**Readme**

Before running the Diversity Browser notebook, a few programs need to be installed.

**Install ANACONDA**

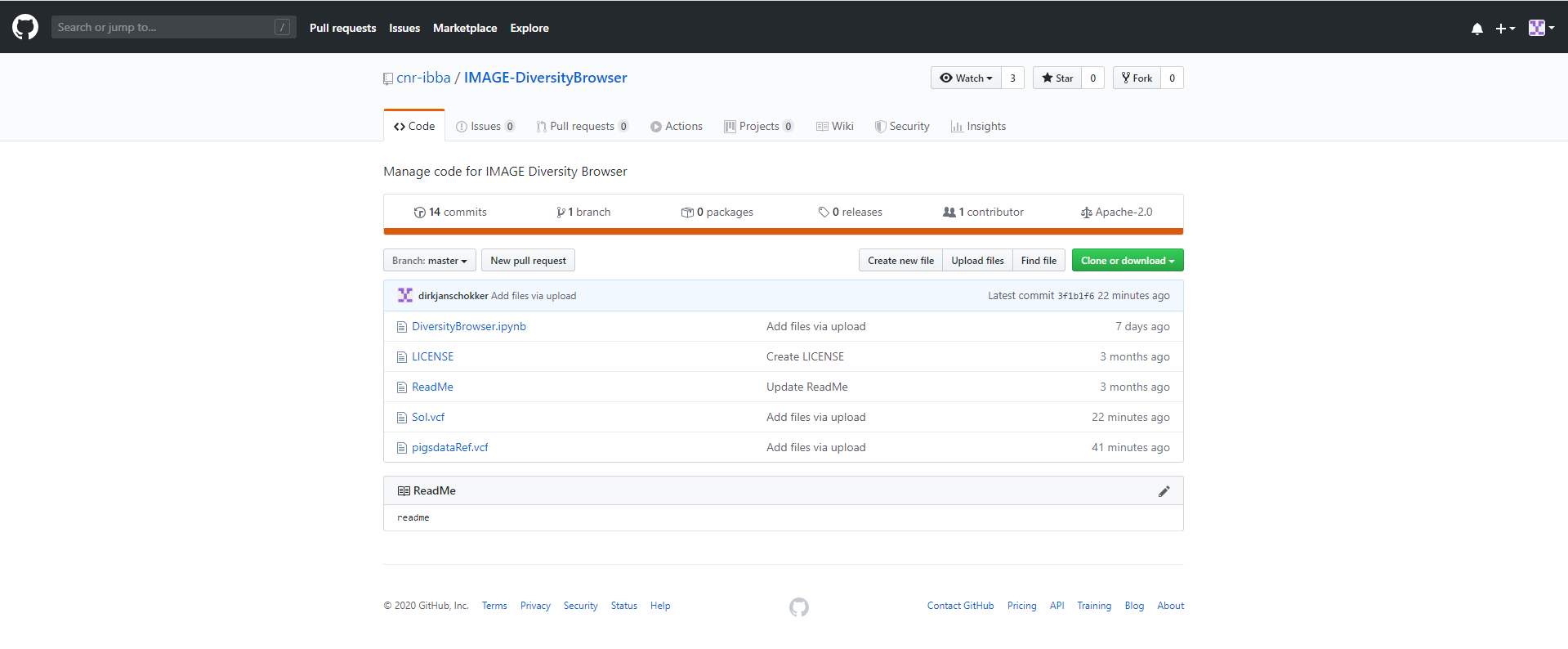
Please follow your platform documentation: <https://www.anaconda.com/distribution/>

**Install PLINK**

Please follow your platform documentation: <https://www.cog-genomics.org/plink/>

If both programs are correctly set-up and installed. The reference and demonstration dataset can be downloaded. Below both datasets are described in more detail and where to find them.

The pig reference dataset can be downloaded at: <https://github.com/cnr-ibba/IMAGE-DiversityBrowser>.



This dataset, IMAGE001\_23\_01\_BPW\_PIG.vcf, contains 149 pigs that were genotyped with the IMAGE ‘*multispecies SNP-chip*’. The ‘*multispecies SNP-chip*’ was specifically designed to handle multiple species, in other words the chip contains around 10,000 SNPs per species, including pigs, cattle, chicken, goat, and sheep.

Please make sure that you save the reference dataset in the folder, where you also saved your samples of interest, i.e. your own animals of interest typed by the ‘*multispecies SNP-chip*’. If you do not have such a file yet, you are free to use the demonstration dataset (SoI.vcf).

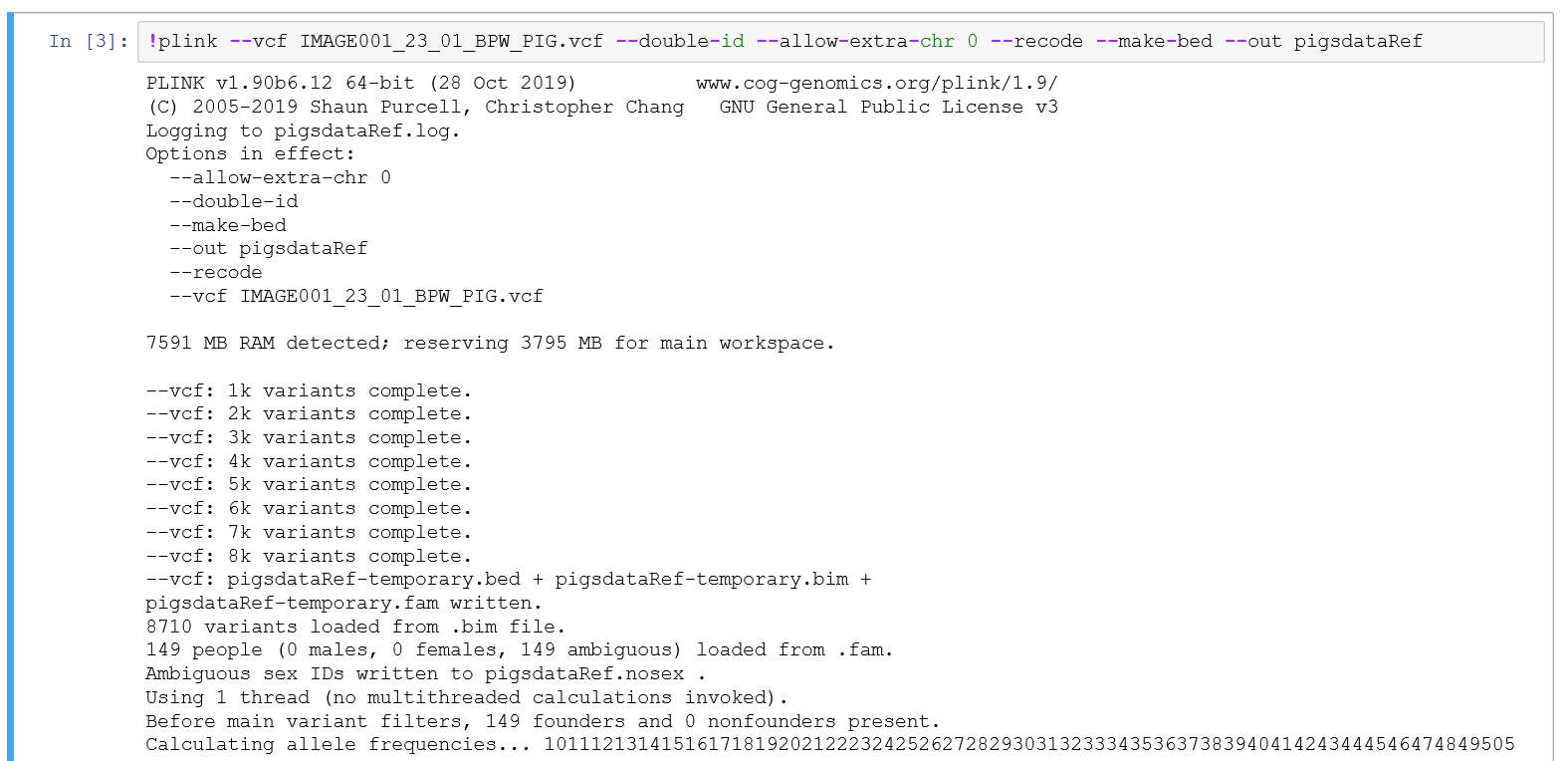
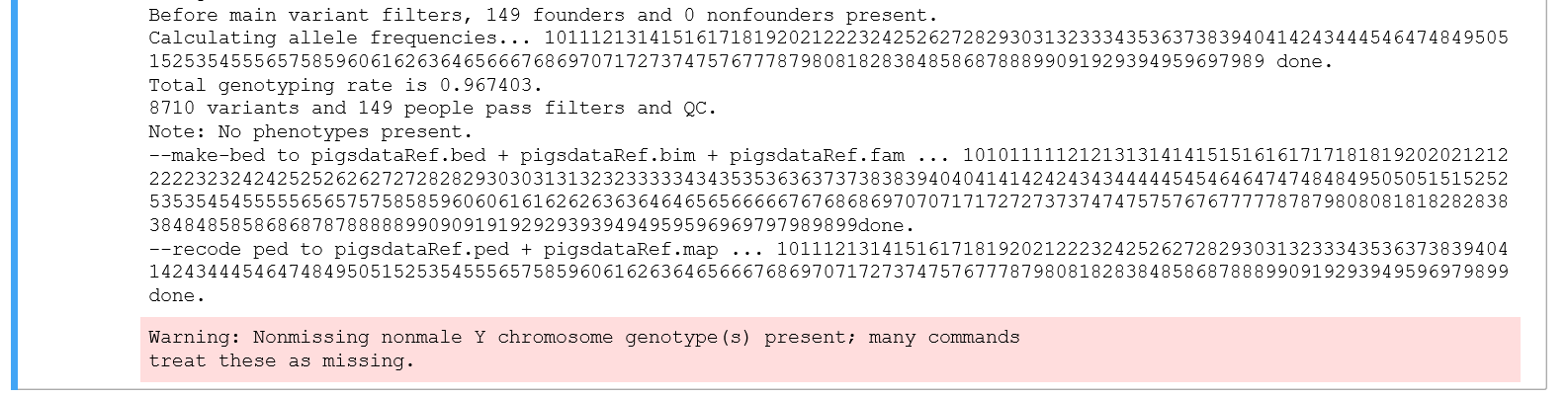
In the example, in our configuration we have set the directory to the following, see figure below.



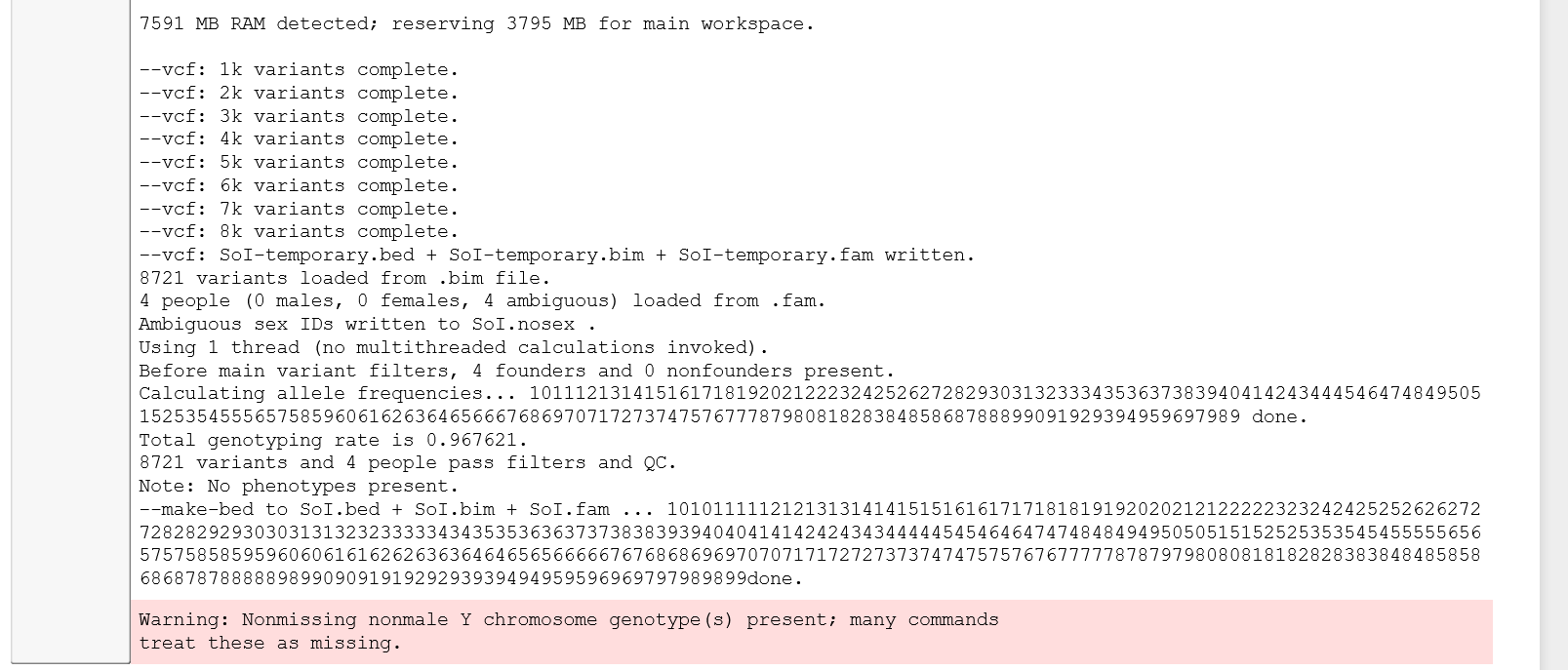
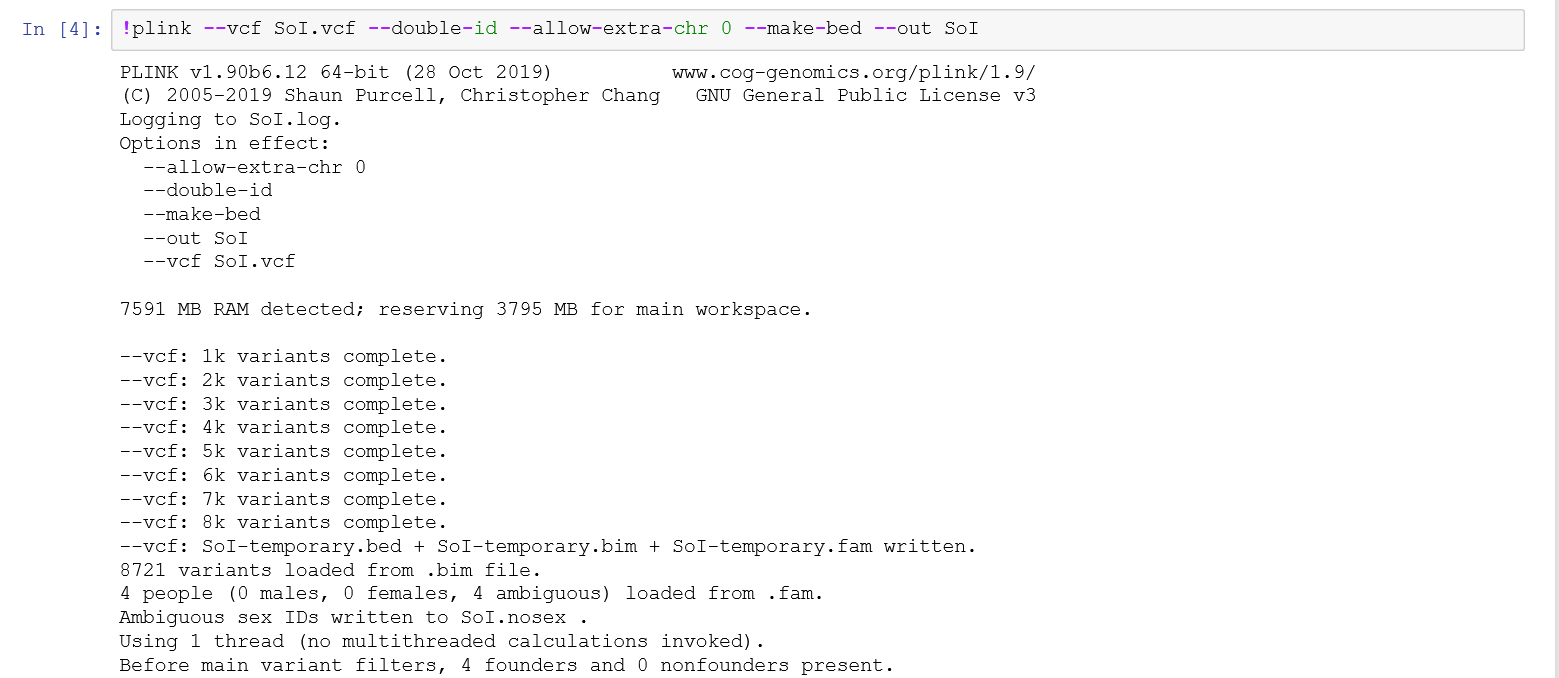
The next step is to run PLINK to create the reference dataset, by the code:

*!plink --vcf IMAGE001\_23\_01\_BPW\_PIG.vcf --double-id --allow-extra-chr 0 --recode --make-bed --out pigsdataRef*

Below also the output that is generated.



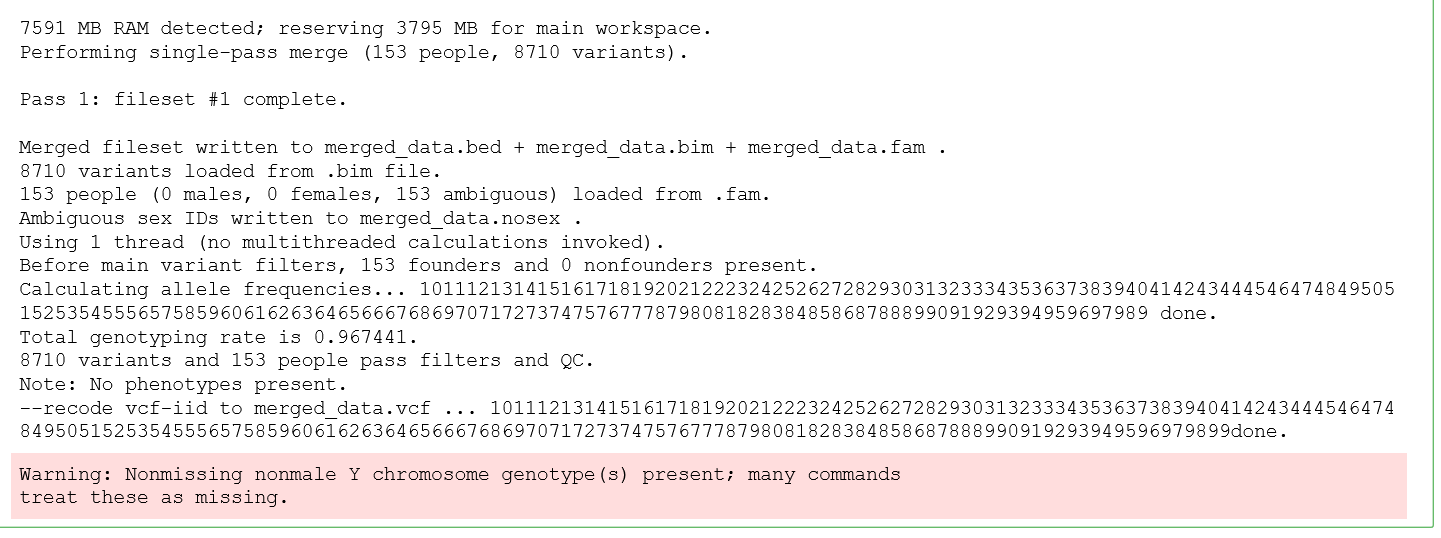
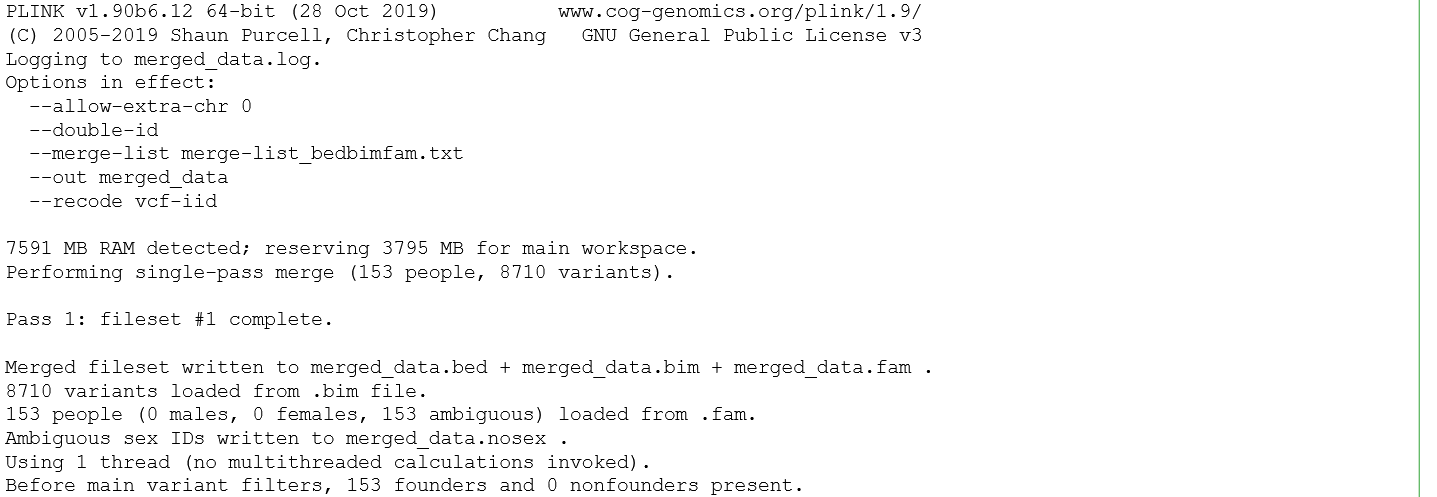
This code will have generated the file pigsDataref.vcf in your folder. The next step is to generate a similar file for the samples/dataset of interest. This is done be the next line of code, see below the code and output generated by the Notebook.



When both reference and sample of interest files are generated, the merging needs to happen. This is done by the following code still within PLINK:

*!plink --merge-list merge-list\_bedbimfam.txt --double-id --allow-extra-chr 0 --recode vcf-iid --out merged\_data*

This will generate the following output, see below a screenshot, as well as the file *merged\_data.vcf*. The latter file will serve as input for the Principal Component Analysis (PCA).



Within PLINK it is possible to employ a PCA (code below), here we have set the number of principal components to 25, and generate a new output file named *PCA\_pigsdataRef*.

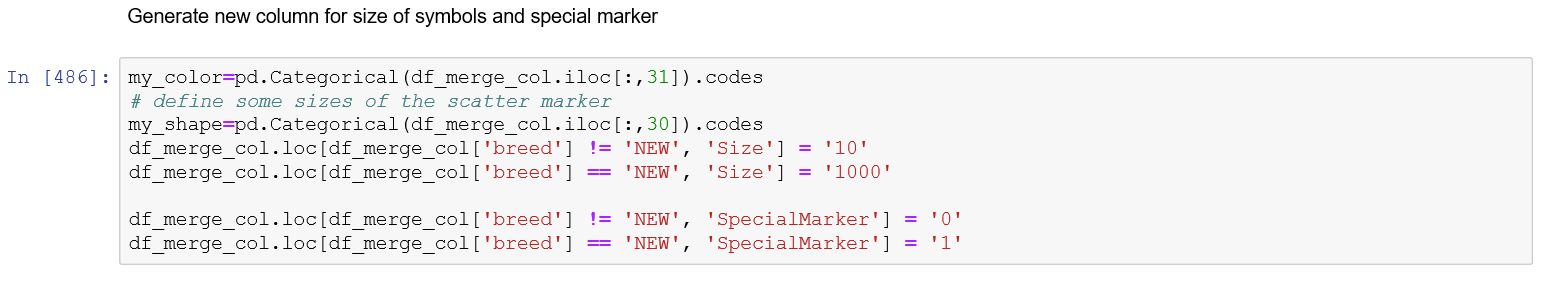
*!plink --vcf merged\_data.vcf --double-id --allow-extra-chr 0 --pca 25 --out PCA\_pigsdataRef*

The output file contains the eigenvectors needed to create the ‘easy-to-interpret’ scatterplot figure.

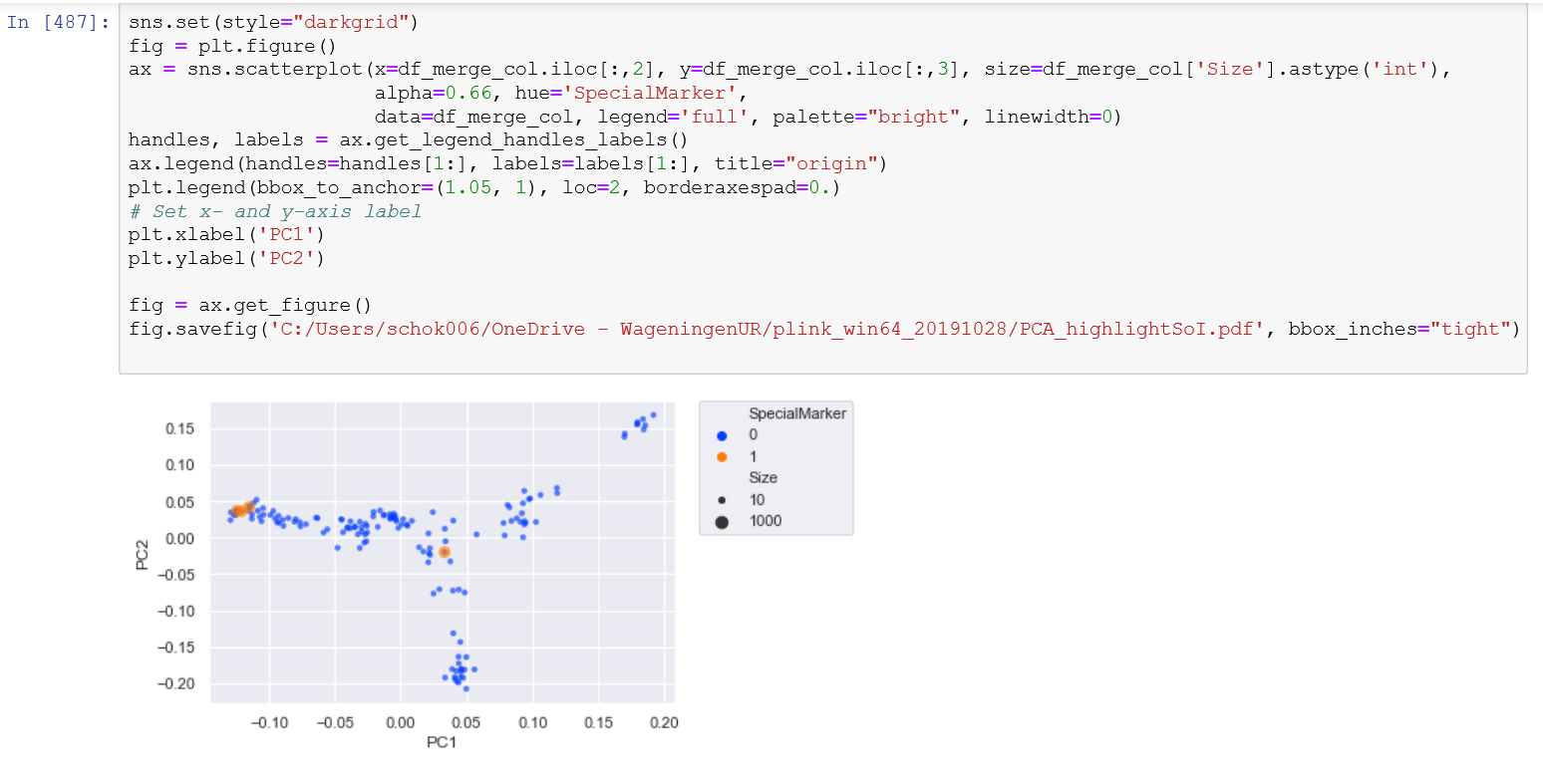
We read the comma-separated file, transform columns to generate an identical format of the PCA-output file, as the phenotype-file. The latter contains the metadata of each sample, here the samples are individual pigs and the metadata contain origin and breed name. This transformation of column names is necessary, for the merge of the PCA-output file and the phenotype-file. Below the screenshot of the Notebook with the actual code.

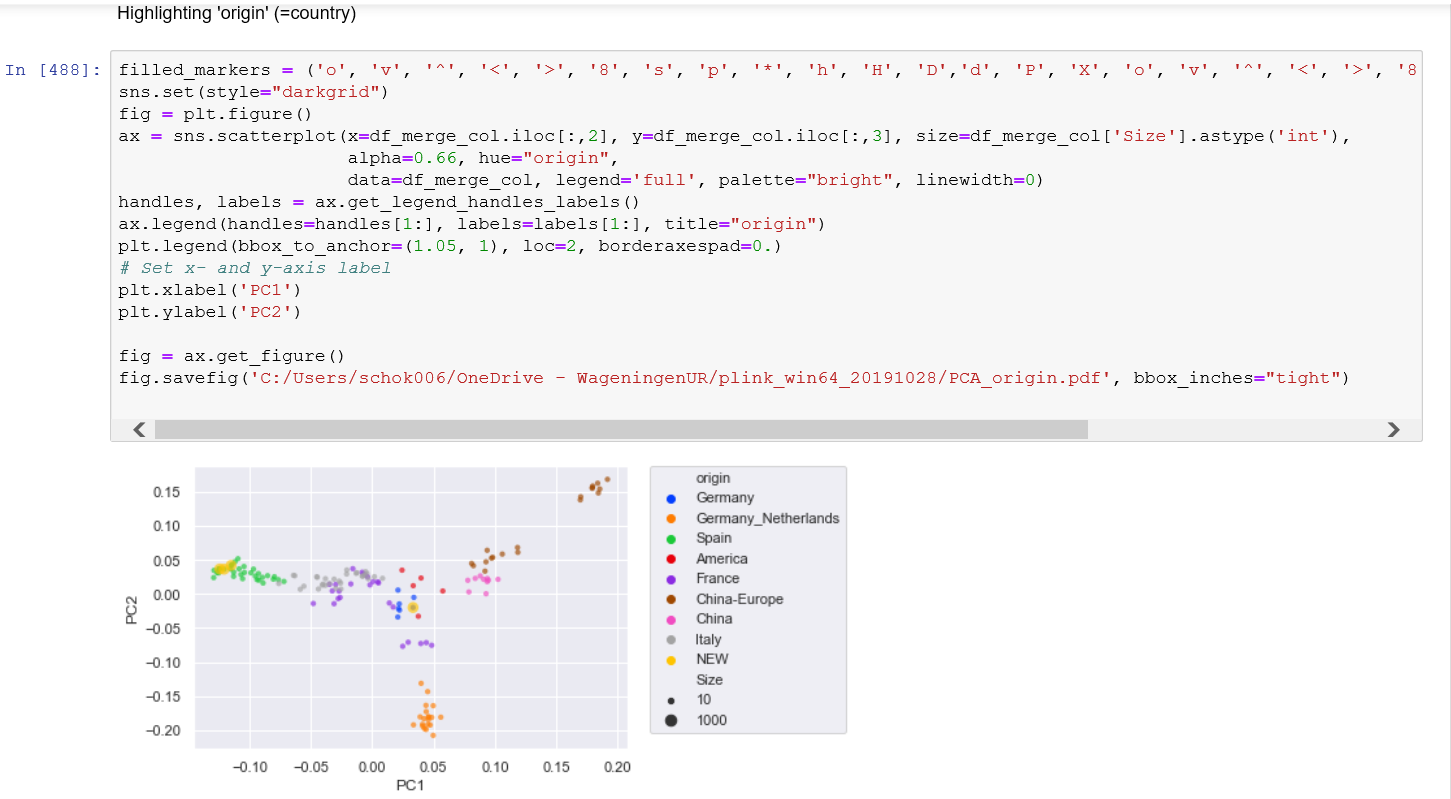


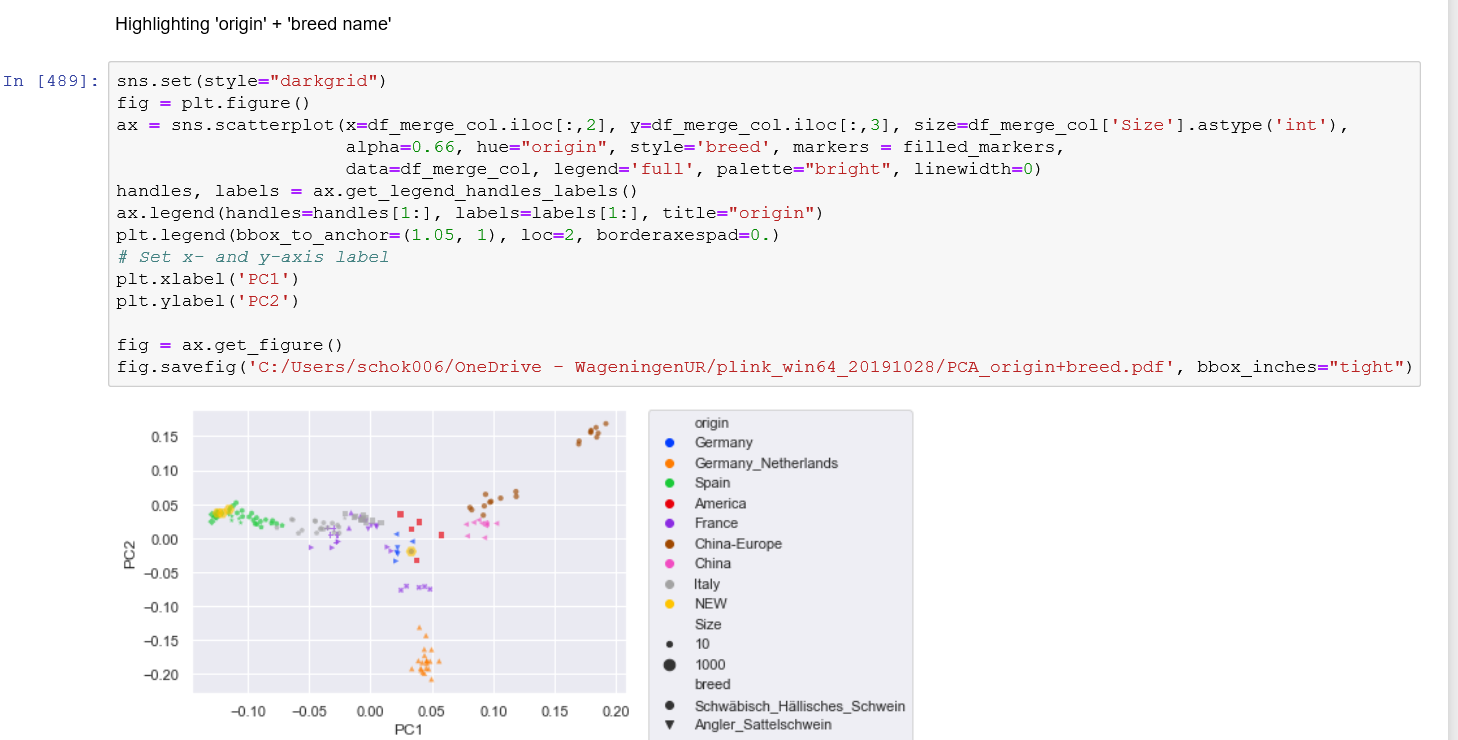
Next step is to generate code that will more easily be translated to different features in the scatterplots we will generate later on. We generate different colors based upon column 31 of the **‘merged’** dataset, which corresponds to the **origin**. We generate different shapes based upon column 30 of the ‘merged’ dataset, which corresponds to the **breed**. And lastly, we generate two new columns based upon the column with the **breed** name, namely *Size* and *SpecialMarker*, both to highlight our sample(s) of interest.



The last three blocks of code are generating different flavors of highlighting the sample(s) of interest with the corresponding metadata (origin and/or breed). Below all three block of code with their corresponding output.







The last piece of code generates a three-dimensional (3D) representation of the scatterplot. This may be used when the first two axes do not give a clear separation of the different origins and/or breeds.

