User manual: Group 1

Seale Rapolai (1098005) Phatho Pukwana (1388857) Cedrick Platt (1500728)

11 May 2018

1 Introduction

This is the user manual for the scientific programme Genesis. This programme is used to create Principal Component Analysis and Admixture plots. This manual will breakdown how to install and use Genesis for users as well as describe technical aspects that they may require further understanding to use Genesis.

Git repository

The Git repository for our project is `https://github.com/Seal12/ELEN3020_ppsd_cps/`

2 User manual

2.1 Installing Genesis

- Clone or Download the git repository above.
- Genesis requires Python 3. Python 3 can be found at: `https://www.python.org/downloads/`.
- Genesis requires numPy, wxPython and matplotlib modules be installed, these can be acquired using pip.

2.2 Running Genesis

Open the location where the git was cloned to in terminal. Thereafter do the following to execute Genesis:

- Linux - Type *python3 startup.py*
- MacOs - Type *python3 startup.py*
- Windows - Type *python startup.py*

2.3 Plotting a Principal Component Analysis Plot

2.3.1 Inputting Data for the Principal Component Analysis Plot

To plot a Principal Component Analysis plot there are two mandatory files you need:

- An evec (Data) file that ends in *.evec*, this is the file that contains the principle component analysis data.
- A phenotype file that ends in *.phe*, this is the file that contains subject identification data.

With these two files, we can plot a new Principal Component Analysis plot. With Genesis open click on the menu item **Plot**, and a drop down menu will appear with two options:

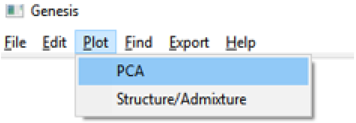


Figure 1: The drop down menu when Plot is clicked on.

Click on PCA, and you will be presented with the following window:

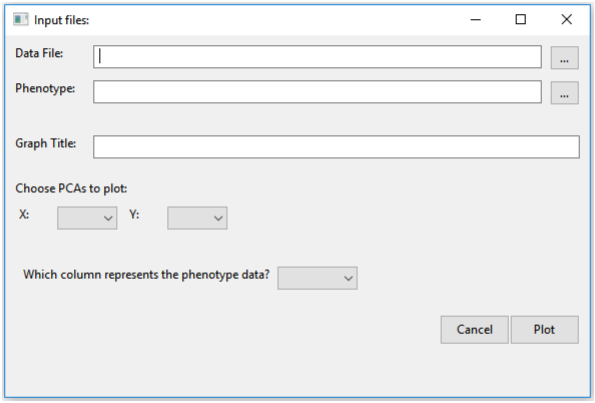


Figure 2: The plot window that appears when PCA is clicked on.

On this window, *Figure 2*, you can input the data files as follows:

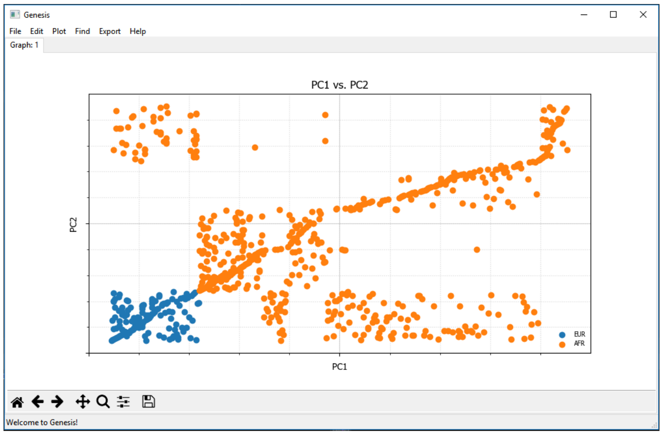
- Click the button with ellipses next to the respective file you want to input.
- A file explorer will appear, navigate to the file you want add.
- The **Data File** takes in the *.evec* file, and the **Phenotype** takes in the *.phe* file.
- The **Graph Title** input takes in a title that will be displayed on the graph. If left left blank, a default title will be used.

With these two files we can now move onto plotting the data inside these files.

2.3.2 Plotting of the Data

From *Figure 2*, we can see we're presented with two axes to choose from under the heading **Choose PCAs to plot**. Clicking on the drop down tab on **X** will allow you to choose which principal component to use as the X axis. Similarly the **Y** drop down tab allows you set the Y axis. Lastly, we need to set which column from our phenotype data file do we want to represent, we do this with the last drop down tab below the the three axes. Once this is set, click the **Plot** button and the plot will be rendered.

2.3.3 The Principal Analysis Plot



Welcome to Genesis!

Figure 3: The newly plotted PCA plot as it appears in Genesis.

As seen in *Figure 3*, this is how a PCA plot appears once the **Plot** button is clicked. In this figure, PC1 has been plotted against PC2, the data for this is included in the *example* file of the git, under *PCA*. *Please refer to section 2.6 for the tool bar.*

2.4 Interacting with the PCA Plot

At this stage of development Genesis currently only supports these interactions by Right clicking on the plot allows you to do the following:

- Set the title of the Plot
- Change the shape and size of the individual groups of the Plot
- Refresh the Plot

2.5 Plotting an Admixture Plot

2.5.1 Inputting Data for the Admixture Plot

To plot an Admixture Plot we need three mandatory files:

- The Q file, this file ends in *.Q.x*, where *x* is the number of ancestries within the file. This file contains the subject ancestry data.

- The Fam file, this file ends in *.fam*, this file contains the subject identification data.
- The Phenotype file, this file ends in *.phe*, this file contains additional identification data for the subjects within the *.fam* file.

With these three files we can plot a new Admixture plot. With Genesis open click on the menu item **Plot**, and a drop down menu will appear with two options, as seen in *Figure 1 on page 2*. On this drop down menu, click **Admixture/Structure**, and you will be presented with the following window:

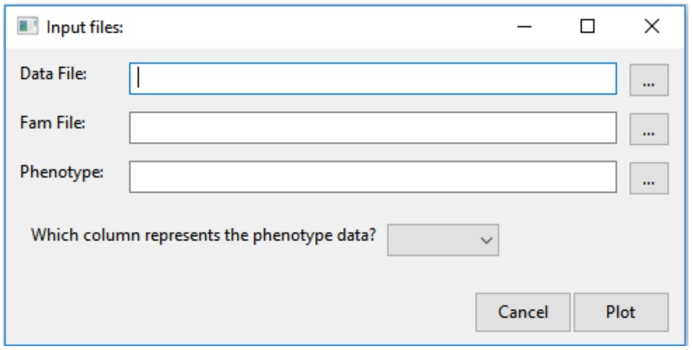


Figure 4: The plot window that appears when Admixture/Structure is clicked on.

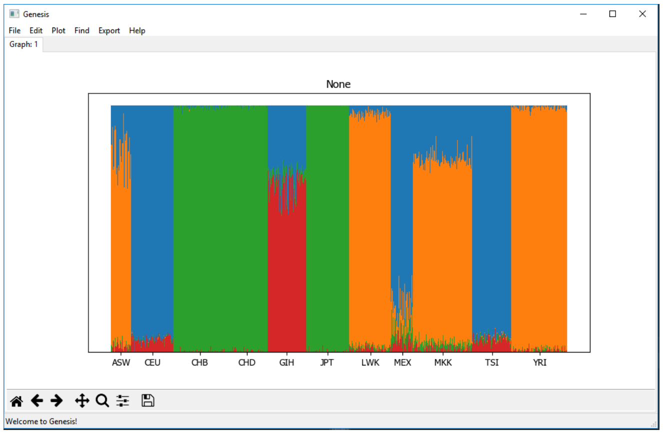
On this window, *Figure 4*, you can input the data files as follows:

- Click the button with ellipses next to the respective file you want to input.
- A file explorer will appear, navigate to the file you want add.
- The **Data File** takes in the *.Q* file, while **Fam File** takes in the *.fam* file, and lastly, the **Phenotype** takes in the *.phe* file.

2.5.2 Plotting of the Data

You select which column in the phenotype file will be used to represent the phenotype data with the drop down tab below the **Phenotype** file input. With the data files set, and the column you want to use, click the **Plot** button to render the plot.

2.5.3 The Admixture Plot



Welcome to Genesis!

Figure 5: The newly plotted Admixture plot as it appears in Genesis.

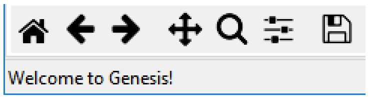
As seen in *Figure 5*, this is how a Admixture plot appears once the **Plot** button is clicked. In this figure, Q4 has been plotted with phenotype column 4 selected, the data for this is included in the *example* file of the git, under *Admix*. *Please refer to section 2.6 for the tool bar.*

2.5.4 Interacting with the Admixture Plot

At the current stage of development, Genesis only supports creating a title for the admixture plot. This is done by right clicking on the plot and selecting **Set Title**.

2.6 Tool Bar

The program currently allows the following control over the view of the graph via these buttons present on both graphs in the form of a tool bar:



Welcome to Genesis!

Figure 6: The buttons to modify the View and export the graph.

- Reset View - clicking the house resets the View to the initial state of the graph.
- Undo Changes to View - clicking on the back arrow allows you to undo the last change to the View.
- Redo Changes to view - clicking on the forward arrow allows you to redo the last change you undid to the View.
- Panning through the graph - clicking on the four-pronged arrow allows you to pan through the graph.

- Zooming into the graph - clicking on the magnifying glass allows you to zoom into the current plot.
- Adjust spaces of the graph - Clicking on the slider allows you to modify the white spaces around the graph.
- Exporting the graph - clicking on the floppy disk allows you to export the graph as an image.

3 Technical manual

3.1 Overview of Design

Genesis was implemented using object-orientated programming principles to separate functionality. Key objects and the basic structure of the program are shown in the diagram below.

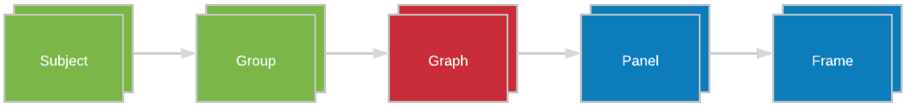


Figure 7: The hierarchy of the program

The specific design methodology used is the Model-View-Controller methodology. Model classes are responsible for how the data is stored and maintained. The Controller classes are responsible for importing and formatting the data they also generate and customise the graphs, the Controller serves as a link between the Model and the View. View classes are responsible for interfacing with the user and providing feedback.

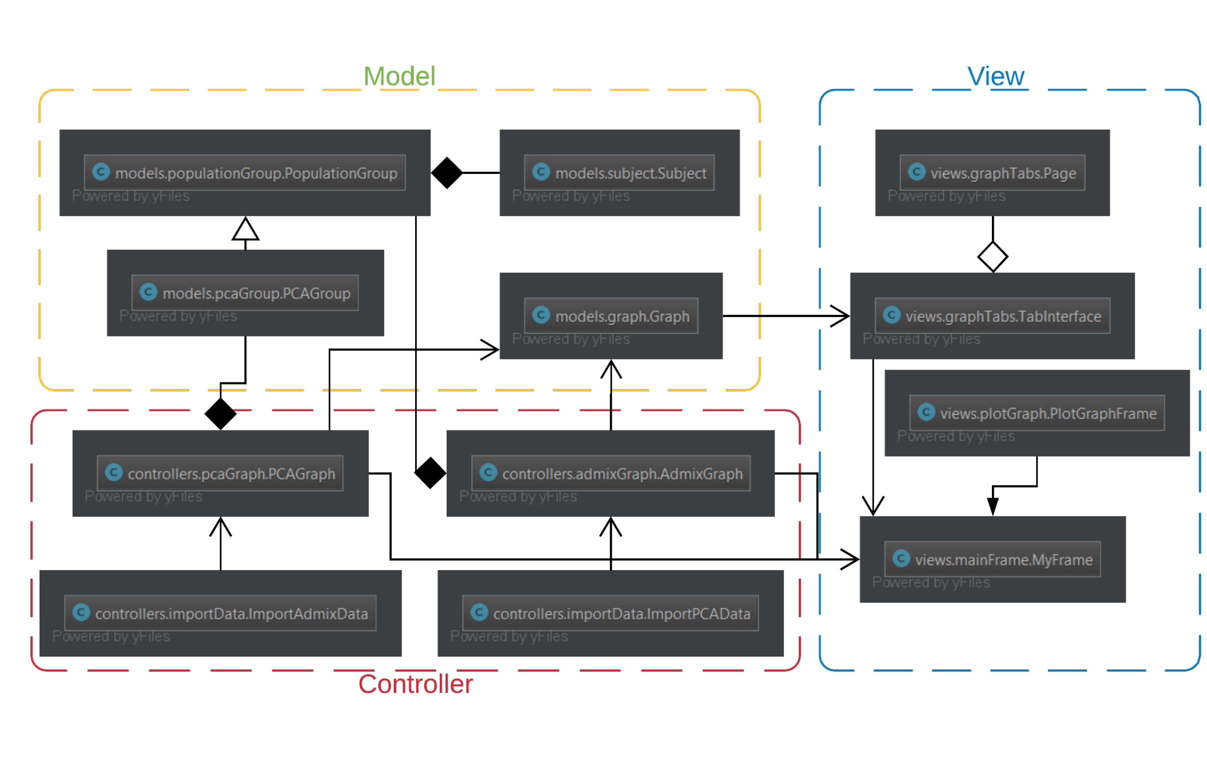


Figure 8: The UML class diagram of Genesis according to the MVC Framework.

Genesis's file directory is also structured after the MVC framework, these folders are named according to the role of the class scripts contained within.

3.2 Design of Key Classes

1. **Model**

   The model directory contains all the classes necessary to store the data to be manipulated, inside this folder are scripts that contain classes, these are key classes related to the model:

   - Subject - this class stores the subject's identification data as well as a list of their ancestry ratios for the Admix or their principal component values for the PCA.
   - PopulationGroup - this class stores the group name and all the subjects' data related to a specific phenotype group as a dictionary
   - PCAGroup - this class inherits from PopulationGroup and the functionality which it adds is group icon information. .
   - Graph - this class stores the graph and the wxPython frame it is attached to and is used to facilitate the tabbed view of each graph in Genesis.

   Below are UML diagrams for these classes, it describes their relation to each other in structuring and storing data:

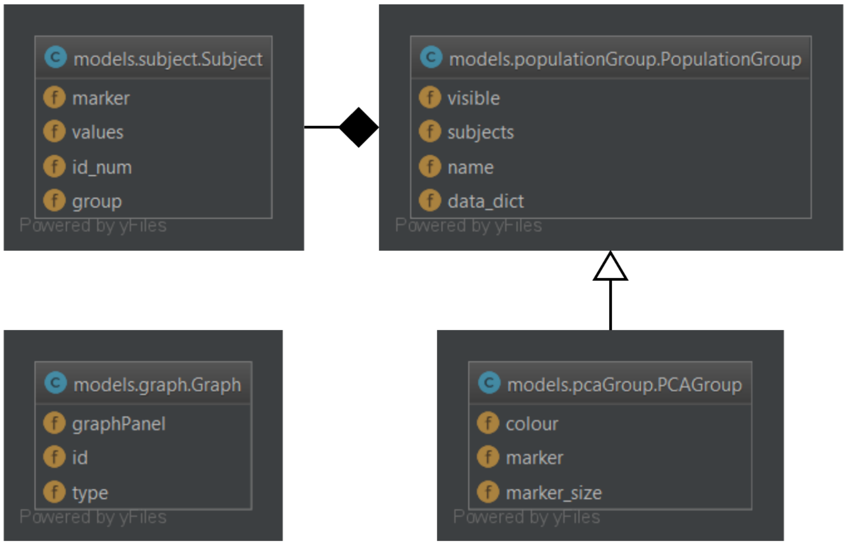


Figure 9: The relationship of the Model classes and their UML diagrams.

2. **Controller**

   The controller directory contains the classes which are responsible for the graph generation and manipulation of the data. The key classes are:

   - ImportAdmixData - this class is responsible for populating the model classes related to the admixture plot like Subject and PopulationGroup, it contains the functions required to format the *.Q* file, *.fam* file and the *.phe* file in a way that the ancestry data will be correctly represented. This includes a normalising function for the ancestry ratios and a function for the data capture of each subject and their population group necessary for the graphing of the admixture plot.
   - ImportPCAData - this class is responsible for populating the model classes relevant to the PCA plot, like Subject and PCAGroup. It contains functions that capture a Subject's identification data like number and their population group to create Subject and PCAGroup Objects as well functions that captures each subject's principal components. This class is used to facilitate the PCA plot.
   - AdmixGraph - this class is responsible for generating the admixture graph, it is a modified bar graph from matplotlib. Each column represents a single subject but the plot is composed of multiple sets of bar graphs plotted on top of each other(one for each ancestry), i.e. in the case of four ancestries four sets of bar graphs will be plotted on the same axes.
   - PCAGraph - this class is responsible for generating the PCA graph, this is a scatter plot from matplotlib. Each group is as a different scatter plot on the same set of axes. Each dot of a group represents an individual and principal components being compared are set to the axises.

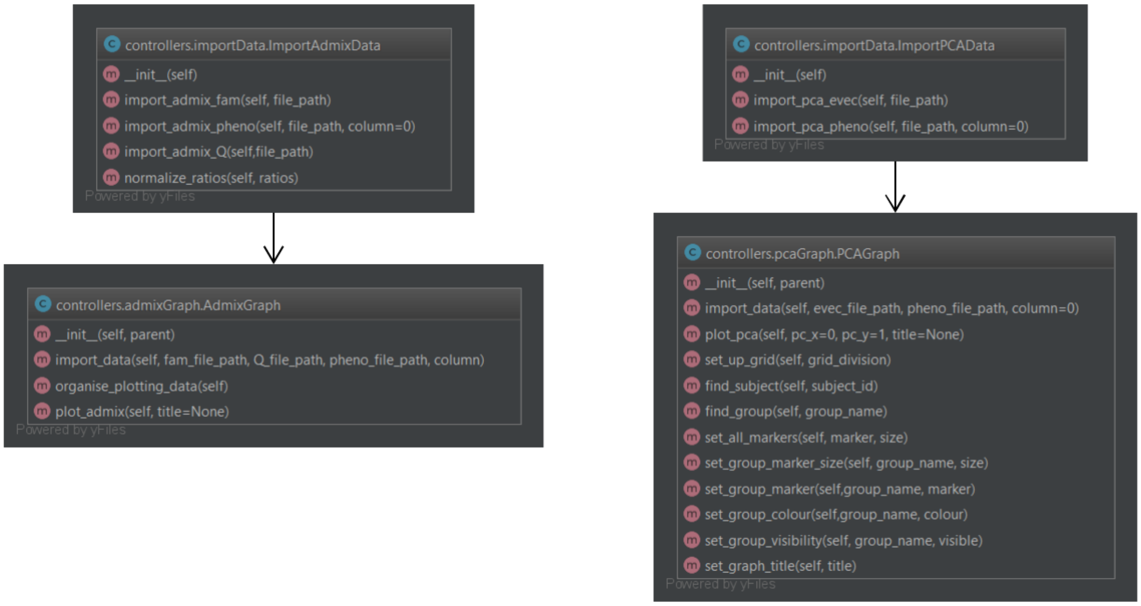

controllers.importData.ImportAdmixData
__init__(self)
import_admix_fam(self, file_path)
import_admix_pheno(self, file_path, column=0)
import_admix_Q(self,file_path)
normalize_ratios(self, ratios)
Powered by yFiles
controllers.admixGraph.AdmixGraph
__init__(self, parent)
import_data(self, fam_file_path, Q_file_path, pheno_file_path, column)
organise_plotting_data(self)
plot_admix(self, title=None)
Powered by yFiles
controllers.importData.ImportPCAData
__init__(self)
import_pca_evec(self, file_path)
import_pca_pheno(self, file_path, column=0)
Powered by yFiles
controllers.pcaGraph.PCAGraph
__init__(self, parent)
import_data(self, evec_file_path, pheno_file_path, column=0)
plot_pca(self, pc_x=0, pc_y=1, title=None)
set_up_grid(self, grid_division)
find_subject(self, subject_id)
find_group(self, group_name)
set_all_markers(self, marker, size)
set_group_marker_size(self, group_name, size)
set_group_marker(self,group_name, marker)
set_group_colour(self,group_name, colour)
set_group_visibility(self, group_name, visible)
set_graph_title(self, title)
Powered by yFiles

Figure 10: The relationship of the Control classes and their UML diagrams.

3. **View**

   Following the Control is the View. From the root directory the folder views contains all the classes that facilitate direct user interaction with the program, these classes are:

   - TabInterface - this class is responsible for creating and removing the tabs that display the different graphs generated on Genesis's main window frame.
   - myFrame - this class is responsible for creating and displaying the different window frames the user will need view to use Genesis, these frames include the menu items and their drop down selection choices.
   - PlotGraphFrame - this class is responsible for creating the window frame that displays after the user has chosen what type of graph to plot, i.e. admixture and prompts the user to select the data via a file browser class and allow them to select which files contain the data they want to plot or alternatively cancel.

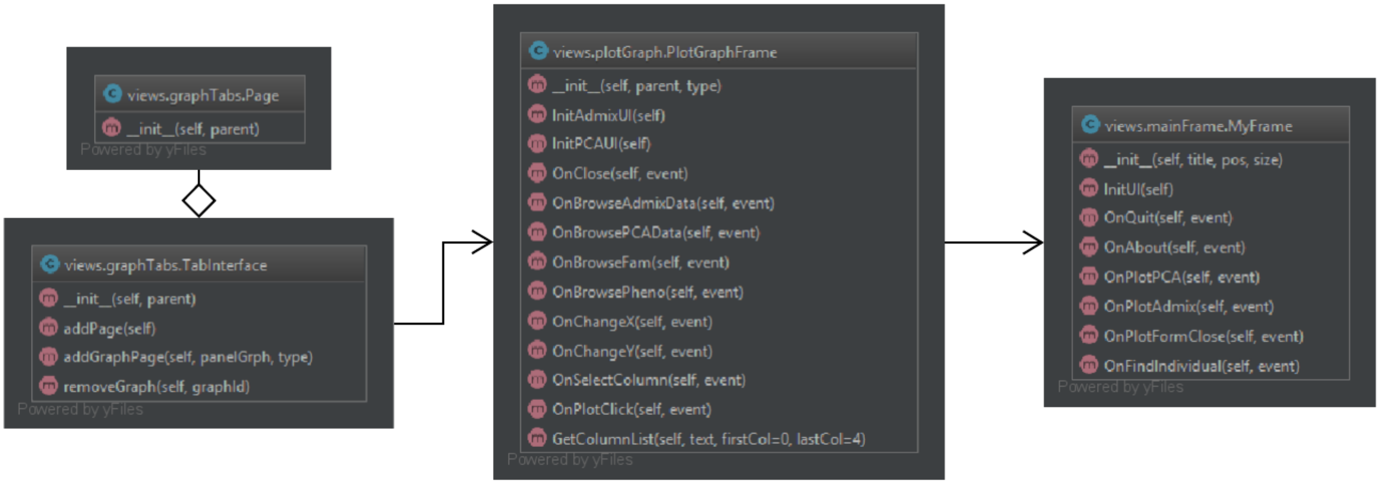


Figure 11: The relationship of the View classes and their UML diagrams.