

# **PopMLvis Platform v1**

## User Manual

March 24, 2024

# Contents

1. Quick start
2. PopMLvis data types
  - 2.1. Genotype, binary plink-formatted, data (\*.bed, \*.fam, \*.bim)
  - 2.2. Correlation/Kinship matrix data
  - 2.3. Principal component result data (PC1, PC2, etc.)
  - 2.4. Admixture result data (ancestry fractions by *admixture*)
3. Importing (PopMLvis data format)
  - 3.1. Pickle file
  - 3.2. Comma Separated Value (CSV) file
  - 3.3. Gene structure files
4. Projections (dimensionality reduction algorithms)
  - 4.1. Principal components analysis (PCA)
  - 4.2. Linear Discriminant analysis (LDA)
  - 4.3. t-Distributed Stochastic Neighbor Embedding (t-SNE)
  - 4.4. Principal component analysis accounting for relatedness between subjects (i.e., PC-Air)
5. Clustering algorithms
  - 5.1. K-means
  - 5.2. Fuzzy c-means
  - 5.3. Hierarchical
6. Admixture analysis
  - 6.1. Admixture result using *Admixture* software
7. Outlier detection
  - 7.1. Statistical metrics

- 7.2. Isolation Forest
- 7.3. Minimum Covariance Determinant
- 7.4. Local Outlier Factor
- 7.5. OneClassSVM
- 8. Visualization
- 9. Exporting
- 10. How to install PopMLvis on your machine?
- 11. Citation and further information
- 12. Technical support
- 13. Future work

## 1. Quick start

PopMLvis is a population genetic visualization application for genome-wide association study (GWAS) data. We provide two versions of PopMLvis: a web-based online version and a computer-based offline version/application. PopMLvis provides a comprehensive interactive environment to explore population structure using genetic data and explore/infer ancestry groups and clusters using various dimensionality reduction algorithms.

## 2. PopMLvis data types

PopMLvis supports different types of input datasets. This gives more flexibility to users on how this tool can be used.

- 2.1. ***Genome-wide association study data***: This is the standard dataset that is used to test the association between genetic variables and disease of interest. The data contains genotypes of subjects and is highly dimensional (thousands of Single Nucleotide Polymorphisms (SNPs) and subjects).
- 2.2. ***Correlation/Kinship matrix data***: This dataset is  $N \times N$  dimension, where  $N$  represents the number of subjects. It contains the genetic correlation/kinship between all pairs of subjects.
- 2.3. ***Principal components data***: These are pre-computed principal components by the user using genotype data or other types of data.
- 2.4. ***Admixture dataset***: This dataset is the result of *Admixture* (or similar models), which contains the admixture fractions of subjects across a predefined number of clusters.

## 3. PopMLvis data format

PopMLvis accepts multiple file formats, which represent gene population data.

- 3.1. ***Comma Separated Value (CSV) file***: common, space or tab-delimited input files are accepted. Headers are required and can include:

IID: it represents the id of a single individual.

PC<sub>1</sub>: the first Principal Component

PC<sub>2</sub>: the second Principal Component

...

PC<sub>N</sub>: the Nth Principal Component

Metadata information: these are extra columns that could be included in the dataset (e.g., Ancestry, Age, Sex, Phenotype status, SNPs, etc.).

- 3.2. **GWAS data:** this is the binary plink format. Three files are required: .bed, .bim, and .fam. In addition, if users want to run PC-air, they should provide a correlation/kinship matrix of all pairs of individuals (space- or comma-delimited). This can be computed by many tools such as plink, GCTA, KING, etc. If the kinship matrix is not provided, PopMLvis uses the identity matrix by default.
- 3.3. **Pickle file:** it is a binary format that can be used to store gene population datasets, including metadata fields. Pickle is used internally by python to serialize objects. It is a faster and more flexible format. However, it is not supported by many programs (applications/software).

Note that many kinship calculators provide outputs in a long format, e.g., in the case of GCTA, the output looks as follows:

IID<sub>1</sub> IID<sub>1</sub> Kinship<sub>1</sub>

IID<sub>1</sub> IID<sub>2</sub> Kinship<sub>2</sub>

IID<sub>1</sub> IID<sub>3</sub> Kinship<sub>3</sub>

etc.

To convert to our matrix format, users can use the following code in R:

```
#####
fam = read.table("data.fam") # data.fam is the fam file used to compute kinship
grm = read.table("kinship.grm") # the long-formatted kinship
list_self = which(grm$V1 == grm$V2) # self-kinship
grm_noself= grm_noself[-c(list_self ),]
out = matrix(NA , nc=nrow(fam) , nr= nrow(fam))
diag(out) = grm[[list_self,4]]
out[upper.tri(out)] = (grm_noself$V4)
out[lower.tri(out)] = t(out)[lower.tri(out)]
write.table(out , "GCTA_matrix",quote=FALSE,row.names=FALSE,col.names=FALSE)
#####
```

## 4. Projections (dimensionality reduction algorithms)

PopMLvis supports multiple dimensionality reduction algorithms, which help visualize the latent structure in GWAS dataset.

- 4.1. ***Principal components analysis (PCA)***: principal components analysis (PCA) is a traditional, well-known, and most used linear transformation technique to visualize the genetic diversity in a dataset. It focuses on capturing the direction of maximum variation in a dataset through these principal components.
- 4.2. ***Principal component analysis accounting for relatedness between subjects (PC-Air)***: It is used to perform a principal components analysis using genome-wide SNP data for the detection of population structure in a sample. Unlike a standard PCA, PC-Air accounts for sample relatedness (*known* or *cryptic*) to provide accurate ancestry inference that is not confounded by family structure.
- 4.3. ***Linear Discriminant analysis (LDA)***: It is a linear transformation technique, like PCA, to find a linear combination of features that best explain the GWAS dataset. It could be categorized as a supervised dimensionality reduction technique, which could be exploited in classifying the dataset simultaneously.

- 4.4. ***t-Distributed Stochastic Neighbor Embedding (t-SNE)***: It is a non-linear transformation technique that is well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions. It tries to preserve the local structure (cluster) of genetic data and capture outliers simultaneously.

## 5. Clustering algorithms

- 5.1. ***K-means***: It is one of the most popular clustering algorithms. It stores k-centroids, which are used to define the clusters (ancestry groups). Then, each data point, which represents an individual, is assigned to the nearest cluster centroid. After that, it calculates the means (updated centroids) of data points in each cluster. This process is repeated until the assignment of data points no longer changes, which means that each subject is assigned to a given cluster (e.g., ancestry group).
- 5.2. ***Fuzzy c-means***: It is similar to K-means, but instead of assigning each data point (i.e., individual) to only one cluster, each data point can belong to many clusters with a weighting percentage. The weighting percentage increases if data points are close to the cluster centroid and decrease if they are far from the centroids.
- 5.3. ***Hierarchical clustering***: The general strategy is to follow a bottom-up approach “*agglomerative*”, where each data point starts in its cluster and pairs of clusters are merged as one moves up the hierarchy. We end up having only one cluster for the whole genotype dataset. Then, based on the user’s decision of how dissimilar clusters should be; a threshold value is applied. A dendrogram “tree-like” is the commonly used representation for hierarchical clustering.

## 6. Admixture analysis

6.1. **Admixture software:** It is one of the widely used admixture algorithms to estimate ancestry fractions of each subject. This is a supervised approach where a predefined number of clusters should be selected by users before running *Admixture*.

## 7. Outlier detection

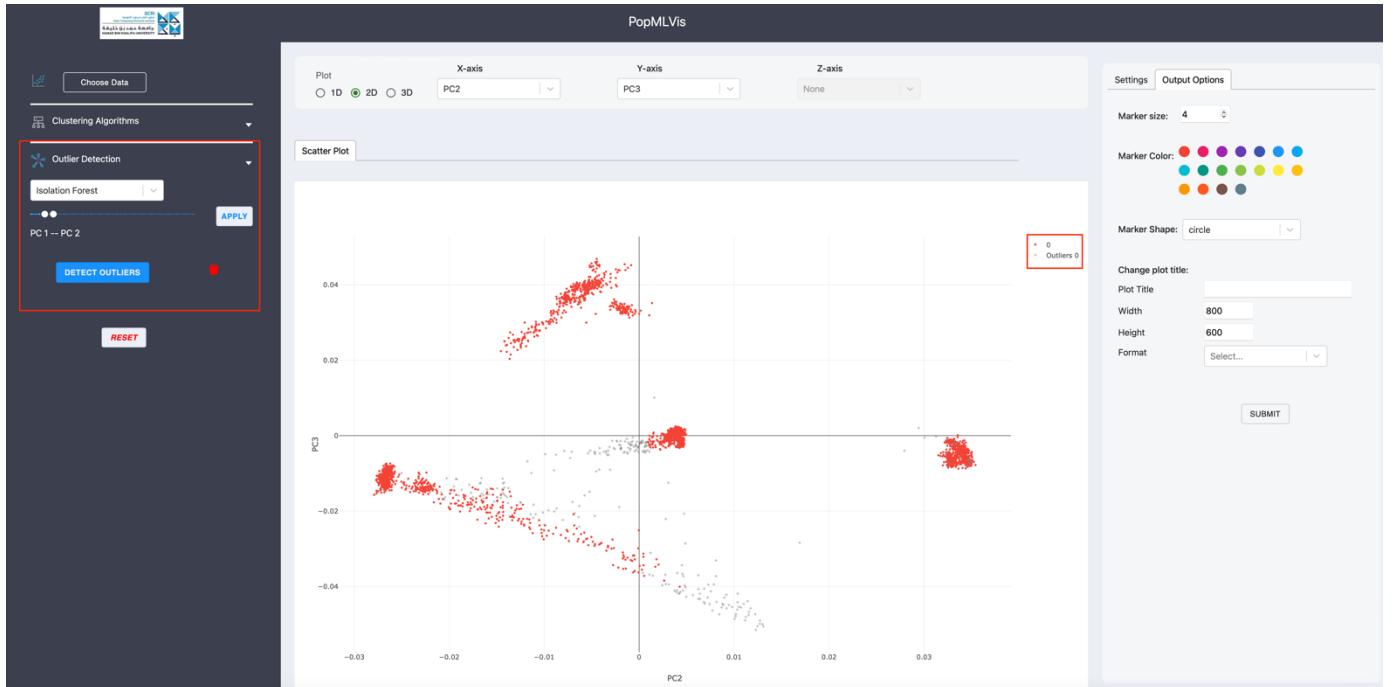
PopMLvis provides multiple outlier detection techniques to flag subjects that could be excluded from downstream analysis:

7.1. **Statistical metrics:** Using principal components (PCs), deviation from the mean ( $\mu$ ) is used to detect outliers ( $\mu \pm 3\sigma$ ,  $\mu \pm 2\sigma$ , etc.), where  $\sigma$  is the standard deviation. Users can define the list of PCs and the standard deviation threshold that can be used for outlier detection ( $\mu \pm 3\sigma$  on PC<sub>1</sub> **AND** PC<sub>2</sub>;  $\mu \pm 3\sigma$  on PC<sub>1</sub> **OR** PC<sub>2</sub>; etc.). **Figure A** shows Statistical metrics detection on a sample dataset.



Figure A: Outlier detection using statistical metrics ( $\mu \pm \sigma$ ).

**7.2. *Isolation Forest*:** This method identifies anomalies by isolating outliers in the data. It is based on a decision-tree algorithm, where it recursively generates partitions on the dataset by randomly selecting a feature and then randomly selecting a split value for the feature (e.g., PC). **Figure B** shows Isolation Forest Outlier detection on a sample dataset.



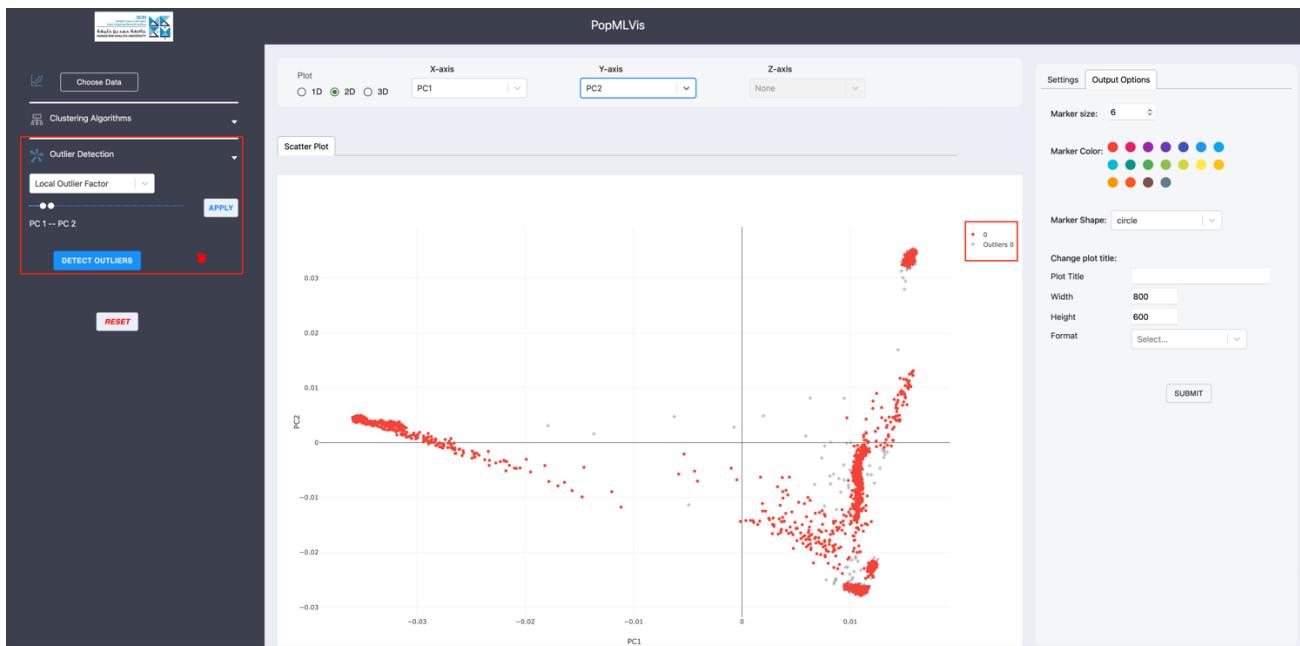
*Figure B: Isolation Forest outlier detection algorithm.*

**7.3. *Minimum Covariance Determinant*:** It estimates the mean and covariance matrix for each subset in the data. Then, it keeps the estimates for the subset whose covariance matrix has the smallest determinant (the most tightly distributed). **Figure C** shows minimum covariance determinant detection on a sample dataset.



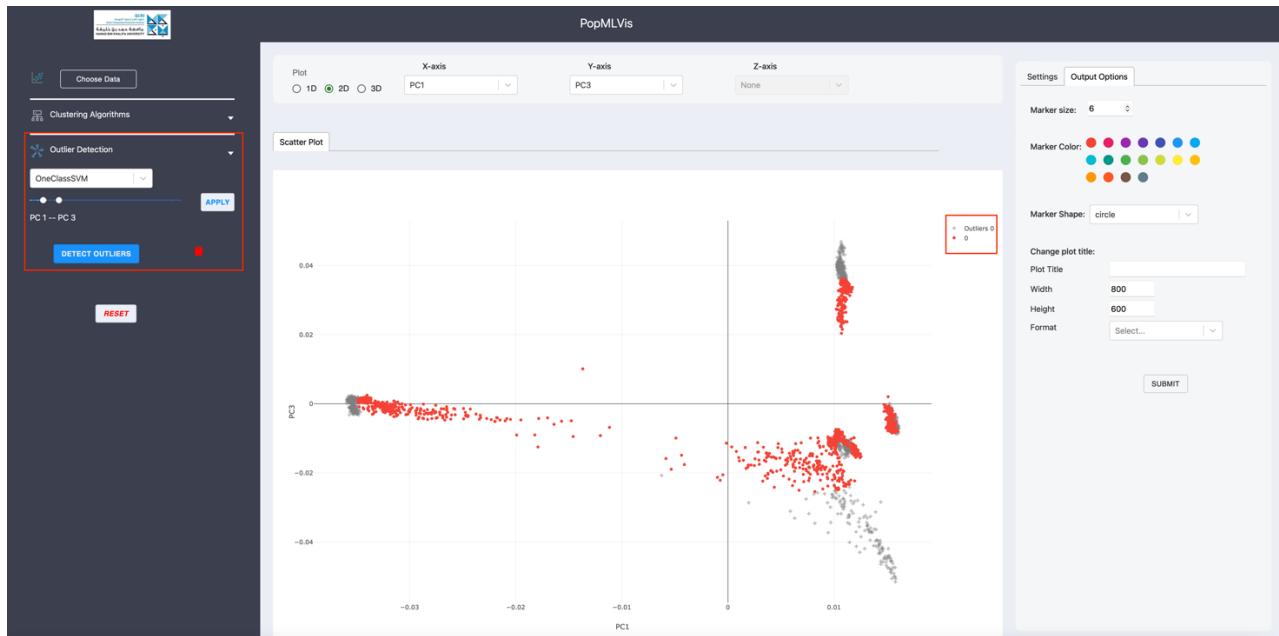
*Figure C: Minimum Covariance Determinant outlier detection algorithm.*

**7.4. Local Outlier Factor:** The anomaly score of each sample is called the Local Outlier Factor. It measures the local deviation of the density for a given sample with respect to its neighbors, where the locality is given by k-nearest neighbors, whose distance is used to estimate the local density. **Figure D** shows Local Outlier Factor detection on a sample dataset.



*Figure D: Local Outlier Factor outlier detection algorithm.*

7.5. **OneClassSVM**: It is a variation of the SVM classification algorithm. The algorithm is modeled as one class, which permits the algorithm to capture the density of the majority class and classifies examples on the extremes of the density function as outliers. **Figure E** shows OneClassSVM detection on a sample dataset.



*Figure E: OneClassSVM outlier detection algorithm.*

## 8. Visualization

In this part, the user will visually experience how the PopMLvis looks like and learn how to use it.

8.1. **Main dashboard:** The main window of PopMLvis overviews all components of the application.

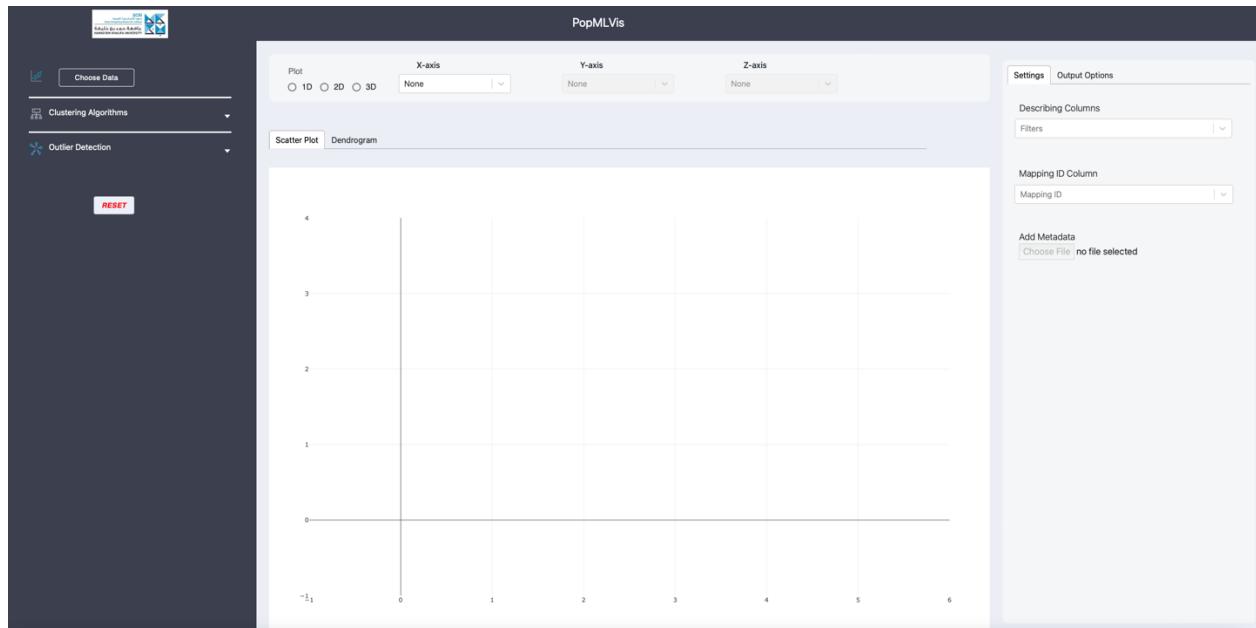


Figure 1: The main window of PopMLvis.

As can be seen in Figure 1, the primary PopMLvis dashboard can be categorized into four panels:

- a. Choose data: We designed this panel in an interactive manner with the user. The users have the choice to either upload their dataset or utilize the provided datasets. This panel also provides the user freedom to choose the dataset type they want, between the PCA data and/or Admixture results.  
For the datasets, we provided two sample datasets (1000 Genomes Project (1KG) and Human Genome Diversity Project (HGDP)) that can be utilized. There is a separate visualization for PCA and Admixture components (see Figure 2).

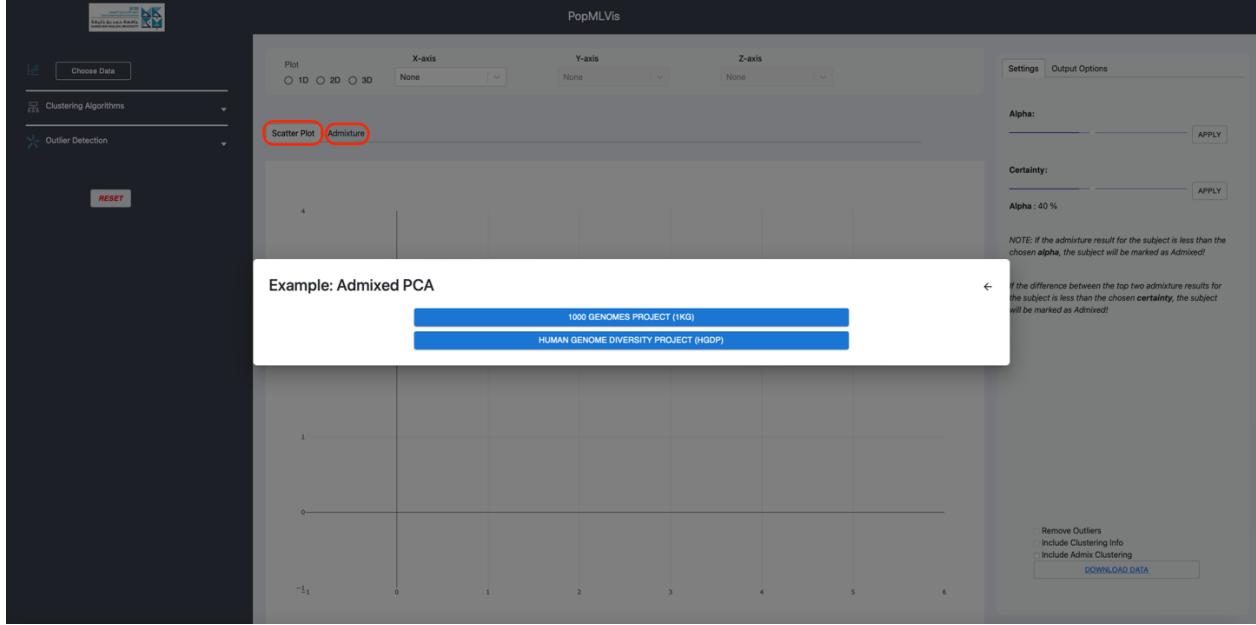
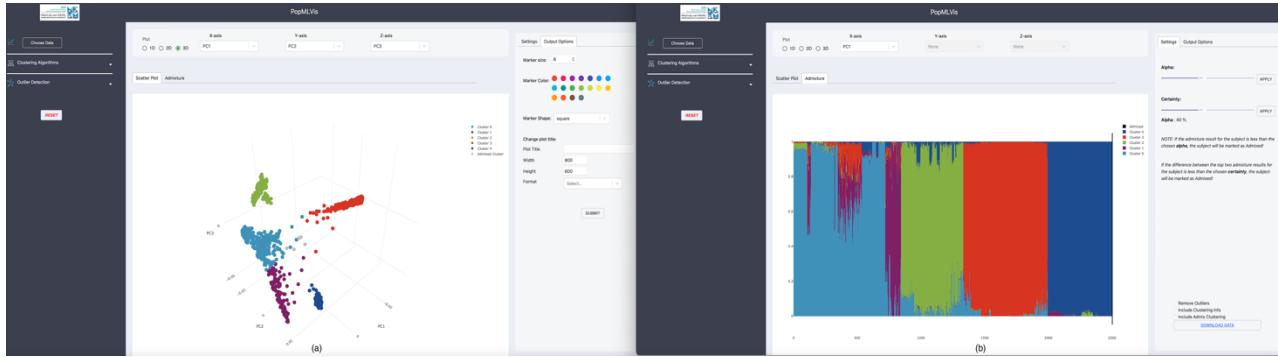


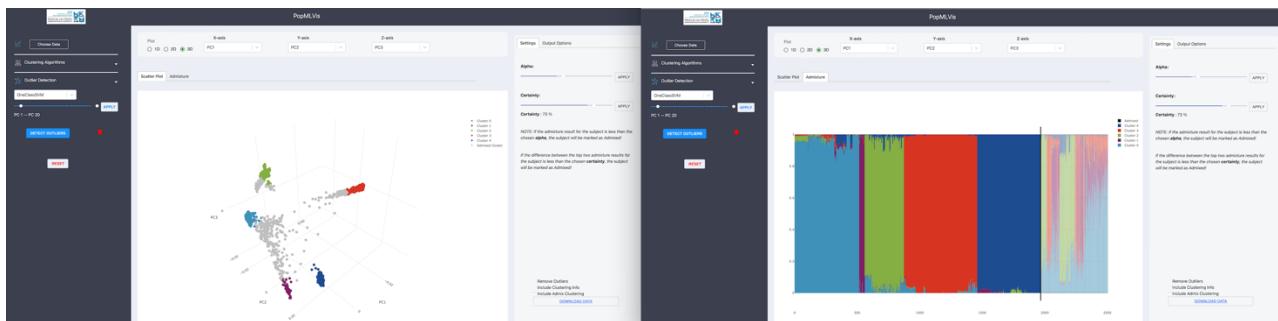
Figure 2: Input panel of PopMLvis.

- PCA and Admixture - It is the combination of the PCA data and the Admixture results. Admixture will try to cluster the data based on their ancestry, and the output will be a set of probabilities  $p_1, p_2, \dots, p_n$  where  $n$  is the number of clusters.
  - We say subject  $s_i$  belongs to cluster  $k$ , if:
    - $p_k = \max(p_1, p_2, \dots, p_k, \dots, p_n)$  and  $p_k > \text{alpha}$ ; or
    - $p_k = \max(p_1, p_2, \dots, p_k, \dots, p_n)$ , and  $p_k - p_j > \text{certainty}$ , where  $p_j = \max(p_1, p_2, \dots, p_{k-1}, p_{k+1}, \dots, p_n)$  i.e the second largest probability.
    - Here, alpha is a regularization parameter that penalizes intermediate ancestry proportions, whereas certainty is how certain an individual belongs to specific ancestry.
  - The visualization result is a bar plot and a scatter plot (see Figure 3). In the scatter plot, each subject is colored by the dominant cluster assigned by admixture, based on alpha/ certainty. The greyed-out subjects are admixed (see Figure 3(a)).
  - In the bar plot, each bar of height 100% corresponds

to one subject, where the stacked colors describe the genetic component proportions of the subject. The greyed-out subjects are admixed (see Figure 3(b)).



*Figure 3: (a) The effect of alpha in PCA scatter plot (b) The effect of alpha in Admixture bar plot.*



*Figure 4: (a) The effect of certainty in PCA scatter plot Figure 4: (b) The effect of certainty in Admixture bar plot*

- b. **Dimensionality Reduction:** If the input data consists of a large number of features, PopMLvis is compatible with performing Dimensionality Reduction algorithms.

As shown in Figure 5, PopMLvis supports 4 dimensionality reduction algorithms to make it possible for the user to analyze high dimensional data more efficiently.

All the options have an expected data type input.

## Select Dimensionality Reduction Method

- PCA (USING CORRELATION MATRIX)
- PC-AIR (USING PLINK FILES AND KINSHIP)
- T-SNE 2D (USING PCA DATA)
- T-SNE 3D (USING PCA DATA)

Figure 5: Two different examples of Dimensionality Reduction algorithms: (a) PCA (expected input: Correlation Matrix/Genetic Relationship Matrix). (b) PC-Air (expected input: PLINK files, optional: Kinship).

### 1. PCA:

- a. **Expected input:** Correlation Matrix (NxN)
- b. **Output:** Low Dimensional Data (Nxk, where k is defined by the user)

### 2. PC-Air:

#### a. **Expected input:**

- i. .bed (PLINK binary biallelic genotype table)
- ii. .bim (PLINK extended MAP file)
- iii. .fam (PLINK sample information file)
- iv. **Kinship(optional)** : A symmetric matrix of pairwise kinship coefficients for every pair of individuals in the sample.

*If the kinship matrix is not provided, the result will be a usual PCA.*

- b. **Output:** Low Dimensional Data (Nxk, where k is defined by the user)

### 3. t-SNE 2D:

- Expected input:** PCA data or Correlation Matrix/GRM
- Output:** 2D data, see Figure 6.

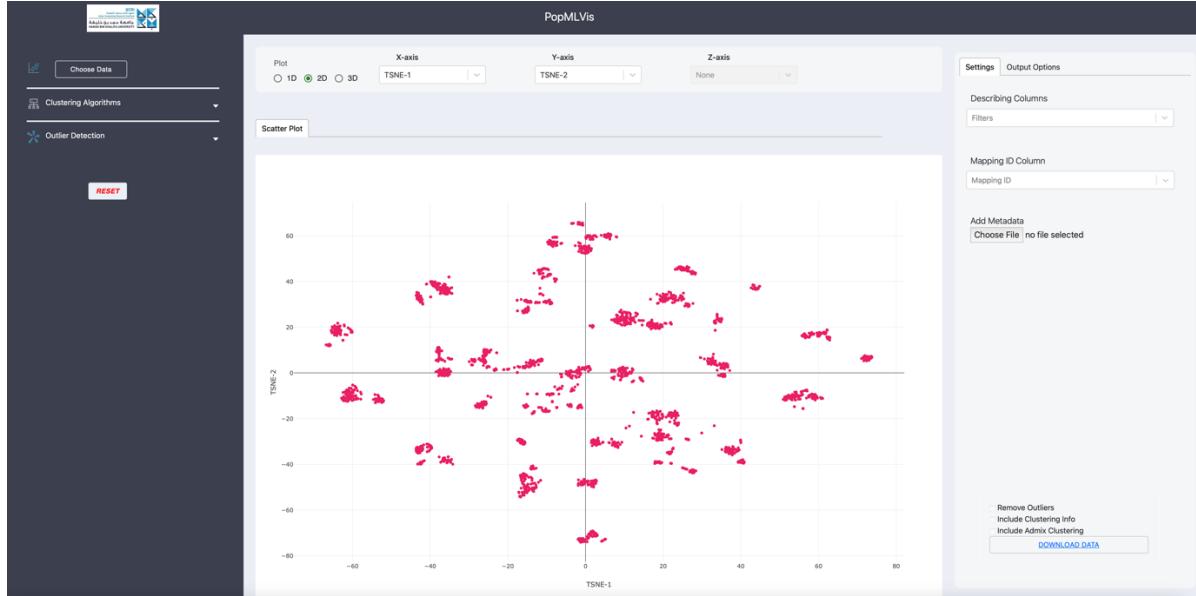


Figure 6: 2D dimensionality reduction of 1000 Genomes Project.

### 4. t-SNE 3D:

- Expected input:** PCA data or Correlation Matrix/GRM
- Output:** 3D data, see Figure 7.

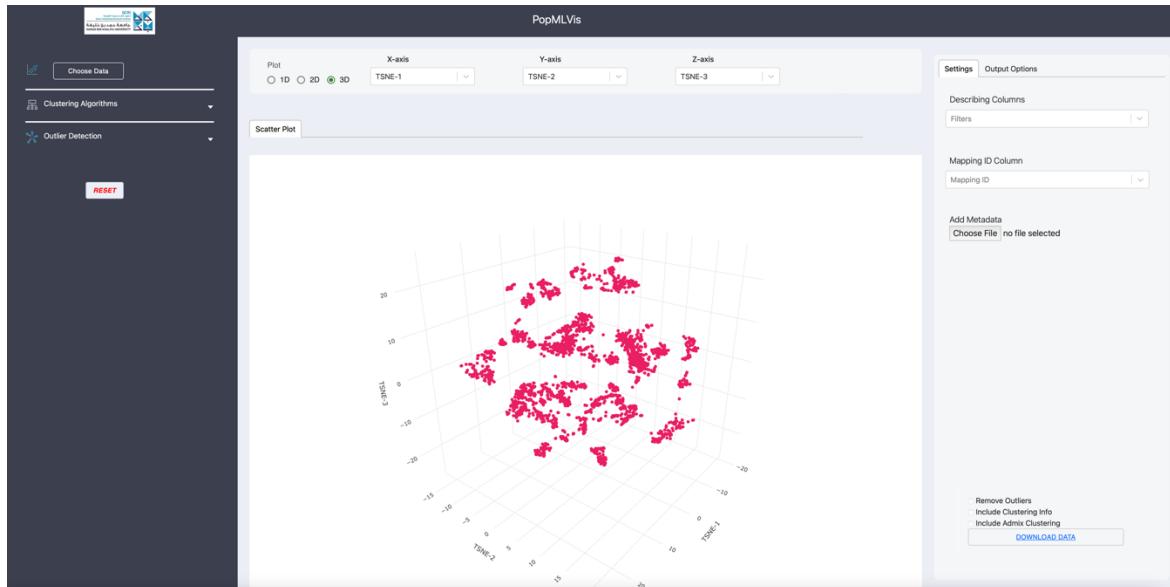


Figure 7: 3D dimensionality reduction of 1000 Genomes Project.

- b. **Visualization panel:** This panel provides the user with different options to choose from in terms of the number of dimensions (1D, 2D, or 3D) and which principal components to be viewed (see Figure 8).

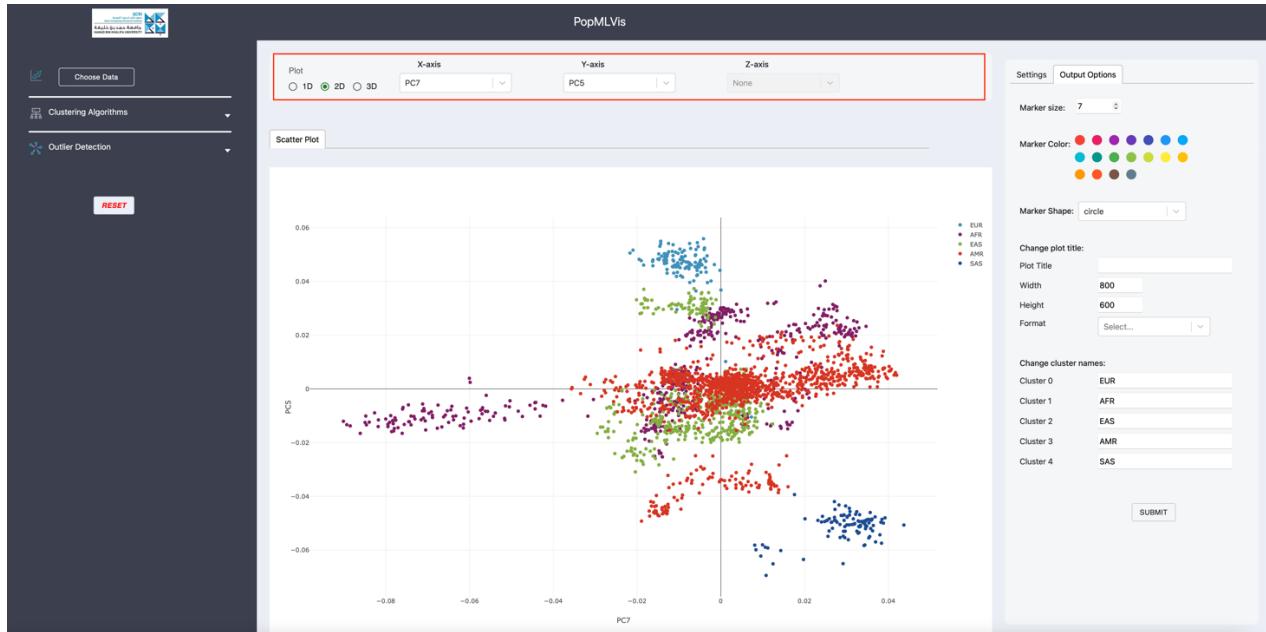


Figure 8: Viewing the principal components (PCAs).

- c. **Clustering panel:** This panel allows the user to apply a variety of clustering algorithms to the uploaded dataset and visualize the results spontaneously. For each algorithm, the user can set the parameters such as the number of clusters (see Figure 9).

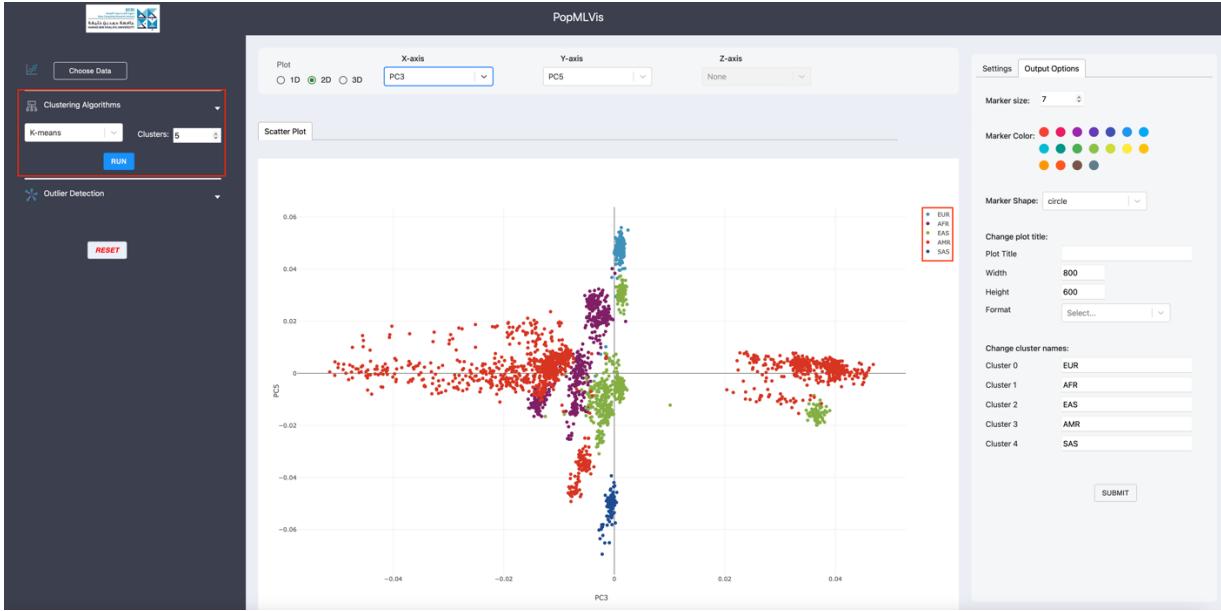


Figure 9: Clustering algorithms panel.

- d. **Outlier detection panel:** The user can specify which principal component they want to remove outliers from, and it is up to the user to choose more than one principal component (see Figure 9). Also, the user can select if they want to do “AND” or “OR” operations when there is more than one principal component. Moreover, the user must decide the deviation (e.g., 1SD, 2SD, etc.) from the mean ( $\mu$ ) to be flagged as an outlier (see Figure 10).



Figure 10: Outlier detection panel.

## 9. Exporting outputs

The download panel has many options for the users to choose from, as depicted in Figure 11.

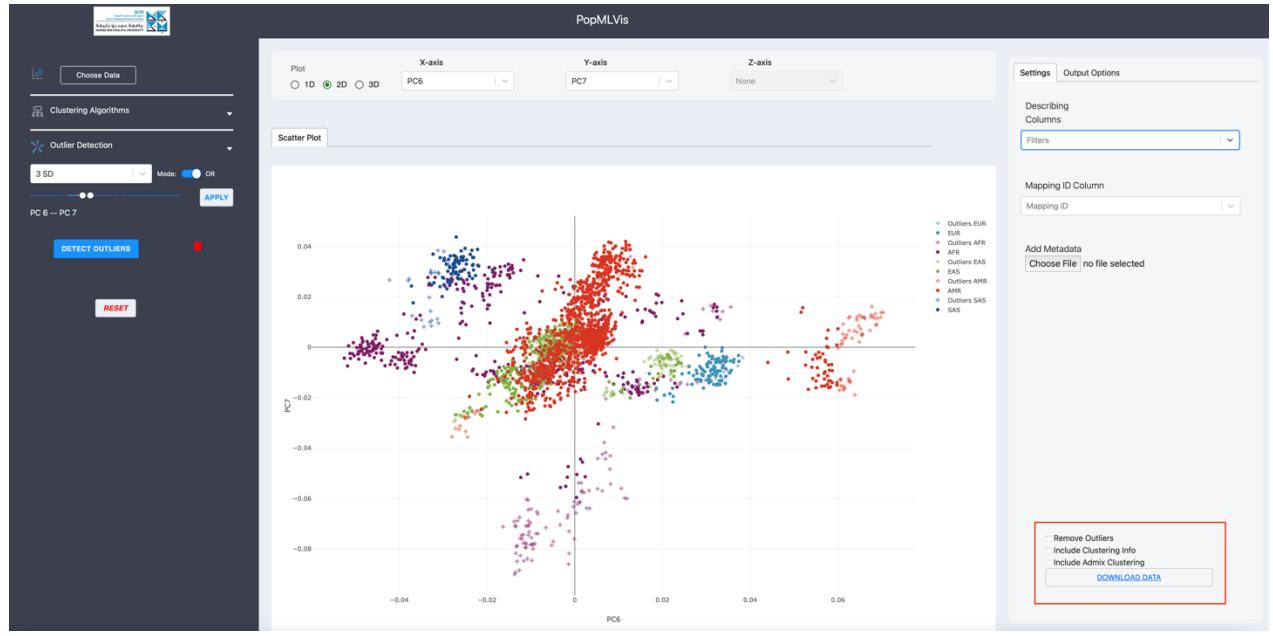


Figure 11: Download Data panel.

After performing the required operations, the user can export the output in a csv file. As you can see in Figure 12, the user can download the data with the following options:

### a.) Removing Outliers

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20	IID	ancestry	cluster
2	0.01125	-0.02719	-0.01224	0.017589	-0.00123	0.006608	0.001872	-0.01893	0.010684	0.014557	-0.0036	-0.02652	0.001872	0.014295	0.026714	-0.01264	0.000767	-0.00563	0.013243	-0.00162	HG00096	EUR	AMR
3	0.011029	-0.02677	-0.01082	0.016826	0.000965	0.007595	0.007671	-0.03098	0.009948	0.005814	0.004331	0.001444	0.008292	0.017359	0.019141	-0.00817	0.001275	0.002693	0.010754	0.005611	HG00097	EUR	AMR
4	0.01131	-0.02687	-0.01176	0.015817	0.001032	0.00609	0.005654	-0.03414	0.01149	0.012079	0.007453	0.000689	0.003533	0.020264	0.021398	-0.01321	0.0078	0.002397	-0.00273	0.003889	HG00099	EUR	AMR
5	0.010922	-0.02702	-0.01218	0.018376	-0.00071	0.0043	-0.00307	-0.0024	0.001356	0.01578	-0.00669	0.040794	-0.01158	0.017124	0.019595	-0.00603	0.008113	0.005191	0.000113	0.009873	HG00100	EUR	AMR
6	0.01119	-0.0268	-0.01239	0.016547	0.001938	0.005802	0.00268	-0.02904	0.008261	0.010632	0.008104	0.007729	0.007059	0.019129	0.039442	-0.01268	0.007005	-0.00064	0.002292	0.005298	HG00101	EUR	AMR

Figure 12: (a) Data excluding outliers.

### b.) Including Clustering Information

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20	IID	ancestry	cluster	outlier
2	-0.02719	-0.01224	0.017589	-0.00123	0.006608	0.001872	-0.01893	0.010684	0.014557	-0.0036	-0.02652	0.001872	0.014295	0.026714	-0.01264	0.000767	-0.00563	0.013243	-0.00162	HG00096	EUR	AMR	0
3	-0.02677	-0.01082	0.016826	0.000965	0.007595	0.007671	-0.03098	0.009948	0.005814	0.004331	0.001444	0.008292	0.017359	0.019141	-0.00817	0.001275	0.002693	0.010754	0.005611	HG00097	EUR	AMR	0
4	-0.02687	-0.01176	0.015817	0.001032	0.00609	0.005654	-0.03414	0.01149	0.012079	0.007453	0.000689	0.003533	0.020264	0.021398	-0.01321	0.0078	0.002397	-0.00273	0.003889	HG00099	EUR	AMR	0
5	-0.02702	-0.01218	0.018376	-0.00071	0.0043	-0.00307	-0.0024	0.001356	0.01578	-0.00669	0.040794	-0.01158	0.017124	0.019595	-0.00603	0.008113	0.005191	0.000113	0.009873	HG00100	EUR	AMR	0
6	-0.0268	-0.01239	0.016547	0.001938	0.005802	0.00268	-0.02904	0.008261	0.010632	0.008104	0.007729	0.007059	0.019129	0.039442	-0.01268	0.007005	-0.00064	0.002292	0.005298	HG00101	EUR	AMR	0
7	-0.02681	-0.01142	0.017157	0.00037	0.007465	0.001083	-0.02782	0.005286	0.008128	0.00306	0.007449	0.003999	0.016458	0.032868	-0.01325	0.006091	0.000315	0.006644	0.012942	HG00102	EUR	AMR	0
8	-0.02661	-0.01177	0.017444	-0.00124	0.004727	0.00506	-0.02233	0.010386	0.015461	0.008195	-0.02155	0.004586	0.007559	0.02404	-0.01152	-0.00152	-0.00195	0.012246	0.001235	HG00103	EUR	AMR	0
9	-0.02652	-0.01137	0.017209	0.001566	0.006101	0.007439	-0.02789	0.008366	0.011046	0.002586	0.006362	0.012514	0.018737	0.027375	-0.00675	0.006601	0.006262	-0.00193	-0.00247	HG00105	EUR	AMR	0

(b)

Figure 12: (b) Including Cluster Information.

### c.) Including Admix Clustering Information

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	IID	ancestry	admixCluster	W
1	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20				
2	0.01125009	-0.0271943	-0.0122364	0.01758942	-0.0012338	0.00660821	0.0018717	-0.0189322	0.01068373	0.0145573	-0.0036003	-0.0265196	0.00187212	0.0142952	0.02671383	-0.0126395	0.00076744	-0.0056344	0.0132435	-0.0016236	HG00096	EUR	0	
3	0.0110286	-0.0267718	-0.0108162	0.0168259	0.000965205	0.0075934	0.0067077	-0.0309776	0.00994801	0.00581364	0.0043137	0.0014379	0.0082924	0.01735922	0.01914092	-0.0081704	0.00127461	0.00269278	0.0107537	0.00561137	HG00097	EUR	0	
4	0.0113101	-0.026875	-0.0117642	0.01581749	0.00103207	0.0066901	0.00565447	-0.0341359	0.01149042	0.01207875	0.0074525	0.00068890	0.0035323	0.02026424	0.02139812	-0.0132036	0.00780003	0.00239737	-0.0027292	0.00388863	HG00099	EUR	0	
5	0.01092207	-0.0270215	-0.0121803	0.01857628	-0.0007059	0.00429880	-0.0030722	-0.0023867	0.0013592	0.0157800	-0.0066930	0.0407835	-0.0115832	0.01712378	0.01959494	-0.0060284	0.00811305	0.00518132	0.0011252	0.00987337	HG00100	EUR	0	
6	0.01118957	-0.0268605	-0.0123876	0.01654869	0.00193759	0.005802	0.00268033	-0.0290419	0.00826079	0.01063151	0.00810376	0.00772927	0.00705054	-0.006412	0.00529848	0.00529848	0.0191295	0.03944185	-0.0126766	0.00705054	0.00229204	HG00101	EUR	0

(c)

Figure 12: (c) Including Admix Cluster Information.

### Illustration Example:

In this illustration example, we will walk through the PopMLvis platform in a step-by-step manner to show how the user can analyze and visualize his/her own dataset. Here, we will utilize the 1000 Genomes Project (1KG) dataset to perform the analysis and visualization.

The first step, the user has to click on the “Choose Data” option on the upper-left corner of the main dashboard, as shown in Figure 1.

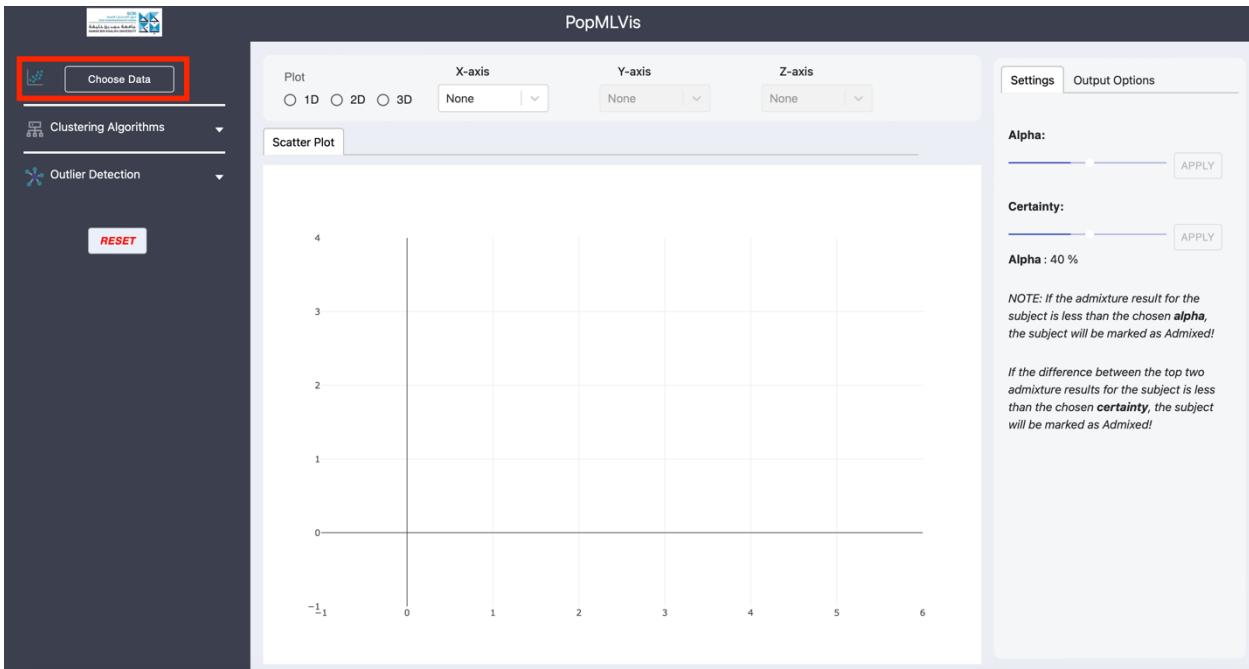
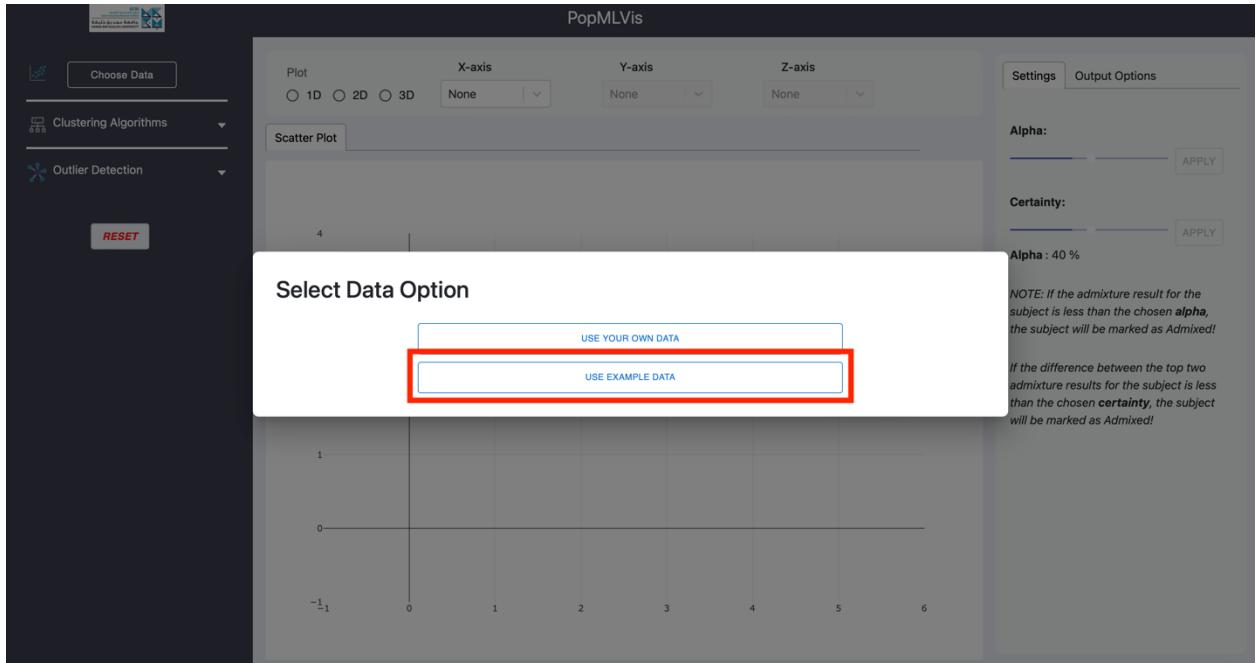


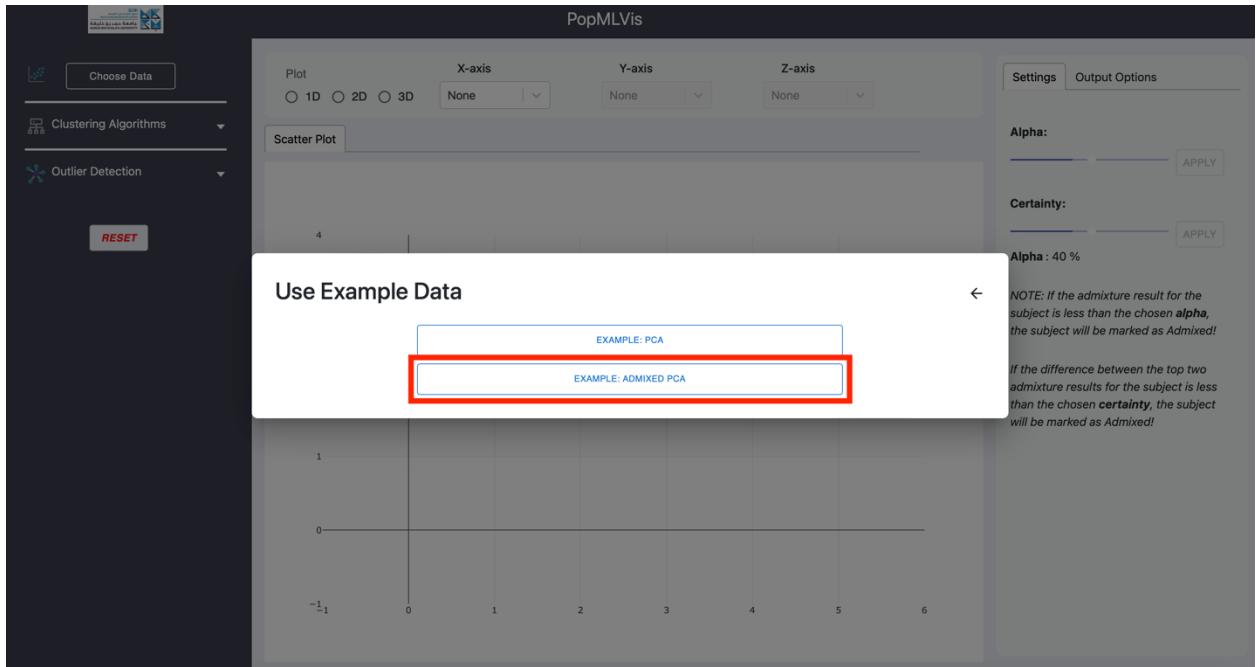
Figure 1: upload the dataset.

Then, a small pop-up window will appear asking the user to either use his/her own dataset or use the example dataset. Here, we will click on the second option, which is use example dataset, as depicted in Figure 2.



*Figure 2: select the dataset.*

After that, a new pop-up window will show with two options, where the user needs only the principal components (PCA), or both the principal components and admixture data. Suppose the user chooses the second option, as depicted in Figure 3.



*Figure 3: use the provided dataset.*

Next, a small pop-up menu appears to provide the user options from which dataset he/she wants to utilize its Admixed and PCA data, as shown in Figure 4.

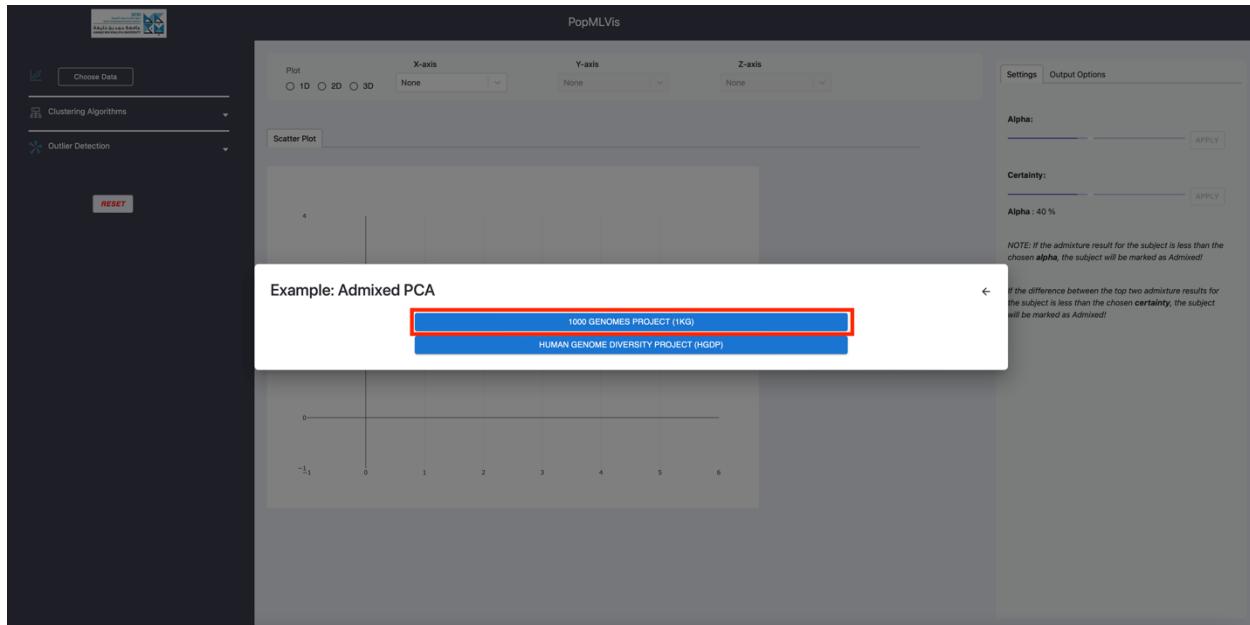


Figure 4: choose from which dataset the user needs admixed and PCA data.

Let's assume the user choose the first option, i.e., 1000 Genomes Project (1KG). Now, if the user clicks on any location outside the small pop-up menu, this activated the main dashboard window, as seen in Figure 5.

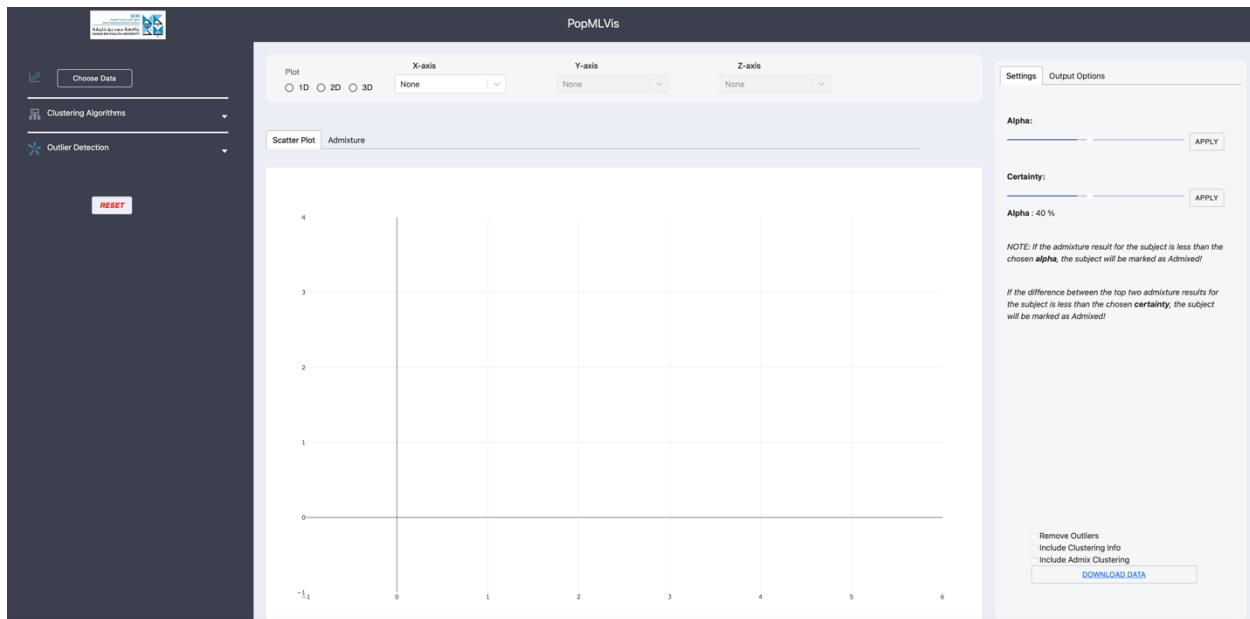
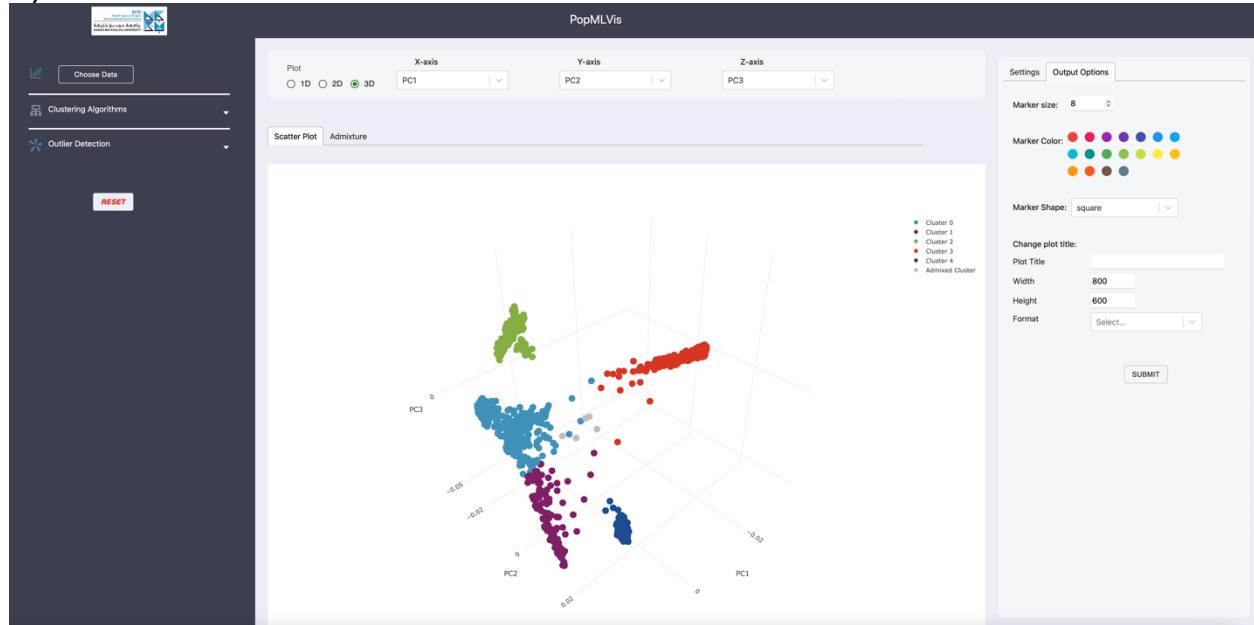


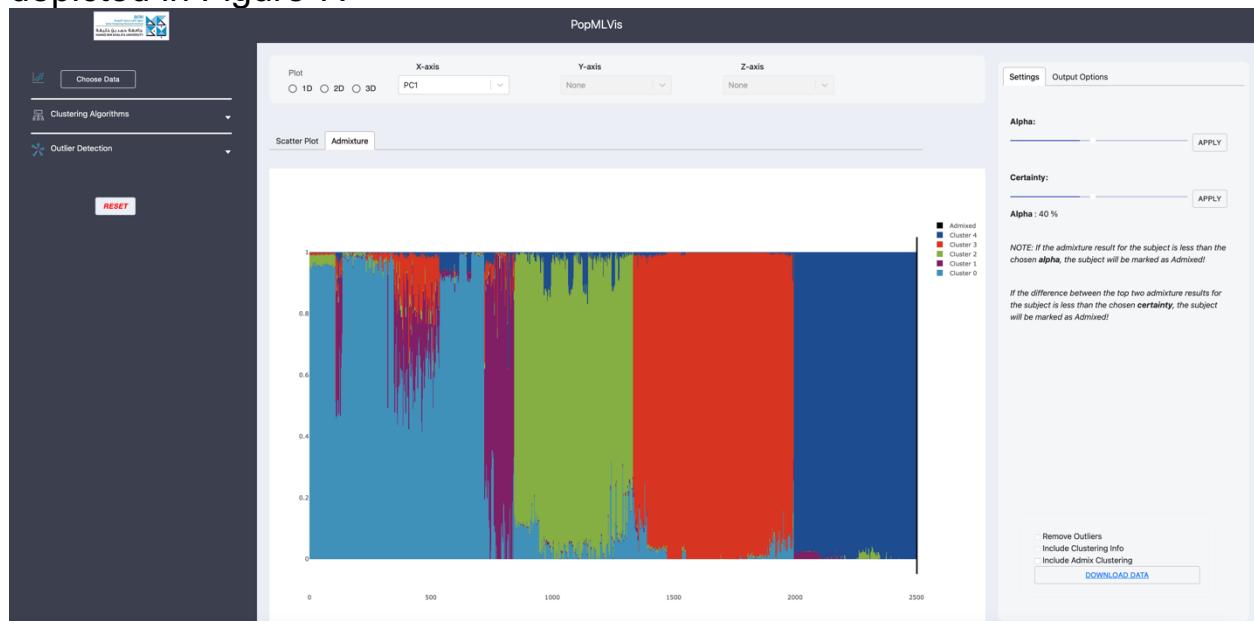
Figure 5: main dashboard window.

Now, both the admixture and PCA data is uploaded in the platform and the user can analyze and visualize them. Suppose the user click on the x-axis drop-down menu and choose any principal component (PC), the corresponding components will appear on the scatter plot window. In a similar manner, the user can choose to visual 2D, or 3D and the corresponding plot will appear automatically (see Figure 6).



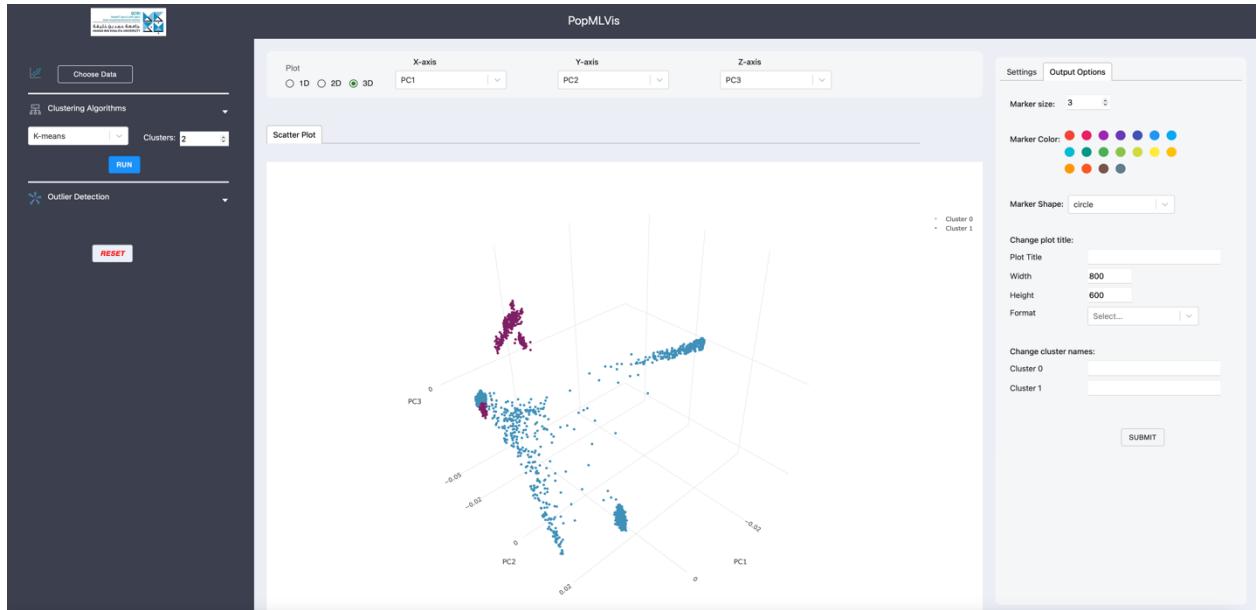
*Figure 6: scatter plot window.*

If the user chooses the Admixture tab, then, the corresponding plot will appear, as depicted in Figure 7.



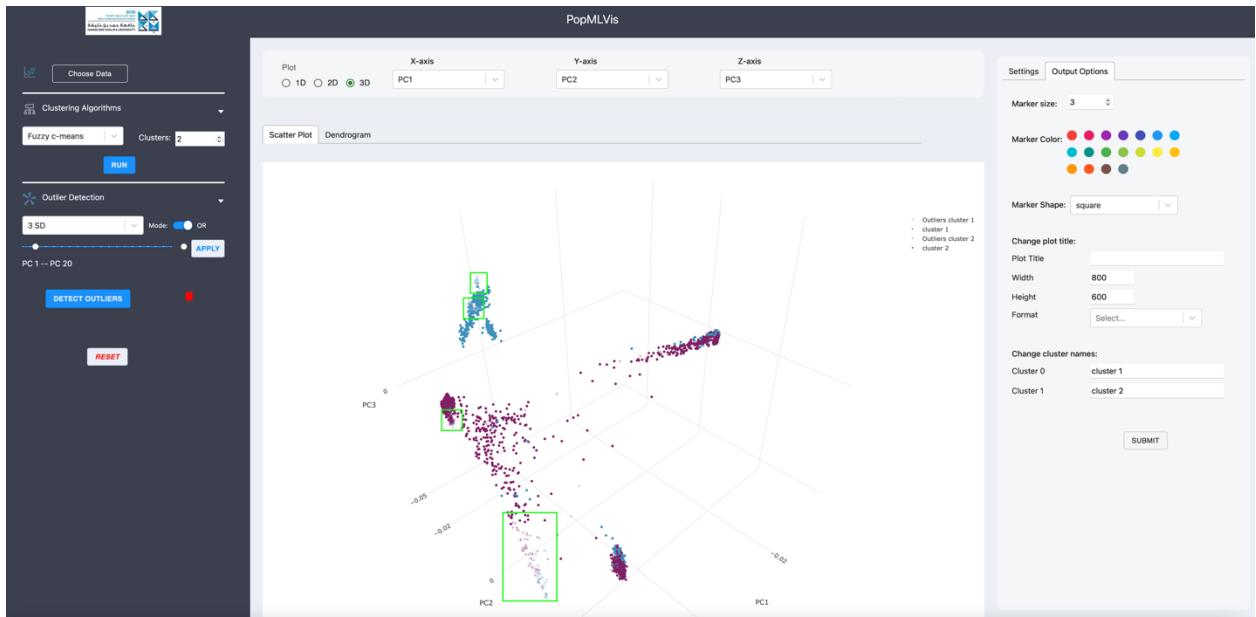
*Figure 7: Admixture window.*

Now, suppose the user wants to perform clustering using K-mean clustering algorithm with two classes (K=2). Then, the clustering results will show immediately on the main dashboard, as depicted in Figure 8.



*Figure 8: K-means clustering.*

In a similar way, suppose the user requires to remove outliers. Hence, based on the user can go to the outlier detection drop-down menu and choose which algorithm wants to use. Once the user chooses the outlier algorithm, he/she can choose from which principal components the outlier needs to be removed. This can be done using the sliding bar. Also, the user has a mode option to choose between AND or OR. Here, AND means the outlier needs to be on all chosen PCs to perform outlier detection on. Contrary, OR means that the outlier needs to be only in one PC from the chosen PCs to perform outlier detection on. Figure 9 shows the detected outliers using 3SD (standard deviation) with an OR mode for all PCs (PC<sub>1</sub> – PC<sub>20</sub>). Here, we highlighted some of the outliers that reside inside the green boxes. Also, the user can exploit the output options on the left hand-side of the window such as changing cluster names.



*Figure 9: outlier detection.*

## 10- How to install PopMLvis on your machine?

In case the user uses confidential data, PopMLvis provides an offline version, where the user can install it on their machine. PopMLvis is supported on Windows, macOS, and Linux operating systems.

The GitHub repository provides an installation script that can be used to install and run the software locally.

### Getting Started

These instructions will cover usage information for the docker container.

### Prerequisites

In order to run this container you'll need docker installed.

- Windows <https://docs.docker.com/windows/started/>
- OS X <https://docs.docker.com/mac/started/>
- Linux <https://docs.docker.com/linux/started/>

If cloning the repository, install .

- git-lfs. <https://github.com/git-lfs/git-lfs/blob/main/README.md>

### Usage

Download the source code by either:

- Cloning the repository from command line
  - `$ git clone https://github.com/qcri/QCAI-PopMLVis.git`
- Or downloading ZIP, see Figure 12.

NOTE: this might take a while because it will download a few big genotype files

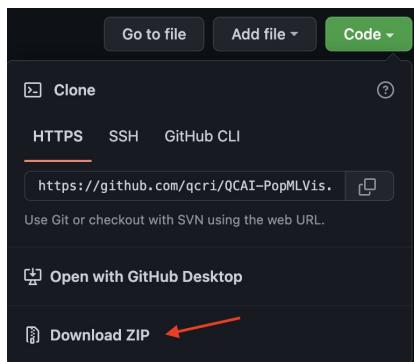


Figure 11

There are two configuration files, *backend/gunicorn\_local.conf*, and *frontend/envfile*. You can choose to modify them if you want to change configs such as the exposed ports, the location of the log outputs, or multiprocessing power.

After that, in your command line, navigate to the folder you downloaded/cloned and run:

```
$ docker-compose up
```

Docker will start fetching the required images. *NOTE: this might take a while*. After the initialization is done, open your web-browser and load <http://localhost:3000>

## 1. Citation and further information

If you find our platform useful, please cite our paper:

## 2. Technical support

In case the user has any question, or suggestion regarding the platform, he/she can send us an inquiry at [popmlvissupport@QCRI.org](mailto:popmlvissupport@QCRI.org)

## 3. Future work