User manual: Group 4

Faheem Moolla   Tofique Rawoot   Keanu Teixeira

11 May 2018

1 Introduction

This manual describes Genesis, a re-implemented version of the original programme for scientists to generate Admixture and PCA graphs based on data inputted. This re-implemented version of Genesis is different from the original programme due to the fact that it was written in the coding language Python 3. It also lacks certain customisation features of the graphs that the original programme had.

Genesis makes use of Genomic data and uses this to explore structural patterns in population. The programme helps the user to gain an understanding on population lineage, gain an understanding on the abnormalities in the data, as well as helps the user to manage Genome-Wide Association Study. Principal Component Analysis is a mathematical function that is used to analyse this genotype information. Admixture graphs are used to examine populations of different/mixed origins and calculate the ratio of these ancestries.

Git repository

The Git repository for our project is `https://github.com/keanutex/firstProject` In this repository the project folder 'Genesis' can be found, which contains the scripts for running the program, with a sub-folder 'Examples' which contains examples of input files that can be used with the Genesis program. The folder 'Timesheet' contains the a Excel Spreadsheet of the minutes that each group member put into creating the programme. The final folder 'Documentation' contains the user and technical manual for the programme.

2 User manual

2.1 Assumptions

This manual assumes that the user of the programme is familiar with Admixture and PCA analysis, tools such as admixture or Eigenstrat, and possibly the use of the original Genesis programme. Error handling within the current programme is not fully implemented, so the user will have to have an understanding on which input is required for each field and which columns from the input files are valid. It is also assumed that the user has Python3 installed.

2.2 Installing Genesis

To install Genesis navigate to the GitHub URL listed above and download the entire project. Python3 must be installed on the users computer. This is usually accompanied with Linux systems, however if Python3 is not found on the PC, use:

```
$ sudo apt-get install python3.6
```

to install Python3.

Different modules need to be installed to run the programme, namely matplotlib and tkinter. To install matplotlib:

```
$ sudo apt-get install python3-matplotlib
```

To install tkinter:

```
$ sudo apt-get install python-tk
```

2.3 Running Genesis

Genesis is currently functional on Linux. To run the programme navigate to the folder downloaded from GitHub and go into the 'Genesis' folder. Then in the console type:

```
$ python3 UI.py
```

This will launch the programme and you will currently be running the latest Genesis programme through the console using the Python3 compiler.

2.4 Input Data

2.4.1 Admixture Plots

Genesis requires two input files and possibly a third (optional) file in order to generate the required Admixture plot:

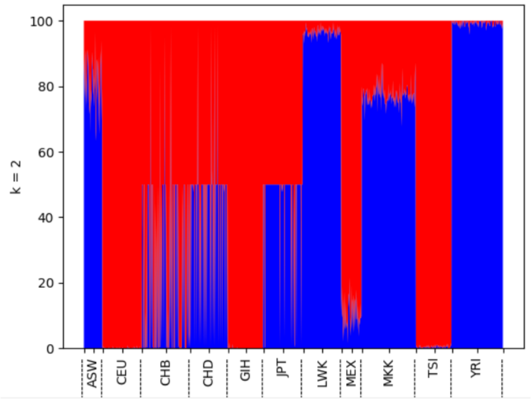
1. A Fam File - The first two columns of the fam file describe and identify the individuals. The first column identifies the family, while the second identifies the individual in question. Every line in the file describes the individual in the same chosen line of the admixture file. Although other columns usually also present in the fam files, these are not required by Genesis.

2. An Admixture Data File - These files are usually produced by programmes such as CLUMPP or Admixture. These files contain a k value that Genesis automatically calculates, which describe the amount of columns in the file. The columns in this file describe the ancestral proportions of each individual.

3. A Phenotype File - This is an optional file. As with the fam file, the first two columns uniquely identify each individual (with a family identifier and an individual identifier).It is essential that the identifiers in the first two columns of the phenotype file match those in the fam file, although it does not have to be in order. If subsequent individuals are produced in the phenotype file, it will not make a difference to Genesis. Every other column in the phenotype file should have other unique identifiers such as sex, language, population group etc.

4. Phenotype Column - Specifying the amount of columns to be used from the phenotype file

Figure 1: Example of an Admixture plot generated by Genesis



2.4.2 PCA Plots

Genesis requires one input file and an optional second file in order to generate the PCA plot:

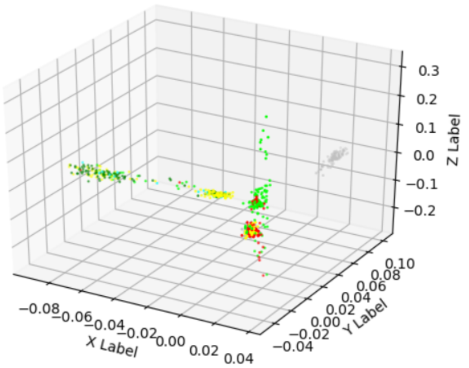
1. A PCA data file - Eigen-Decomposition is used. This takes the distance matrix and produces eigenvalues and eigenvectors from it. The eigenvalues are used for the dimensions, while the eigenvectors give the positions of each individual in that space.

2. A phenotype file - the use of this file is the same as with the Admixture plots.

3. Four Eigenvector Columns - Specifying which eigenvector columns to be used from the data file

4. Phenotype Column - Specifying the amount of columns to be used from the phenotype file

Figure 2: Example of a PCA plot generated by Genesis



2.5 User Interface

Once the Genesis programme starts running, a UI will appear. From here, the user has the option to create a new Admixture plot or a new PCA plot, as well as the options to load and old plot or save the current plot.

2.5.1 Admixture Plot

Admixture plots are created by inputting an admixture data file, a fam file, as well as an optional third phenotype file. To input these files, click **New Admixture Project**. On the screen that pops us, click **Import Data File**. Navigate to where this admixture file is and select it. Next, click **Import Fam File**. Do the same as with the Data File. A option is then given to **Import Phenotype File**. If this is required, click the box and navigate again to where the phenotype file is and select it. In the text box specify the column needed from the Phenotype file. Then click **Set Column**. Once all this is done, click **Finish** at the bottom of the pop-up box.

2.5.2 PCA Plot

PCA Plots are established by inputting a Data File as well as a Phenotype File. To input these files, click **New PCA Project**. On the screen that pops us, click **Import Data File**. Navigate to where this PCA file is and select it. Next, click **Import Phenotype File**. Do the same as with the data file. The user is then prompted to input the Eigenvector columns as well as the phenotype columns in use. Once these are set, click **Set Columns** and then click **Finish**.

2.6 Save PCA Project

This will open a new window that allows the user to save the current PCA project that is open. The different input file's directories are saved within a text file as well as the Eigenvector columns and Phenotype Column the user inputted. This save file is saved within the project directory.

2.7 Load PCA Project

This creates a file explorer window which allows the user to navigate through their PC and select a save file that was previously created, which contains the directories and columns of the previously saved project. In this current implementation, the input files used for the project cannot be moved, or else the project will not properly be loaded.

2.8 Save Admixture Project

Similar to the 'Save PCA Project', this button will create a new window that allows the user to save their current, open Admixture Project to a text file. The input files directories and the Phenotype column are saved.

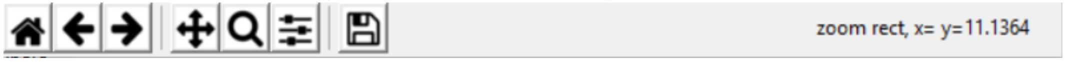
2.9 Load Admixture Project

Similar to the 'Load PCA Project', this button will open a file navigator to selected an Admixture text save file to be loaded. In this current implementation, the input files used for the project cannot be moved, or else the project will not properly be loaded.

2.10 Using the Admixture/PCA Plots

Once the plots are generated, they can be interacted with. A pop-up showing the plot will appear. The graphs contain a tool-bar at the bottom from which to interact with the graphs.

Figure 3: Tool-bar for Graph Interaction



zoom rect, x= y=11.1364

- The first button on the tool-bar is a home button, where the graph can be reset to its original plot after being manipulated.
- The second and third buttons allow for the user to go back or forth respectively to the previous or next views of the graph after being interacted with.
- The fourth button allows for the screen to be panned according to the axes of the graph. This allows for the graph to be viewed from different angles.
- The following button with the magnifying glass zooms the graph in, in the shape of a rectangle, depending on where the user clicks on the graph.
- The following button on the tool-bar gives the user the freedom to configure subplots. When this button is clicked, another pop-up box will appear which will allow the user to change the size and spacing of the graphs by clicking on the bars. When the reset button is pressed, these changes will be applied.
- The last button allows the user to save the plot in a chosen directory as a PNG.

3 Technical manual

3.1 Overview of Design

The main design pattern used for the design of Genesis was the MVC model.

3.1.1 Model

These were used to implement the logic for the programmes' domain. The models used were separated into two classes; the Admixture class, called admixPlotter, and the PCA class, called PCAPlotter. The classes contained functions which created the graphs and updated graphs based on user interaction (from the Controller). The functions were used to receive data from the Controller, such as which files were to be uploaded and the columns of the files that were used. The Model then took these files and data for the columns and sent it to the functions within the PCA and Admixture classes to generate a plot based on this.

3.1.2 View

The View is used to display the information to the user of the programme. The UI class is the core class of the View, as this displays the programme data and information to the user. A UI was generated which allows the user to establish new PCA or Admixture graphs, and then interact with these graphs as described in Section 2.10. This is all created from the model data, that is sent to the view as a result of the user input in the Controller. All the functions in the UI class are used to display a UI, which then allows the user to give input (files and columns). The functions then receive data from the Model (PCA and Admixture classes) and then display this as a View in the form of a graph.

3.1.3 Controller

The Controller handles user interaction, selects a view to render from the UI, and works with the model. The Controller in the case of Genesis are the functions in the UI, which receives the input from the user (after the user upload the files) and then thus manipulates these and sends them to the Model. The selection of columns used are also data taken from the user and then sent to the Model. The View works hand-in-hand with the Controller as the user input is taken from the UI.

3.2 Key classes

3.2.1 UI Class

This section of the programme handles the UI and could also be considered the 'Main' class of the program, in that the instances of the PCA and Admixture classes are instantiated in this class, and all functions are subsequently run through this class. The class handles the UI of the program, displaying labels, buttons, entries, etc, using the tKinter module of python. When the programme is run, a main Window is created which contains most of the functionality of the programme, with button-presses allowing for added functionality. Within this class, data is loaded into the Admixture and PCA instantiated objects (from user input) and graphs are subsequently plotted in a separate window which the user can interact with. Projects are also saved and loaded from the interface into text files.

3.2.2 Admixture Class

The admixPlotter class is used to plot admixture charts from data obtained in the UI class. This class has four core functions, namely: ReadFamFile function, ReadDataFile function, ReadPhenoFile function, and the PlotGraph function.

The readFamFile stores the data of the fam file in a list, and is only utilised in a sub function to determine if the number of lines in all three files are equal in length and that there are no inconsistencies. The ReadDataFile function stores the data in a list of columns, and calculates the percentages to be used in the plotting of the Admixture graph. The readPhenoFile function, accesses the column of data selected by the user in the UI, which determines the categories and allocates colours for them. The PlotGraph function is used to plot the graph from the data obtained including labels and colours.

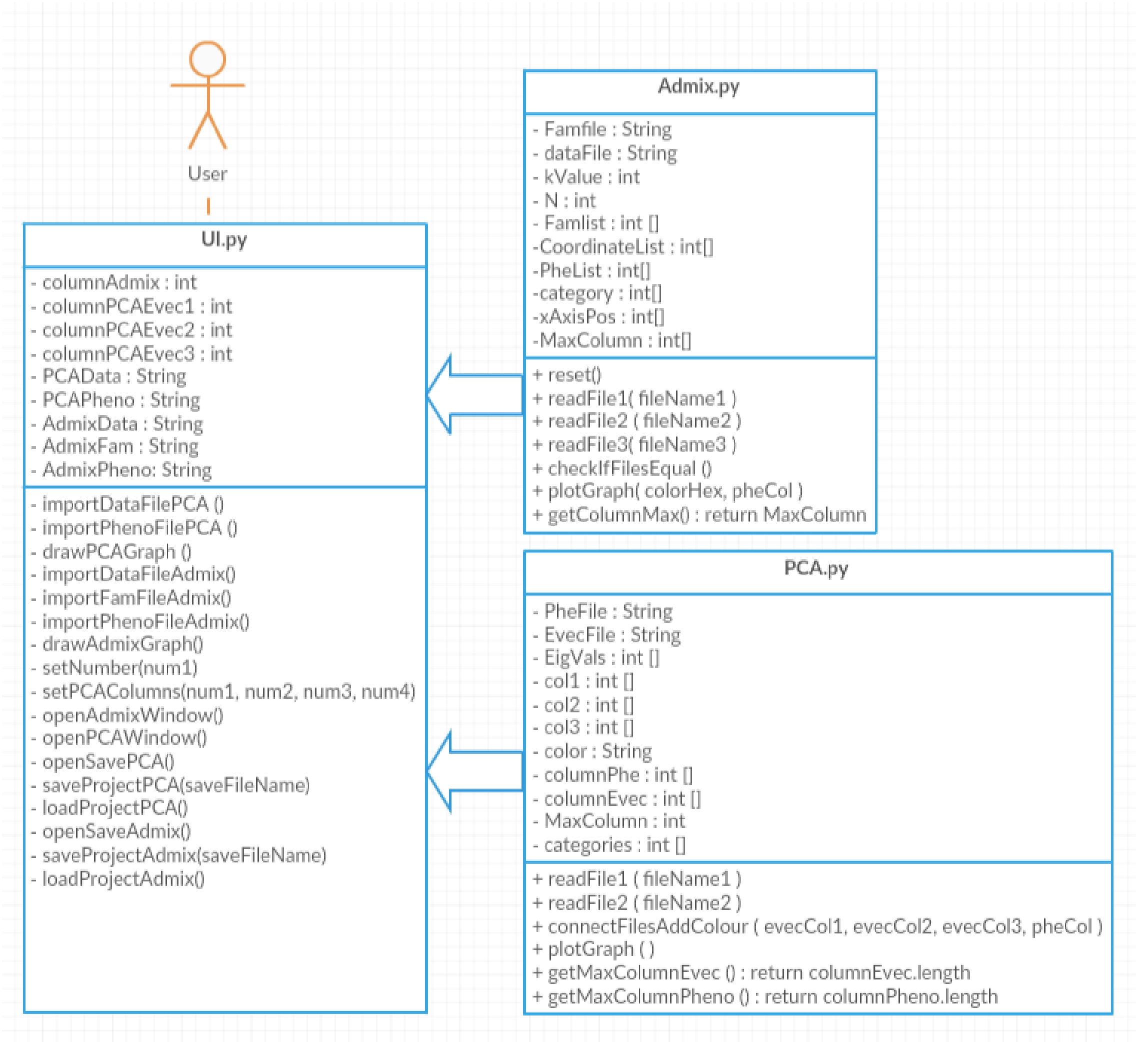
3.2.3 PCA Class

The PCAPlotter class is used to chart PCA graphs using functions in the class and data attained from the UI class. This class uses similar functions to read the data from the Evec and Phenotype files and store the data in lists. The class uses a connectFilesAddColour function to compare the Eigen-values of both files and relate them to one another. The function also catagorises the data determined by the phenotype file and allocates colours corresponding to the data categories. The PlotGraph function will plot a scatter-plot from the results obtained from the above mentioned functions and provide axes labels and a descriptive legend.

3.3 UML Diagram

The UML diagram to describe the use of classes can be seen in Figure 4.

Figure 4: UML Diagram for Genesis



Group document (physical hand-in) signed by:

Faheem Moolla:

Keanu Teixeira:

Tofique Rawoot: