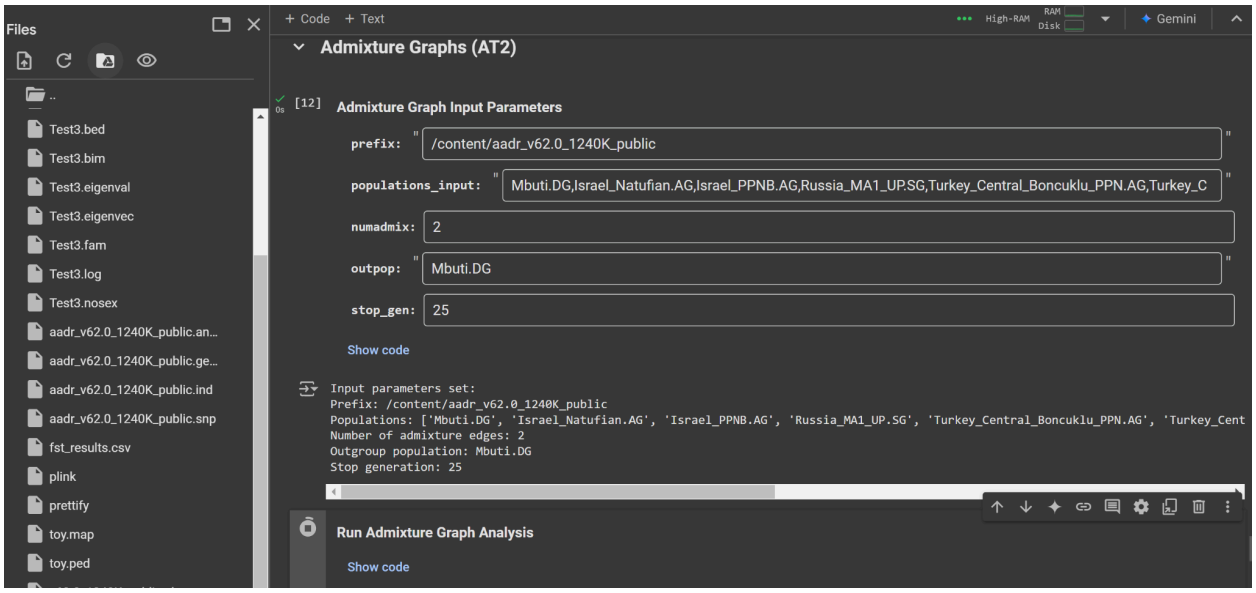
# Computing Fst


In this example, we can point to the 1240K AADR dataset and compute Fst (read **this** and **this** for details). I want to estimate the population differentiation between IBS.DG and two other groups (comma-separated). We can leave **bootstrapping** off for now (sometimes it throws errors). We can leave **adjust_pseudohaploid** on for now. If you are estimating Fst for too many populations it can result in a RAM runtime error, refer to General Troubleshooting below for advice. We can always split it Fst estimates between pops and combine the outputs later if RAM is an issue.



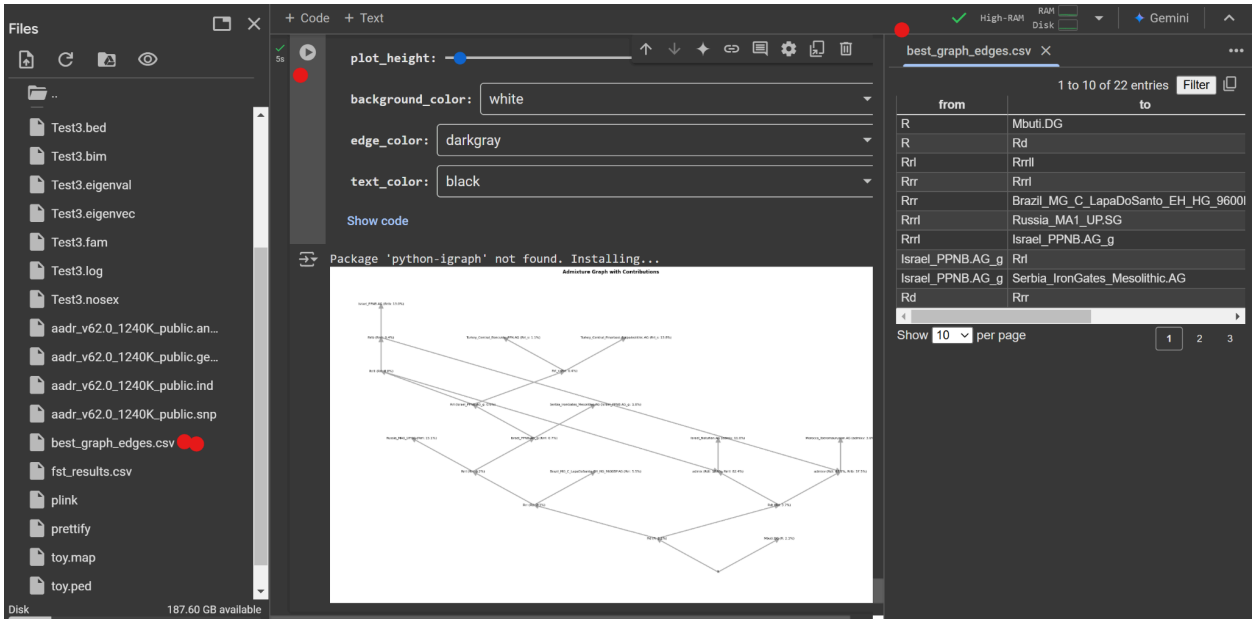

Fst is based on allele frequencies so it will be highly sensitive to the amount of coverage. We can open and Download the results of this Fst run using File Explorer.

## Plotting Admixture Graphs




Refer [here] for usage. Admixture graphs work great for predicting relationships and contributions of admixture between populations. Personally, I do not use this feature but perhaps someone else would so I have included it in the notebook. In this example, we point to our dataset prefix filepath and I pasted a random Right list to make a random graph for illustrative purposes.

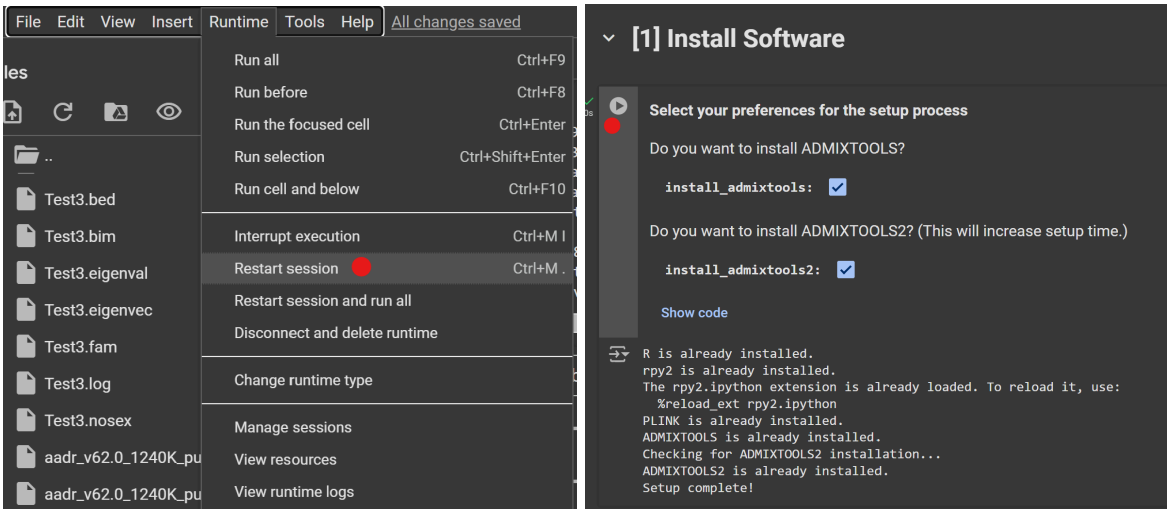
Choose an appropriate (or best guess) of admixture events. Here I chose 2. Choose an appropriate outgroup (here I chose Mbuti.DG). The **stop_gen** parameter you can set higher to simulate graphs and improve the score (however it will take more time).




We run the first code cell to establish the parameters then the next code cell to Run the graph simulations and ranking. The next code cell you can use to adjust the plot. We can also open the *.csv* file to see contributions and what ghost populations were created. The code cell will create an inverted tree of your best-ranked admixture graph. See Documentation for theory, usage, and interpretation.

## General Troubleshooting

Runtime Crash due to RAM Limit:




Sometimes the analysis you are doing causes a runtime error due to limited RAM on a free standard Colab runtime. In this case, a crash can cause some tools to be unloaded, and you can **Restart the session** and perform **Step 1** again to load those important packages.

For most non-runtime related errors, it will usually be due to a typo in the prefix filepath or a sample name. Sometimes unchecking (or checking) additional parameters can fix the issue; read the Documentation to see what scenarios might apply to your particular combination of samples (e.g. using pseudohaploid samples).

## Acknowledgements

ARA, Hands, Luka, and X for major contributions to the development of the notebook.

Sam, Ape, regh, qebap, Herman, Builder, rodolfo, and Jorge for testing, suggestions, and sample data.