BMC Bioinformatics

SOFTWARE

Open Access

PopMLvis: a tool for analysis and visualization of population structure using genotype data from genome-wide association studies

Mohamed Elshrif[1*], Keivin Isufaj[1], Khalid Kunji[1] and Mohamad Saad[1*]

*Correspondence: melshrif@hbku.edu.qa; msaad@hbku.edu.qa

[1] Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

Abstract

One of the aims of population genetics is to identify genetic differences/similarities among individuals of multiple ancestries. Many approaches including principal component analysis, clustering, and maximum likelihood techniques can be used to assign individuals to a given ancestry based on their genetic makeup. Although there are several tools that implement such algorithms, there is a lack of interactive visual platforms to run a variety of algorithms in one place. Therefore, we developed PopMLvis, a platform that offers an interactive environment to visualize genetic similarity data using several algorithms, and generate figures that can be easily integrated into scientific articles.

Keywords: Population genetics, Bioinformatics software, Machine learning, GWAS analysis

Introduction

Population structure can be inferred from Genome-Wide Association Study (GWAS) data and focuses on the genetic variation within and between populations by investigating the distributions of alleles and how their frequencies change over time [1]. Sophisticated algorithms implemented in standalone software are often used to infer population structure. A widely used tool is *ADMIXTURE* [2], which relies on maximum likelihood techniques [3]. Many of these software provide complementary results, but, to the best of our knowledge, there is a lack of a system that seamlessly visualizes the outputs of multiple software jointly. Another issue is that many softwares such as *ADMIXTURE* [2], *FASTSTRUCTURE* [3], *STRUCTURE* [4], and *STRUCTURESELECTOR* [5] rarely provide graphical outputs. Moreover, users cannot easily exploit existing additional related information (e.g., sex, disease status, known ancestry, etc.) while analyzing and interpreting their outputs like *ClustVis* [6].

Here, we considered all aforementioned drawbacks and developed an interactive platform, named PopMLvis, which carries out a wide range of tasks that a user may need to infer population structure using GWAS data. PopMLvis is flexible as: (1) It supports a variety of input datasets, i.e., raw genotype data, Principal Components (PCs), and admixture membership coefficient matrix; (2) It performs dimensionality reduction using Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and PC-Air, which is a principlal component analysis that accounts for relatedness through the genetic relationship matrix (GRM) [7, 8]; (3) It performs various clustering algorithms (e.g., K-means and Hierarchical Clustering); (4) It detects outliers using Isolation Forest, OneClassSVM, and other metrics; (5) It offers an interactive and zoomable friendly graphical user interface; (6) It produces publication-ready figures in various types and resolutions. In addition, PopMLvis allows users to: (7) Download output files generated within PopMLvis with all required information that are ready for downstream analysis (e.g., association testing); (8) Link metadata with obtained clustering results; and (9) Integrate estimated genetic diversity indices generated by genetic structure programs (e.g., *ADMIXTURE*) and the clustering results. Since PopMLvis has a modular design, it is easy to add new modules (e.g., classification) or a new algorithm to the existing modules (e.g., uniform manifold approximation). PopMLvis is a secure web-based platform. Due to potential privacy concerns, we provide an offline version that can be installed locally. PopMLvis can be easily used without the need to write any script, which makes it more accessible to researchers.

Implementation

PopMLvis consists of three main panels, each with unique functionalities that the user can perform as depicted in Fig. 1.

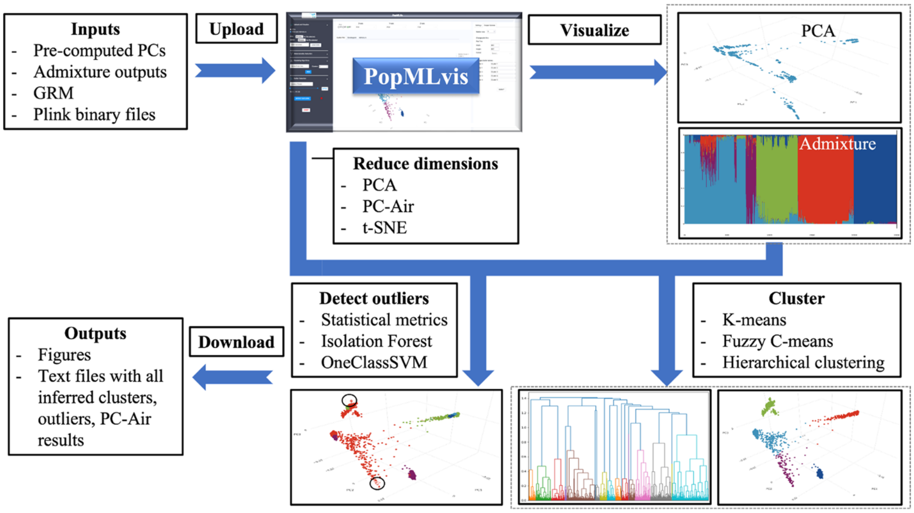


Fig. 1 PopMLvis pipeline/workflow: (1) Upload and visualize PCA and Admixture results; (2) Dimensionality reduction: PCA, PC-Air, t-SNE 2D and 3D; (3) Clustering: K-means, Fuzzy C-means, and Hierarchical Clustering; (4) Detecting outliers: Isolation Forest, local Outlier Factor, and Statistical measures; and (5) Download graphical plots and datasheets

1. *Input and Machine Learning (ML) Panel:* This panel is composed of three modules:
    - *Choose data:* The first step, users can choose to use their own data, or example data, which is provided along with PopMLvis platform. Then, users can specify the type of data to upload, if they prefer to use their own data, which could be raw or processed data. For processed data, it includes PCA and/or admixture outputs (i.e., fraction of ancestral origins as obtained by admixture tools [2–5]. Data will be immediately reflected on the visualization panel after the upload. For raw data, it includes genotype data, GRM, projected dataset, and PCA. Here, the user can perform dimensionality reduction as well. Moreover, PopMLvis supports PCA, PC-Air, and t-SNE 2D and 3D. Also, t-SNE can be run on top of PCA results to visualize the data in a more reduced space.
    - *Clustering algorithms module:* This module includes the K-means, Fuzzy C-means, and Hierarchical Clustering algorithms. The Fuzzy C-means algorithm is suitable when admixture exists between individuals, and these individuals can belong to multiple clusters/ancestries.
    - *Outlier detection module:* PopMLvis integrates outlier detection algorithms based on statistical metrics (mean, standard deviation, and covariance matrix) and machine learning techniques such as OneClassSVM, Local Outlier Factor, and Isolation Forest.
2. *Visualization panel:* This panel supports three interactive plot types: (1) Scatter plots: 1D, 2D, 3D, zoom in/out, legend and label naming, download, etc.; (2) Admixture bar charts: the user can investigate the estimated ancestral fraction for each individual with different certainty values; and finally, (3) Dendrograms to visualize the hierarchal clustering of the data. Scatter plots and admixture bar charts are linked together, so a change in one plot will be reflected in the other plot.
3. *Option panel:* This panel provides the users with an option to include additional information on individuals such as sex, age, disease status, etc. This can be reflected on the plots with color/shape differences. The user has flexibility to define plot name, labels, resolution, etc. This makes the PopMLvis graphical outputs ready to be integrated in scientific articles.

PopMLvis system architecture

The architecture of PopMLvis consists of three main components (see Fig. 2):

1. *Front-End:* The front-end is built using *ReactJS*. React makes our data visualization attractive and efficient. All communication with the back-end is achieved through *REST APIs*, benefiting from promise based *HTTP* clients for the browser. The website is compatible with different screen sizes, making the visualization dynamic.
2. *Back-End:* The back-end on the server side is served as a *REST API* and was developed using *Flask*. We used *Gunicorn* as a pre-fork worker model, where the master manages a set of workers. The number of workers corresponds with the number of concurrent requests that our back-end can handle. Gunicorn should only need 4–12 worker processes to handle hundreds or thousands of requests per second. *Python*

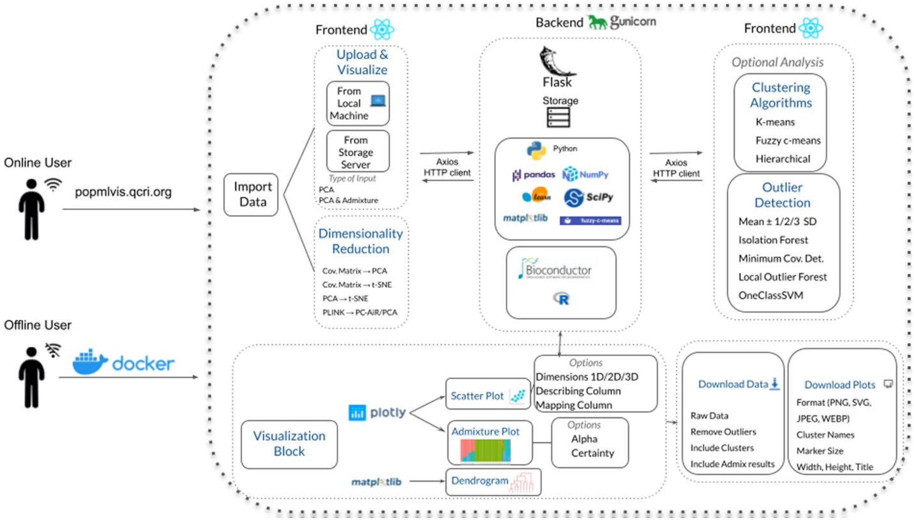


Fig. 2 PopMLvis schematic architecture

was used for the machine learning and computational algorithms. Numpy, Pandas, Matplotlib, Scikit-learn and Scipy are among the libraries that were used. To integrate the *PC-Air R package*, we needed to add another layer of communication between Python and R. In this case, Flask would serve as a middle layer, serving the front-end request to R and waiting for its response, to send it back to the front-end.

3. *Data layer:* PopMLvis can handle several types of data with various file extensions, including plink binary data (.bed,.fam,.bim), pre-computed PCA results, Genomic Relationship Matrix (GRM), and admixture results. Most of the data will never be stored on the server. It will be either encrypted inside the body of the request using HTTPS protocol, or used only on the front-end. The *choose data* tab will keep the data in the front-end only. When settings change, no requests will be made to the back-end. The clustering algorithms and outlier detection modules require the data to be sent to the back-end for computation, but results will be returned back to the user without storing or keeping any trace of it. Because of various encryptions and file extensions, the dimensionality reduction uploads are stored locally with encrypted filenames, processed, and results are communicated to the user. All gathered data will be cleaned through a job scheduler, CRON.

Conclusion

We developed PopMLvis, an interactive tool that supports many useful features for researchers to analyze and visualize population structure using genotype data. Our user-friendly platform offers important functionalities including: supporting a variety of input datasets, interactive visualization, performing dimensionality reduction, clustering, and outlier detection. In addition, for privacy purposes, we developed a web-based as well as a standalone version of the platform.

Methods

PopMLvis data types

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

- *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).
- *Correlation/Kinship matrix data:* This dataset is NxN dimension, where N represents the number of subjects. It contains the genetic correlation/ kinship between all pairs of subjects.
- *Principal component analysis data:* These are pre-computed principal components by the user using genotype data or other types of data.
- *Admixture data:* This dataset is the result of ADMIXTURE (or similar tools/models), which contains the admixture fractions of subjects across a predefined number of clusters.

PopMLvis data format

PopMLvis accepts multiple file formats:

- *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_1: the 1st Principal Component

  PC_2: the 2nd Principal Component.

  PC_N: 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.).
- *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.
- *Pickle file:* It is a binary format that can be used to store genetic 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/softwares).

Note that many kinship calculators provide outputs in a long format, e.g., in the case of GCTA [9], the output looks as follows:

IID$_1$ IID$_1$ Kinship$_1$
IID$_1$ IID$_2$ Kinship$_2$
IID$_1$ IID$_3$ Kinship$_3$
etc.

To convert it 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)
######
```

Projections (dimensionality reduction algorithms)

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

- *Principal components analysis (PCA):* principal components analysis 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.
- *Principal components 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 the standard PCA, PC-Air accounts for sample relatedness (known or cryptic) to provide accurate ancestry inference that is not confounded by family structure.
- *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.
- *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.

Clustering algorithms

- *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).
- *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 decreases if they are far from the centroids.
- *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.

Admixture analysis

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.

Outlier detection

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

- *Statistical metrics:* Using principal components (PCs), deviation from the mean (μ) is used to detect outliers ($\mu \pm 3\sigma$, $\mu \pm 2\sigma$, etc.), where σ 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_1 and PC_2; $\mu \pm 3\sigma$ on PC_1 or PC_2; etc.).
- *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).
- *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).
- *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.

- *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.

Visualization

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

As can be seen in Fig. 3, the primary PopMLvis dashboard can be categorized into four panels:

- *Choose data:* We designed this panel in an interactive manner with the user. The users has the choice to either upload their dataset, or utilize the provided datasets.

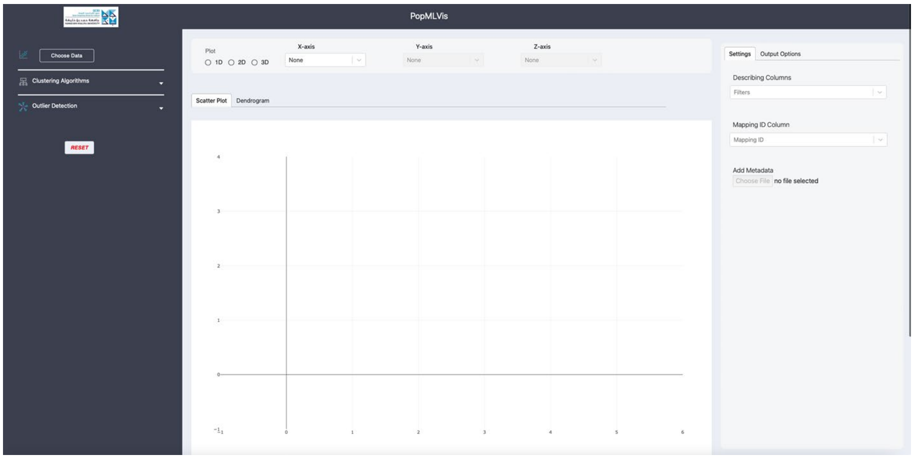
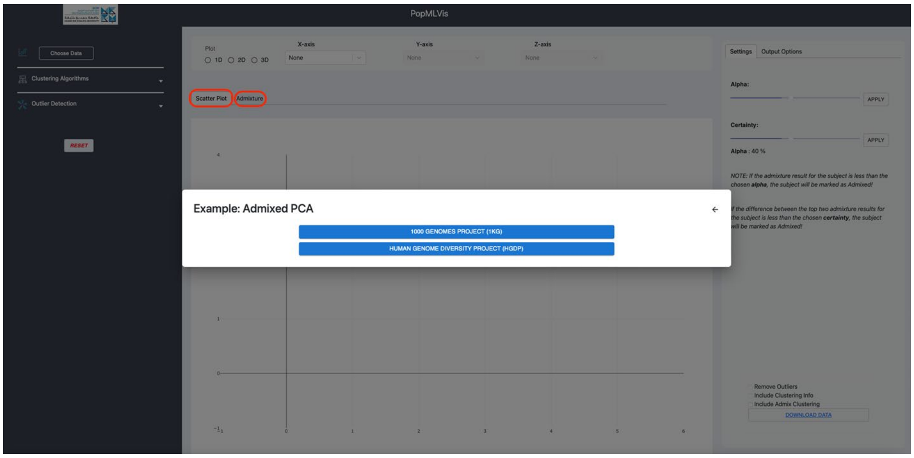


Fig. 3 The main window of PopMLvis



Alpha : 40 %

*NOTE: If the admixture result for the subject is less than the chosen **alpha**, the subject will be marked as Admixed!*

*If the difference between the top two admixture results for the subject is less than the chosen **certainty**, the subject will be marked as Admixed!*

Fig. 4 Input panel of PopMLvis

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 Fig. 4).

- *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, .. p_n where n is the number of clusters.
    - We say a subject s_i belongs to cluster k, if
    - $p_k = \max (p_1, p_2, \ldots p_k, \ldots, p_n)$ and $p_k > \text{alpha}$; or
    - $p_k = \max (p_1, p_2, \ldots p_k, \ldots, p_n)$, and $p_k - p_j > \text{certainty}$,

    where $p_j = \max (p_1, p_2, \ldots p_{k-1}, p_{k+1}, \ldots, p_n)$, i.e. the second largest probability.

    - Here, alpha is a regularization parameter that penalizes intermediate ancestry proportions, whereas certianty is how certain an individual belongs to specific ancestry.

    - The visualization result is a scatter plot and a bar plot (see Fig. 5).

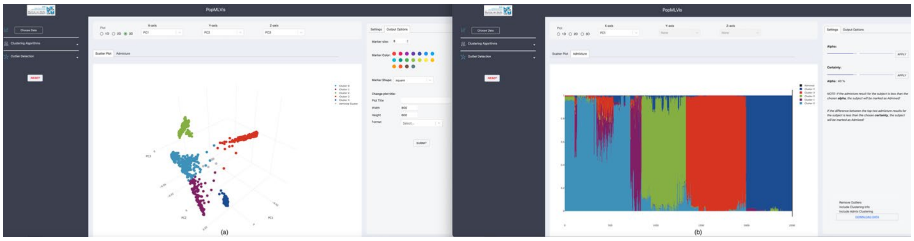


Fig. 5 **a** The effect of alpha and certainty in PCA scatter plot and, **b** The effect of alpha and certainty in Admixture bar plot

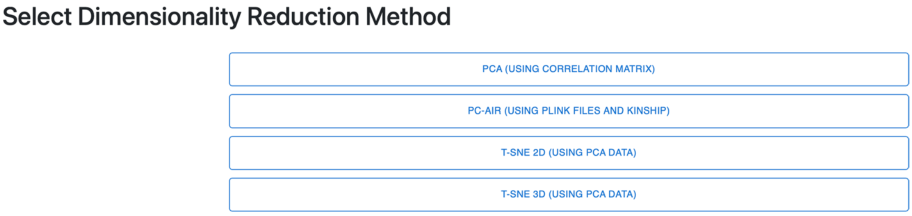

Select Dimensionality Reduction Method
PCA (USING CORRELATION MATRIX)
PC-AIR (USING PLINK FILES AND KINSHIP)
T-SNE 2D (USING PCA DATA)
T-SNE 3D (USING PCA DATA)

Fig. 6 Four different examples of dimensionality reduction algorithms. **a** PCA (expected input: Correlation Matrix/Genetic Relationship Matrix). **b** PC-Air (expected input: PLINK files, optional: Kinship). **c** t-SNE 2D (using PCA data). **d** t-SNE 3D (using PCA data)

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 Fig. 5a).

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 Fig. 5b).

- Dimensionality Reduction: If the input data consists of a large number of features, PopMLvis is compatible with performing dimensionality reduction algorithms. As shown in Fig. 6, PopMLvis supports four 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.
  - PCA

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

- PC-Air:

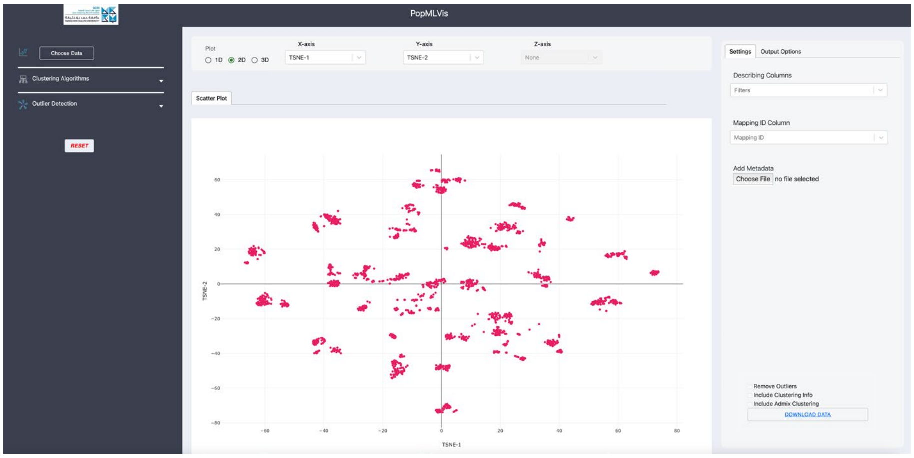
  a. Expected input:

  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.


Plot ○ 1D ● 2D ○ 3D
X-axis TSNE-1
Y-axis TSNE-2
Z-axis None

Describing Columns Filters
Mapping ID Column Mapping ID
Add Metadata Choose File no file selected

Remove Outliers
Include Clustering Info
Include Admix Clustering

Fig. 7: 2D dimensionality reduction of 1000 Genomes Project.

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

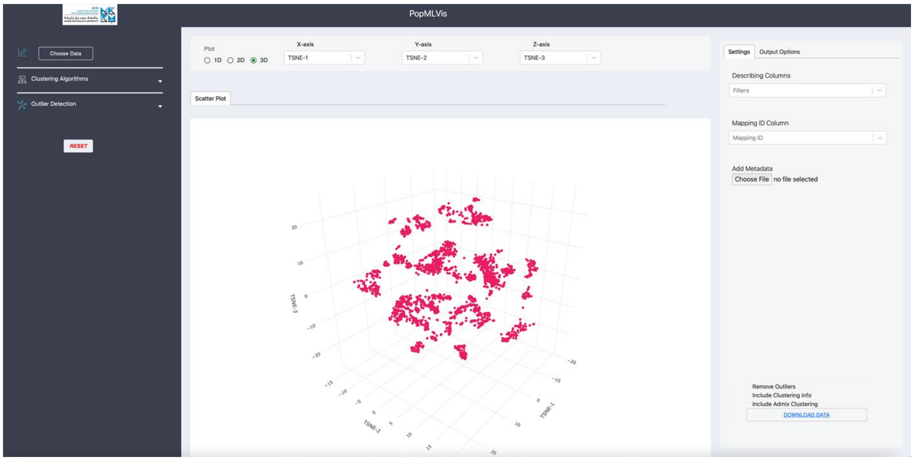
- t-SNE 2D:

a. Expected input: PCA data or Correlation Matrix/GRM.
b. Output: 2D data, see Fig. 7.

- t-SNE 3D

a. Expected input: PCA data or Correlation Matrix/GRM.
b. Output: 3D data, see Fig. 8.

  - *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 Fig. 9).



Plot ○ 1D ○ 2D ● 3D

X-axis TSNE-1

Y-axis TSNE-2

Z-axis TSNE-3

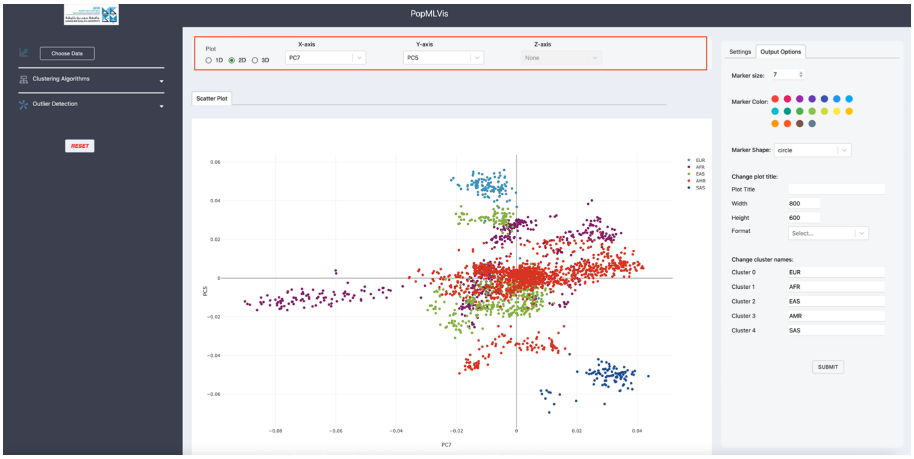
Describing Columns Filters

Mapping ID Column Mapping ID

Add Metadata Choose File no file selected

□ Remove Outliers
□ Include Clustering Info
□ Include Admix Clustering

Fig. 8: 3D dimensionality reduction of 1000 Genomes Project



Plot ○ 1D ● 2D ○ 3D

X-axis PC7

Y-axis PC5

Z-axis None

Marker size: 7

Marker Color:

Marker Shape: circle

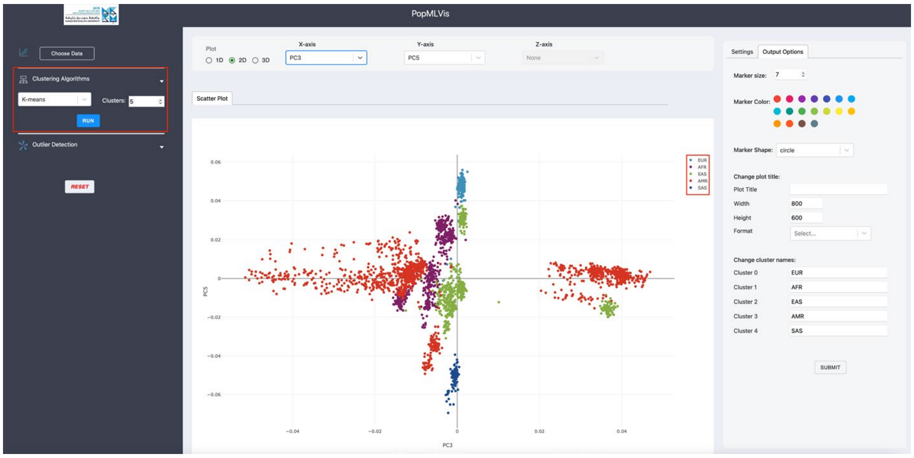
Change plot title:

Plot Title	
Width	800
Height	600
Format	Select...

Change cluster names:

Cluster 0	EUR
Cluster 1	AFR
Cluster 2	EAS
Cluster 3	AMR
Cluster 4	SAS

Fig. 9 Viewing the principal components



Plot ○ 1D ● 2D ○ 3D

X-axis PC3

Y-axis PC5

Z-axis None

K-means

Clusters: 5

Marker size: 7

Marker Shape: circle

Change plot title:

Plot Title

Width 800

Height 600

Format Select...

Change cluster names:

Cluster 0 EUR

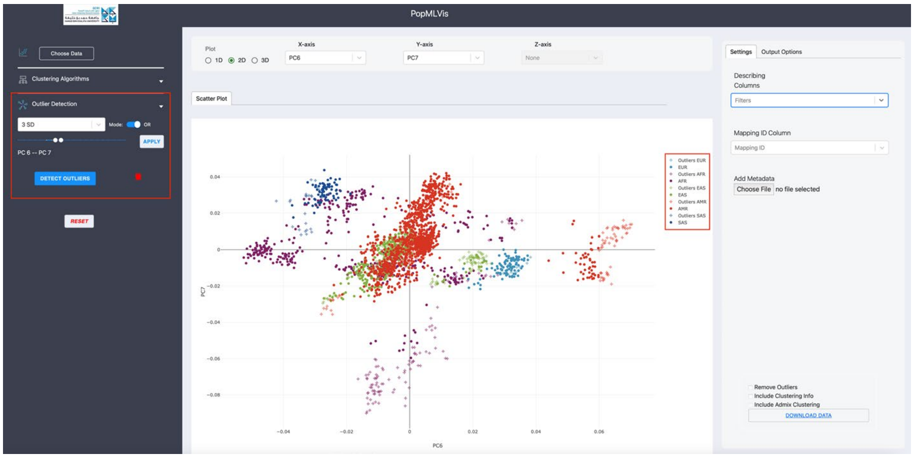
Cluster 1 AFR

Cluster 2 EAS

Cluster 3 AMR

Cluster 4 SAS

Fig. 10 Clustering algorithms panel



Plot ○ 1D ● 2D ○ 3D

X-axis PC6

Y-axis PC7

Z-axis None

3 SD

Mode: OR

PC 6 -- PC 7

Describing Columns Filters

Mapping ID Column Mapping ID

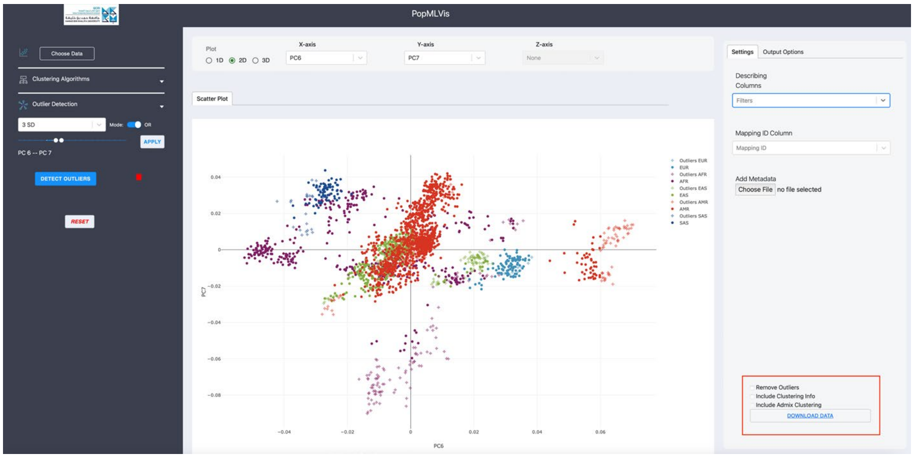
Add Metadata Choose File no file selected

□ Remove Outliers

□ Include Clustering Info

□ Include Admix Clustering

Fig. 11 Outlier detection panel



Plot ○ 1D ● 2D ○ 3D

X-axis PC6

Y-axis PC7

Z-axis None

3 SD

Mode: OR

PC 6 -- PC 7

Describing Columns Filters

Mapping ID Column Mapping ID

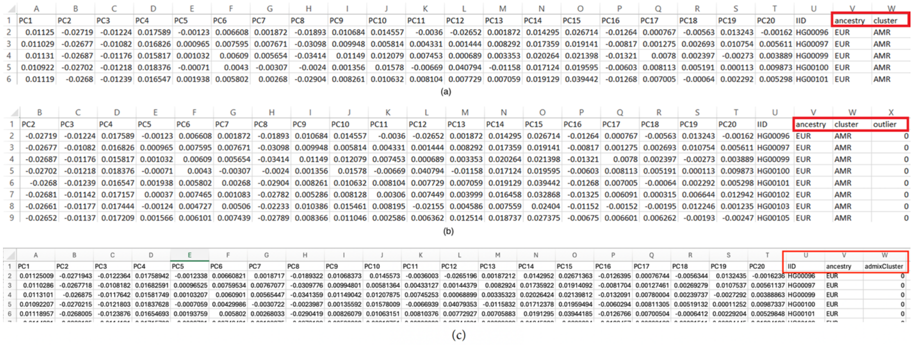
Add Metadata Choose File no file selected

□ Remove Outliers

□ Include Clustering Info

□ Include Admix Clustering

Fig. 12 Download data panel



	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	0.003353	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

(a)

	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	0.003353	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)

	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	admixCluster
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	0.02671363	-0.0126395	0.00076744	-0.0056344	0.0132435	-0.0016236	HG00096	EUR	0
3	0.0110286	-0.0267718	-0.0108182	0.01682591	0.00096525	0.00759534	0.00767077	-0.0309776	0.00994801	0.00581364	0.00433127	0.00144379	0.0082924	0.01735922	0.01914092	-0.0081704	0.00127461	0.00269279	0.0107537	0.00561137	HG00097	EUR	0
4	0.0113101	-0.026875	-0.0117642	0.01581749	0.00103207	0.0060901	0.00565447	-0.0341359	0.01149042	0.01207875	0.00745253	0.00068899	0.00335323	0.02026424	0.02139812	-0.0132091	0.00780004	0.00239737	-0.0027292	0.00388863	HG00099	EUR	0
5	0.01092207	-0.0270215	-0.0121803	0.01837628	-0.0007059	0.00429986	-0.0030722	-0.0023987	0.00135592	0.01578009	-0.0066939	0.04079353	-0.0115832	0.01712378	0.01959494	-0.0060294	0.00811305	0.00519132	0.00011252	0.00987337	HG00100	EUR	0
6	0.01118957	-0.0268005	-0.0123876	0.01654693	0.00193759	0.005802	0.00268033	-0.0290419	0.00826079	0.01063151	0.00810376	0.00772927	0.00705883	0.0191295	0.03944185	-0.0126766	0.00700504	-0.0006412	0.00229204	0.00529848	HG00101	EUR	0

(c)

Fig. 13 **a** Data excluding outliers. **b** Including cluster information, and **c** Including Admix clustering information

- *Clustering panel:* This panel allows the user to apply a variety of clustering algorithms to the uploaded dataset and visualizes the results spontaneously. For each algorithm, the user can set the parameters such as the number of clusters (see Fig. 10).
- *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. Also, the user can select if he/she wants to do "AND" or "OR" operations when there is more than one principal component. Moreover, the user has to decide the deviation from the mean (μ) to be flagged as an outlier (e.g., 1SD, 2SD, etc.) as shown in Fig. 11.

Exporting outputs

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

- *Removing outliers*: As shown in Fig. 13a, the user can exclude outlier samples.
- *Include clustering information:* As depicted in Fig. 13b, the user can add clustering information to the exported dataset.
- *Include Admix clustering:* As depicted in Fig. 13c, the user can add Admix clustering data to the exported dataset.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05908-1.

Additional file 1.

Author contributions

M.E. and M.S. conceived the study. M.E., K.I. and K. K. built the system. M.S. provided the 1000 Genomes Project and Human Genome Diversity Project datasets. M.E., K. I. and M.S. performed the analysis. M.E., K.I. and M.S. drafted the manuscript. M.E., K.I., K.K., and M.S. revised the manuscript and approved it. All authors tested the system. M.E. and M.S. coordinated the study.

Funding
Authors are funded by the Qatar Computing Research Institute. No external funding was received.

Data availability
PopMLvis is freely available and can be accessed online at https://popmlvis.qcri.org. For data privacy issues, the source code of PopMLvis can be downloaded and installed from the GitHub repository https://github.com/qcri/QCAI-PopMLVis. The user manual and tutorial can be found online at the PopMLvis website.

Declarations

Ethics approval and consent to participate
N/A.

Availability and requirements
Project name: PopMLvis, Project home page: https://popmlvis.qcri.org, Operating system(s): Platform independent (PopMLvis is supported on Windows, macOS, and Linux operating systems), Programming language: Python, C++, R, and Flask, Other requirements: Python 3.9 or higher, License: MIT license, Any restrictions to use by non-academics: licence needed.

Competing interests
All authors declare no conflict of interest.

Received: 26 September 2023 Accepted: 20 August 2024
Published online: 11 September 2024

References

1. McCandlish DM, Stoltzfus A. Modeling evolution using the probability of fixation: history and implications. Q Rev Biol. 2014;89(3):225–52.
2. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655–64.
3. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945–59.
4. Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics. 2014;197(2):573–89.
5. Li YL, Liu JX. StructureSelector: a web-based software to select and visualize the optimal number of clusters using multiple methods. Mol Ecol Resour. 2018;18(1):176–7.
6. Metsalu T, Vilo J. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Res. 2015;43(W1):W566–70.
7. Conomos MP, Miller MB, Thornton TA. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet Epidemiol. 2015;39(4):276–93.
8. Gogarten SM, Sofer T, Chen H, Yu C, Brody JA, Thornton TA, Rice KM, Conomos MP. Genetic association testing using the GENESIS R/Bioconductor package. Bioinformatics. 2019;35(24):5346–8.
9. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.