

The 'Genesis' Manual

Version 0.2.5

Robert Buchmann and Scott Hazelhurst

Copyright © 2015, University of the Witwatersrand, Johannesburg

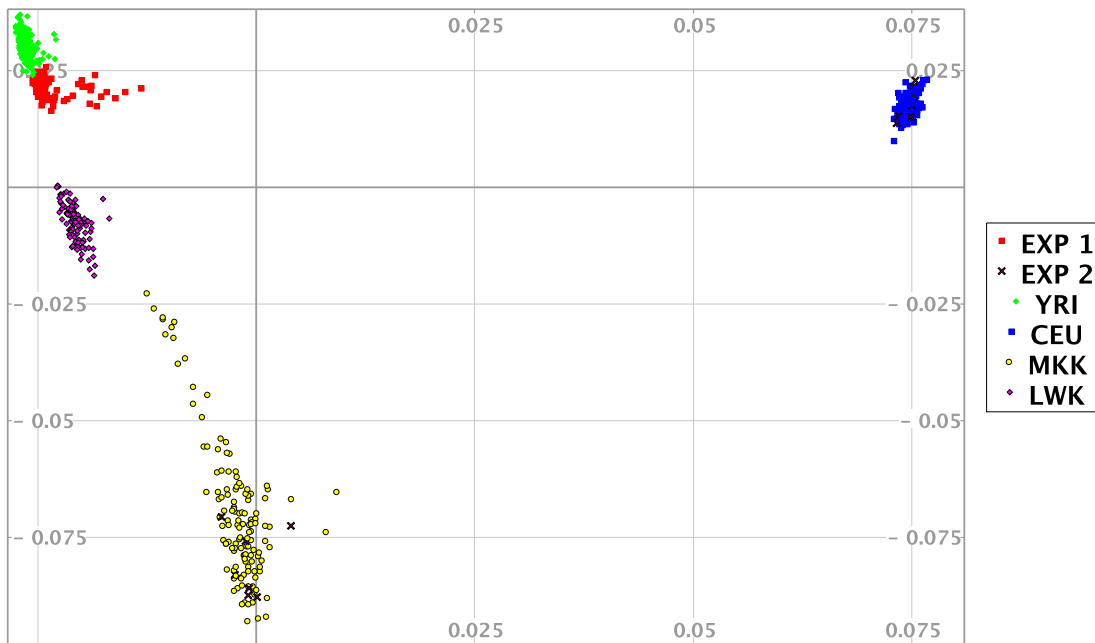
1 Description

1.1 Introduction

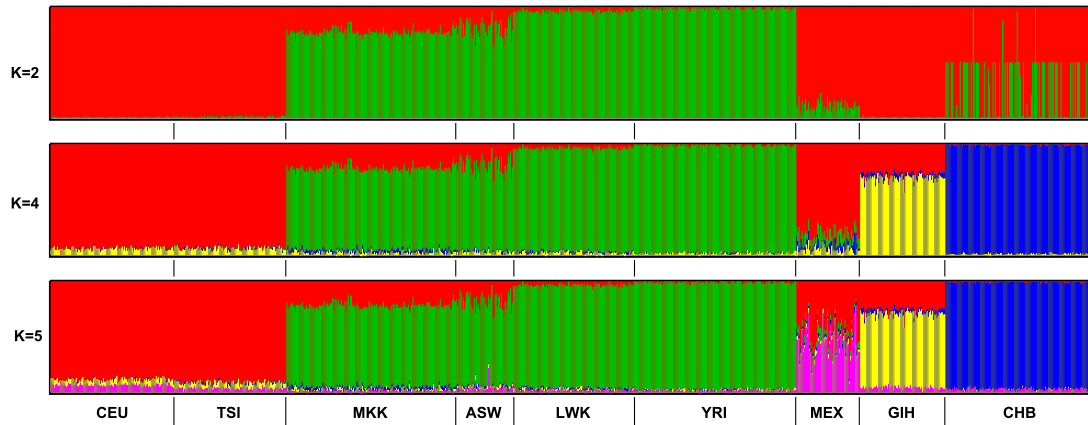
This manual describes **Genesis**, a program created for scientists to generate PCA (Principal Component Analysis) and structure/admixture graphs from data outputted by common tools such as **eigenstrat** [Pritchard et al. 2000] and the **SNPRelate** [Zheng et al 2012] package for PCAs and **Admixture** [Alexander et al. 2009] and **CLUMPP** [Jakobsson and Rosenberg 2007] for admixtures.

Genesis was developed with user-friendliness in mind as other tools can be complex to use and lack certain features. All elements of the graphs that would need to be edited can be done so using a graphical user Interface where the graphs themselves are interactive and different elements can be viewed and changed at the click of the mouse. All this saves the time that scientists would rather be spending doing more important things.

Principal Component Analysis is a mathematical and statistical procedure that can be used to analyse genotype data. The differences between samples' genotype data can be used to project each sample into a p -dimensional space, where the p axes are uncorrelated. For realistic data, typically p is 4 or less and often only the most important two dimensions are used. Programs such as **eigenstrat** produce the the PCs, and **Genesis** produces them. An example is found below:



Admixture mappings are used to analyse populations of mixed ancestry and determine the ratios of proposed different ancestries. These ratios can then displayed in stacked bar graphs as structure/admixture graphs.



1.2 Assumptions

This manual assumes that the reader is familiar with structure and PCA analysis of genotype data, and has used tools such as admixture, Eigenstrat and/or plink [Purcell et al 2007; Purcell and Chang 2014].

1.3 Licence

Genesis was written by Robert W Buchmann, and copyright is owned by the University of the Witwatersrand, Johannesburg. The code is released under the Affero General Public Licence version 3. **Genesis** uses the iText Software Corporation's iText library also released under the Affero General Public Licence.

2 Installing and Running Genesis

2.1 Installing Genesis

The latest code can be found at <http://www.bioinf.wits.ac.za/software/genesis>

Genesis does not special installation and the file can simply be executed using Java. It does, however, **require the Java SE Runtime Environment 1.7 or higher to be installed on the user's system**. The latest version of the Java SE Runtime Environment 1.7 can be found at: <http://www.oracle.com/technetwork/java/javase/downloads/jre7-downloads-1880261.html>

To check if you already have Java installed, you can open your Operating System's command line interface and enter

```
java -version
```

If Java is installed, one line of the response should read

```
Java(TM) SE Runtime Environment (build 1.x.0_*)
```

where x is the version number (and should be 7 or higher) and the * is not important.

Alternatively, Windows users can open Control Panel and then open **Add/Remove Programs** and check the list for Java SE Runtime Environment (in windows XP and earlier) or open **Programs** then **Programs and Features** and if Java x is in the list, then Java is installed where x is the version number (in Windows Vista, 7 or 8).

Mac OSX requires X11 to be installed. This can be downloaded from <http://xquartz.macosforge.org/landing/>.

Genesis is compatible with the 32 and 64 bit varieties of Windows, Mac OSX and Linux.

If you are using a version of Java other than Oracle's, you may need to install the SWT libraries separately. On Ubuntu, installing openjdk, the following should work

```
apt-get install jre-default
```

OR (but NOT both)

```
apt-get install libswt-gtk-3-java openjdk-7-jre-headless
```

2.2 Running Genesis

Windows and Linux If the Java SE Runtime Environment 1.7 is installed, the user should be able to open Genesis by double-clicking the file. If this does not work, the user will have to launch the jar manually through a command line editor (cmd in windows or terminal in linux) using the following command.

```
java -jar Genesis.jar
```

For Linux and MacOSX the recommended configuration is to unzip, cd into the genesis-distrib directory and then say

```
sudo mv Genesis.jar misc/genesis /usr/local/bin
```

```
sudo chmod a+x /usr/local/bin/genesis
```

and then, the program can be run by executing the command **genesis**

OS X: Ensure that the Java SE Runtime Environment 1.7 is installed. In OSX, the user must launch the file through the command line with an extra argument as follows

```
java -XstartOnFirstThread -jar Genesis.jar
```

Wrapper script A simple wrapper script `genesis` is provided to aid the use of the program which works for Linux or MacOS X. Installation of the JAR and wrapper can be done by running `install.sh`. By default, the script installs into `/usr/local/bin`. If you want it elsewhere, specify the directory name, e.g. `./install.sh /Users/scott/bin`. You may need `sudo` privileges.

Memory usage: If you are building up very complex charts (e.g, in a large study doing admixture charts for 6 or 7 values of K on the same Genesis chart), you may require extra memory when running Genesis.

```
java -Xmx2048m -Xms521m -jar Genesis.jar
```

2.3 Known bugs

- On MacOS X, when the program is first run the SWT menu does not appear on the menu bar. If you make another application active and then come back to Genesis, the menu appears. This appears to be a bug of SWT on MacOS X.
- Using some versions of SWT, occasionally when opening a previous project the program crashes with an obscure error message. Update your SWT to a later version. We try to package the latest available version with Genesis. (This appears to have gone away with the latest version of SWT).

2.4 Supported input data formats

For structure charts, `Genesis` supports the output of the `admixture` and `CLUMPP` programs. In addition, we provide a python script that will convert the output of the `structure` program into the appropriate input format. See See Chapter 3 [Creating Structure Plots], page 6.

For PCA charts, `Genesis` supports the output of the `eigenstrat` and `plink` programs and the `SNPRelate` R package. In addition, we provide Python scripts that will convert the output of the `fastpca` program into the appropriate format. See See Chapter 4 [Creating PCA Plots], page 13.

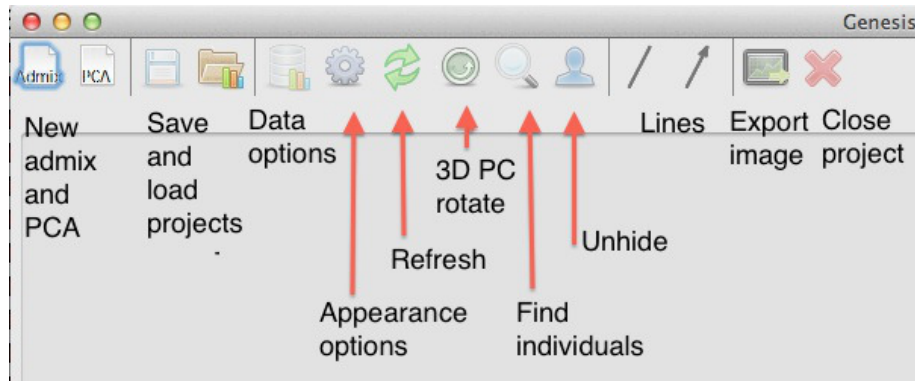
We shall try to add to the suite of conversion scripts, but we intend that future versions of `Genesis` will automatically detect the correct input format and deal appropriately.

2.5 Menu structure

When `Genesis` runs, a menu appears at the top. The menu items are *File*, *Graph* and *Help*. *File* allows creation and saving of graphs, and exporting image files. *Graph* allows changing data and appearance options. The manual is available under *help*.

For added usability the icons with the key menu features appear at the top of the *Genesis* window. If the cursor is placed over the icon, a brief description of the menu item is given.

The menu items are briefly described below and the functions are described in detail in next sections.



From left to right, the menu items are:

- Admix: create a new Admixture chart
- PCA: create a new PCA chart
- Save Project: this saves the current charts being created in a Genesis format file, which incorporates all changes the user has made. This does not save images, but the internal state of the program for later reloading.
- Load Project: this loads a previously saved Genesis project. Note that the New Admix and PCA options are used to create charts from raw output of the admixture and PCA tools. The Load Project option loads previously saved charts.
- Data Options: allows the user to provide constraints on the data files used. For example, the user can specify which column in the file contains the phenotype data, or which principal components should be displayed.
- Appearance Options: gives the user flexibility in determining how what the charts look like. Examples are: size of margin, choice of font and size, headings and so on.
- Refresh.
- Show/Hide 3D Rotate Panel: if the user is showing a 3D PC chart, simple rotation is supported.
- Search for individual: the user can specify the ID of an individual in the study and that individual's data is highlighted in the chart.
- Unhide: individuals or groups can be hidden from the display. This option allows unhiding.
- Line: Simple lines can be drawn
- Arrow: Simple arrows can be drawn
- Export: The picture can be exported to a PNG, SVG, or PDF file.
- Close Project: The current chart is closed. If not previously saved, it is deleted.

3 Structure Plots

3.1 Data input format

Genesis requires two input files and an optional third file:

- An *admixture* file which contains on each line the estimated ancestral proportions of each individual. Typically, this would be produced by a program like Admixture (e.g, and admixture Q file), or CLUMPP, the output formats of which **Genesis** supports natively. For example, an Admixture Q file for $K=4$, contains four columns. Provided the input is a legal format, **Genesis** will automatically work out what input file it is, and what the K value is. Instructions on using the `structure2CLUMPP` script which can be used for Structure input files is described later.
- A plink-style *fam* file. The first two columns must uniquely identify the individuals (the first column is usually the family identifier, the second the individual identifier). See the plink manual for further description. The j -th line of this fam file describes the j -th individual in the admixture file. This all that **Genesis** requires, although *fam* files typically have other columns as well.
- Optionally, a phenotype file: again, the first two columns indentify the individual, and subsequent columns label the individuals (e.g., case/control, male/female, population group, language, etc). The individual identifiers should be the same as in the fam file: each individual in the fam file should also be in the phenotype file, but there is no problem in having individuals in the phenotype file who are not in the fam file. The order of entries in the phenotype file need not be the same as in the fam file. You could use the same file as the fam and the phenotype file.

3.2 Inputting Data

Genesis produces structure charts by taking input from Admixture or CLUMPP tools, together with a PLINK-style fam and optionally a phenotype file which would contain population labels.

To input these files, click **File** → **New Admixture** or the **New Admixture** button on the toolbar. On the screen that opens, click **Import Data File** and navigate to the admixture data file outputted by the Admixture/CLUMPP tools. Then click **Import Fam File** and navigate to the fam data file. Finally, click **Import Pheno File** and navigate to the phenotype data file.

You can import multiple data files into the same project by clicking **Import Data File** again. These files can be imported in any order relative to the importing of the fam and pheno file.

In the drop-down menu select the column of the phenotype file that will be used to group the data. To draw the graph, click **Finish** or click **Next** to access the Appearance Options menu (See below...).

Editing Phenotype Column/Importing Additional Data Files

If a graph has already been plotted, and you wish to change the column of the phenotype data used to group the data or import additional data files, you can access the initial menu by clicking the **Data Options** button on the toolbar, or clicking **Graph** → **Data Options**.

3.3 Appearance Options

The Appearance Options menu can be accessed through the **New Admixture** button by clicking **Next** after importing the files or by clicking the **Appearance Options** button on the toolbar, or clicking **Graph** → **Appearance Options**.

- **Heading**
To set/change the heading, open the **Appearance Options** menu and type the heading into the text box that says “Set Heading.” To change the font of the heading, click **Select Heading Font** and select the font of choice.
- **Border**
To show/hide the border open the **Appearance Options** menu and (un)check the **Show Border** checkbox.
- **Population Group Labels**
To show/hide the population group labels, open the **Appearance Options** menu and (un)check the **Show Population Group Labels** checkbox. To change the font of the population group labels, click **Select Population Group Label Font** and select the font of choice.
- **Graph Size and Spacing**
To change the size and spacing of a structure plot, open the **Appearance Options** menu and select values for the **Set Graph Height**, **Set Thickness of Each Subject** and **Set Distance Between Graphs** boxes.
- **Horizontal/Vertical Orientation**
To set the orientation of the structure plot, and select **Horizontal** or **Vertical** from the bottom drop down menu.

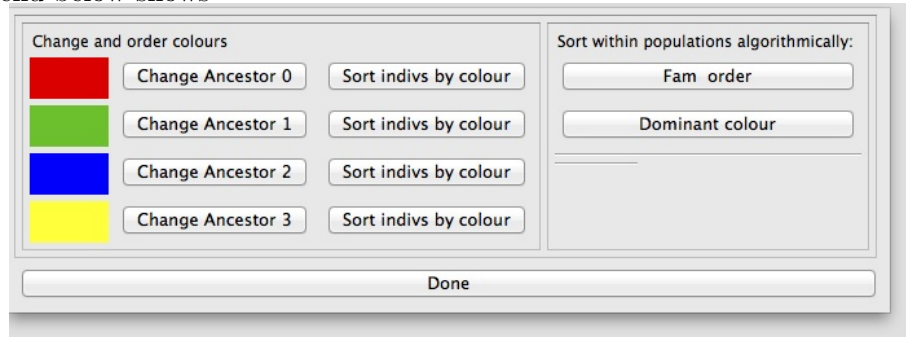
3.4 Interacting with the Graph

Other settings and options can be changed by interacting with the graph. Certain elements can be clicked to view or modify their options.

- **Selecting Subjects**
To select an individual subject on the structure plot, click on the subject on the plot. This will bring up a subject menu where you can view the data about the subject (from the phenotype data file) and gives the option to hide the subject from the plot.
- **Population Group Name and Order**
To select a population group, click on the population group label for the group. This will bring up a menu where you can change the Name of the group and also shift the group left/right (or up/down) in the graph.
- **Changing the colours and order of colours and individuals**
This option allows you to
 - change colours (e.g., change red to blue)
 - change the order of the colours (e.g., on a horizontal chart, move the red segments from below the blue segments to above the red segments)
 - order the individuals within a population group by an ancestral group.

Right click anywhere on a graph and select **Population Group Options**.

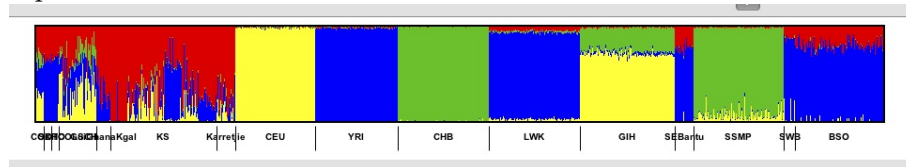
The menu below shows



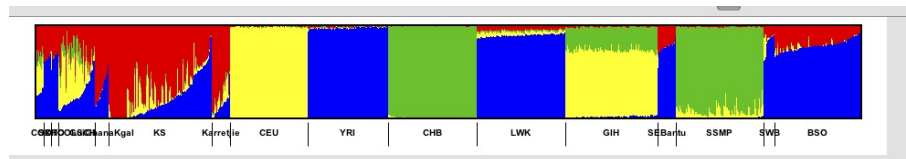
From the next menu, click the button next to the colour you wish to change and from the next dialog, click the **Change \dots** button to change the colour or the **Shift Up** and **Shift Down** buttons (or the **Shift Left** and **Shift Right** buttons) to change the order of the colours (i.e., on a horizontal graph, moving colours up and down; on a vertical graph moving them left and right).

Clicking on a “Sort indivs by colour” button, will sort the *individuals* in each population group in order (left to right on a horizontal graph; top to bottom on vertical graph) so that the individual with the least ancestry represented by that colour is first, and the one with the greatest is last.

For example if we have



and click on the *Sort indivs* next to the blue colour then we get.



Clicking on “Dominant colour” orders each population group in a similar way, but instead of sorting each population group on the same colour, Genesis will compute the dominant ancestral component for that group and then order the individuals in that population group by that component.

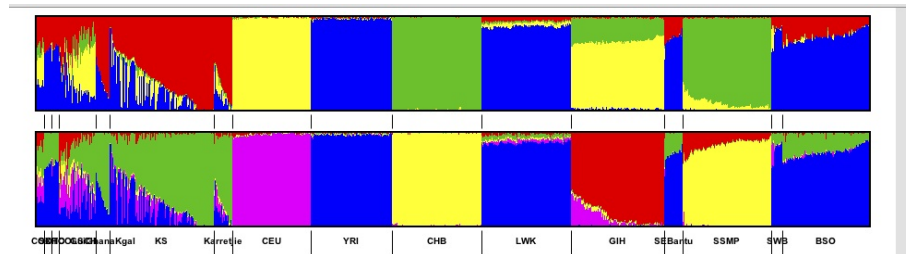
The *Fam order* option is essentially an undo: the individuals are displayed in the relative order that they appear in the fam file.

The ordering of individuals by component is not strict. If two individuals have a very similar but not equal primary component, then a secondary component is used (this is more likely to lead to put similar individuals close to each other). Note that these options are being refined and behaviour may change.

- Recolouring a graph based on another graph

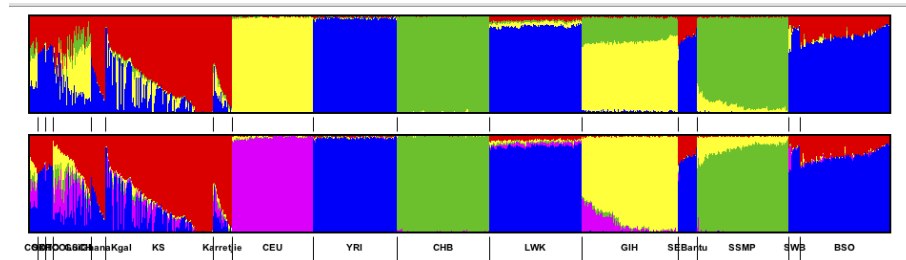
A common use case is to have structure charts for multiple values of K . When multiple graphs are imported some of the most glaring apparent differences will not be due to different numbers of colours but by arbitrary choices of which colours were used. Even though different values of K are used, typically a column from one value of K will correspond to a column from another value of K . However, since the order of the columns is arbitrary the colours are arbitrary. Genesis has an option which tries to match the colours of of one graph with respect to another graph.

For example a common case might be something like:



To recolour a graph, right click anywhere on a graph and select **Population Group Options** and choose either *Colour this graph similar to next* or *Colour this graph similar to previous* (these options only appear if it makes sense). Genesis will recolour.

In our example we will get



Note that this option does not guarantee to recolour optimally. Rather it is intended to reduce the load on the user by making sensible choices for obvious colouring. (Indeed it is not clear what “optimal” means, or best to do colouring. For example, if a “yellow” colour splits into “yellow” and “magenta” which new component should be yellow and which magenta may require insight into populatio dynamics). A very simple heuristic is currently used, which works well in practice. However, this will improve in time.

- Searching for a Subject

To search for a subject in the graph by name, click the **Search for individual** button in the toolbar. In the dialog, enter the Name (first, last or both) of the individual you wish to find and click **Ok**. If the individual was found in the data, it will be selected and the subject dialog for that individual will open. If the individual was not found, a message will displayed.

3.5 Annotating the Graph

- Creating a Label

To create a label and annotate the graph, right click on the graph where the label is to be placed, and click **Create Label at Mouse Pointer**. Then enter the label's text and click **OK** to place the label.

- Editing or Deleting a Label

To reposition a label, click on it and drag to where it should go. If you press the shift key while dragging the label there is a snap to grid feature. To edit a label that has been created, click on the label to bring up the Label dialog. From this dialog you can edit the label's text, reposition the label or delete the label.

- Hiding individuals

To hide a subject from the graph, click on the subject on the graph to bring up the subject dialog. From this menu you can choose to hide the subject from the graph.

- Hiding Groups

To hide a population group from the graph, click on the group in the key to bring up the population group dialog and check the box labelled **Hide this group from the graph**.

- Showing Hidden Subjects and groups

To reshown a subject that has been hidden, click on the **Select Hidden Individuals or Groups** button and from the drop down menu that pops up, select the name of the subject or group that you wish to show. From the subject dialog you can now uncheck the **Hide this Individual from the Graph** or **Hide this group from the Graph** checkbox.

- Drawing lines and arrows

Genesis provides very simple functionality for adding lines and arrows. Click on the icon and place the line or arrow by (a) clicking on the point where the line/arrow should start, (b) *dragging* to the point where the line should finish, and (c) releasing the mouse. If you press the shift key down at the same time, then only vertical or horizontal lines can be produced.

Lines and arrows can be deleted or moved – if you press the shift key down at the same time as moving there is a “snap to grid” feature. Note that lines cannot be shortened or lengthened, and nor can the orientation change.

3.6 Useful scripts

3.6.1 structure2CLUMPP: Wrapper script for the Structure tool

Genesis supports Admixture Q files and CLUMPP output files natively. CLUMPP's output format is a derivative of the Structure tool's output. (Thus our naming convention is somewhat inaccurate, since CLUMPP format is really a sub-part of the Structure format). When Structure runs, it produces log information, summary information about the population and the inferred ancestry of individuals. This inferred ancestry is what we want and can be found in the middle of the output file.

The script **structure2CLUMPP** takes one mandatory argument, the file name of a structure output file.

```
python structure2CLUMPP testdata1.out_f
```

By default, output is placed on standard output. The `--outbase` flag can be used to set the base of the output name. The script appends to this base the suffix $K.Q$, where K is the number of columns in the ancestry file (so as to be consistent with *admixture*).

```
python structure2CLUMPP --outbase data testdata1.out_f
```

If there are 4 columns in the file `testdata1.out_f`, then this will produce a file `data.4.Q`.

The script also has a flag `--produce-fam` which can be used to produce a bare bones fam file if needed.

The full usage is

```
usage: structure2CLUMPP [-h] [--outbase OUTBASE] [--produce-fam FAM_NAME] N
```

produce admixture style output from structure output

positional arguments:

```
N          structure file
```

optional arguments:

```
-h, --help          show this help message and exit
--outbase OUTBASE  output file name base (default output to standard out)
--produce-fam FAM_NAME
                   produce fam file
```

Technical details. The output of the structure program contains various information, including log and FST data. Immediately after the lines that start

Inferred ancestry of individuals:

```
Label (%Miss) Pop: Inferred clusters
```

follow the inferred ancestry of each individual followed by blank line. This is what we want. An extract might look something like this:

```
7      7      (0)    1 :  0.017 0.014 0.970
8      8      (0)    1 :  0.009 0.005 0.986
9      9      (0)    1 :  0.353 0.116 0.531
```

Here $K=3$, and the ancestry of each individual can be found to the right of the colon (for our purposes we can ignore what's to the left of the colon). The `structure2CLUMPP` script extracts out this part of the output from the `structure` program and produces output which could be output from CLUMPP. The `Genesis` program can read this input.

3.6.2 `sortfamQwithin.py` — ordering individuals by colour

Although `Genesis` directly supports some ordering by individuals, we have a script that has more sophisticated functionality. In time this will be migrated into `Genesis`. Usually individuals are ordered in the fam file by group, but within the group the ordering is arbitrary. In admixed populations this may lead to confusing pictures because individuals who are adjacent to each other may have different admixtures so one gets a very jagged chart where a pattern is hard to discern. The script `sortfamQwithin.py` consistently sorts a fam and Q file(s) so that within each group, the individuals are sorted by the dominant ancestral population for that group. This script is documented here:

<http://www.bioinf.wits.ac.za/software/poputils/>

3.6.3 Other scripts

There are other useful scripts like *fams2phe* and *popifyfam.py* which can be used to create phenotype files in <http://www.bioinf.wits.ac.za/software/poputils/>

4 PCA Plots

4.1 Input data format

Genesis takes as input one mandatory file, and one optional file:

- The compulsory input file is a PCA file. In this version of the code, threeinput formats are supported: eigenstrat, plink and SNPRelate.
- A phenotype file (see the description in the previous section), which labels each individual. In addition to plink-style phenotype files, eigenstrat pca.evec files are legal as input (the phenotype is often in the last column).

We have scripts that convert from other popular PCA formats (PLINK, flashpca) to a format the Genesis understands. These scripts are discussed in the section *Advice on data formats* below. We hope that in future versions of Genesis that this will be handled natively.

4.2 Data input

To input Eigenstrat files, click **File** → **New PCA** or the **New PCA** button on the toolbar. On the screen that opens, click **Import Data File** and navigate to the PCA data file outputted by the Eigenstrat software. Then optionally click **Import Phenotype File** and navigate to the phenotype data file. To input SNPRelate Data, click **File** → **New PCA** or the **New PCA** button on the toolbar. On the screen that opens, click **Import Data File** and navigate to the PCA data file outputted by the Relate package. The relate package file includes the phenotype information in the data file.

In the drop-down menus, select the 2 or 3 PCAs to plot as the axes and select the column of the phenotype file that will be used to group the data. To draw the graph, click **Finish** or click **Next** to access the Appearance Options menu (See below...).

- Editing PCAs/Phenotype Column

If a graph has already been plotted, and you wish to change the PCAs to be the axes or the column of the phenotype data to represent the data, you can access the initial menu by clicking the **Data Options** button on the toolbar, or clicking **Graph** → **Data Options**.

- Appearance Options

The Appearance Options menu can be accessed through the **New PCA** button by clicking **Next** after importing the files or by clicking the **Appearance Options** button on the toolbar, or clicking **Graph** → **Appearance Options**.

- Heading

To set/change the heading, open the **Appearance Options** menu and type the heading into the text box that says “Set Heading.” To change the font of the heading, click **Select Heading Font** and select the font of choice.

- Border

To show/hide the border open the **Appearance Options** menu and (un)check the **Show Border** checkbox.

- **Axes**
To show/hide the axes, axis labels, grid and scale, open the **Appearance Options** menu and (un)check the relevant checkboxes.
- **Key Options**
To set the position of the key, open the **Appearance Options** menu and select the key position from the drop down menu. To hide the key, select **No Key** from the drop down menu. To change the font of the key, click **Select Key Font** and select the font of choice.
Other settings and options can be changed by interacting with the graph. Certain elements can be clicked to view or modify their options.
- **Selecting Subjects**
To select an individual subject on the structure plot, click on the subject on the plot. This will bring up a subject menu where you can view the data about the subject (from the phenotype data file) and gives options to change the subject's icon as well as to delete the subject and hide or place the subject on top.
- **Changing Subject's Icon**
To change an individual subject's icon, click the subject on the graph and from the dialog that opens, select the icon shape and colour. Selecting the shape **Default** will set the shape of the icon to the group's shape and checking **Clear all icon data specific to this individual** will set the icon's shape and colour to the group's.
- **Placing a Subject on Top**
To place an individual subject on top, click the subject on the graph and from the dialog that opens, check the **Place this individual on top** checkbox and click **Done**.
- **Changing the Population Group Names**
To change a population group's name, click on the group in the key. This will bring up the Population Group dialog. From here you can set the group name and click **Done**.
- **Changing the Population Group Icons**
To change a population group's icon, click on the group in the key. This will bring up the Population Group dialog. From here you can set the shape and colour of the icon and click **Done**.
- **Changing the Order of Population Groups in the Key**
To change a the order of the population groups in the key, click on the group in the key. This will bring up the Population Group dialog. From here you can click **Shift Up** or **Shift Down** shift the group's order in the key.

4.3 Annotating the Graph

- **Creating a Label**
To create a label and annotate the graph, right click on the graph where the label is to be placed, and click **Create Label at Mouse Pointer**. Then enter the label's text and click **OK** to place the label.
- **Editing or Deleting a Label**
To edit a label that has been created, click on the label to bring up the Label dialog. From this dialog you can edit the label's text, reposition the label or delete the label.

Labels can be moved: pressing the shift key while dragging creates a snap to grid feature.

- Drawing lines and arrows

Genesis provides very simple functionality for adding lines and arrows. Click on the icon and place the line or arrow by (a) clicking on the point where the line/arrow should start, (b) *dragging* to the point where the line should finish, and (c) releasing the mouse. Currently no line editing functionality is provided, other than being able to delete a line.

4.4 Hiding Subjects and Population Groups from the Graph

- Hiding Subjects

To hide a subject from the graph, click on the subject on the graph to bring up the subject dialog. From this menu you can choose to hide the subject from the graph.

- Hiding Groups

To hide a population group from the graph, click on the group in the key to bring up the population group dialog and check the box labelled **Hide this group from the graph**.

- Showing Hidden Subjects and Groups

To reshown a subject that has been hidden, click on the **Select Hidden Individuals or Groups** button and from the drop down menu that pops up, select the name of the subject or group that you wish to show. From the subject dialog you can now uncheck the **Hide this Individual from the Graph** or **Hide this group from the Graph** checkbox.

4.5 Searching for a Subject

To search for a subject in the graph by name, click the **Search for individual** button in the toolbar. In the dialog, enter the Name (first, last or both) of the individual you wish to find and click **Ok**. If the individual was found in the data, it will be selected and the subject dialog for that individual will open. If the individual was not found, a message will display.

4.6 Rotating the Graph

To rotate a 3D PCA plot, click the **Show/Hide 3D PCA Rotate Panel** button in the toolbar. This will bring up the rotate panel which contains a slider which can be dragged to rotate the graph about the z-axis.

4.7 Advice on data formats

Eigenstrat is directly supported by the **Genesis**.

4.7.1 SNPRelate

The **SNPRelate** R package of Zheng *et al* [2012] can be used to do PC-analysis. However, since it is an R-package there is no **SNPRelate** default format since output is fully programmable in R. We support the following output: a file that contains the eigenvalues, followed by the eigenvectors, produced using the following R commands.


```
pca <- snpgdsPCA(genofile,snp.id=snpset)
write.table(pca$eigenval,"pca.rel",sep="\t",quote=FALSE)

tab1 <- data.frame(sample.id = pca$sample.id,
  pop = factor(pop_code)[match(pca$sample.id, sample.id)],
  EV1 = pca$eigenvect[,1],
  EV2 = pca$eigenvect[,2],
  EV3 = pca$eigenvect[,3],
  EV4 = pca$eigenvect[,4],
  EV5 = pca$eigenvect[,5],
  EV6 = pca$eigenvect[,6],
  EV7 = pca$eigenvect[,7],
  EV8 = pca$eigenvect[,8],
  EV9 = pca$eigenvect[,9],
  EV10 = pca$eigenvect[,10],
  stringsAsFactors = FALSE)

write.table(tab1,"pca.rel",sep="\t",quote=FALSE,append=TRUE)
```

4.7.2 flashpca

FlashPCA is designed to perform PCA on very large data set. It takes as input a plink BED and BIM file and produces eigenvectors or principal components. We have a script *flashpca2evec* which converts the data into a format that **Genesis** can read. Because the flashpca output has no information about the sample IDs, *flashpca2evec* also needs the fam file as input. This script **requires Python 2.7**.

By default, flashpca calls its output files *eigenvalues.txt* and *eigenvectors.txt* and this is (by default) what *flash2pca* expects. For example:

```
flashpca2evec --fam data.fam --out data.evec
```

However, if the files have other names, the appropriate flags can be used

```
flashpca2evec --fam data.fam --eigenval file1.evals --eigenvec sample.csv --out data.evec
```

4.7.3 plink2evec

PLINK 2 [Purcell and Chang 2014] (and its alpha release plink 1.9) supports PCA directly. Genesis can handle these files natively but assumes that the default plink's default naming convention is used (e.g., a *.eigenvec* suffix). If this is not followed, Genesis will not be able to recognise the file type. Thus plink2evec is bundled for that purpose.

plink2evec converts the plink output files into the format that Genesis can read.

By default PLINK calls its output files *plink.eigenval* and *plink.eigenvec* and this is (by default) what *plink2evec* expects. For example:

```
plink2evec --out result.pca.evec
```

However, if the files have other names, the appropriate flags can be used

```
plink2evec --eigenval file1.evals --eigenvec sample.csv --out data.evec
```

And if as is common in plink usage, the eigenvectors and eigenvalues file was specified by using the *plink* --out flag, then *plink2evec* can use its --bfile flag

```
plink --bfile sample --pca --out sample  
plink2vec --bfile sample --out sample.pca.evec
```

5 Saving, Loading and Exporting Projects

5.1 Saving a Project

To save a project currently in use, click the **Save Project** button on the toolbar or click **File** → **Save Project**. Enter a filename and navigate to the location you wish to save the file and click **Save**.

We recommend regular saving. Although, Genesis is stable, like other software it may crash and saving is easy.

Please note that in the Genesis GUI, that you may have several projects open at once. Saving only saves the current project!

5.2 Loading a Project

To load a project previously saved, click the **Load Project** button on the toolbar or click **File** → **Load Project**. Navigate to and select the file you wish to load and click **Load**.

5.3 Exporting Graphs

Once you are satisfied with the graph you are currently working on, you can export the graph as an image file (.png) or in a PDF. To do so, click the **Export** button on the toolbar or click **File** → **Export**.

From the dialog that opens, select whether to export as a PNG, SVG image or as a PDF, then navigate to the location you wish to save the file and click **OK**.

5.3.1 Notes on exporting as a PDF

Genesis will export the graph as it is on the screen in front of you, and will scale this image to fit on the width of an A4 page. This may cause larger admixture project's images to be of too high detail to view all the detail on the monitor of your computer all at once. This is normal and expected. The detail will be stored there - it may be necessary to zoom in to view it all - and if the image is printed, the detail will be there (assuming the printer is of high enough quality).

5.3.2 Notes on exporting as a PNG image

PNG (Portable Network Graphics) is a cross-platform image file format that can be handled by almost all good image viewers and editors. When exporting as a PNG file, note that the image will be the same size as it is on the screen as you export it. This means that resizing the Genesis window between PCA exports will create different sized images (Admixture sizes are determined by the options in the **Appearance Options**). This is important as PNG images do not scale up, i.e. a smaller image will become blurry if it is stretched too big. Keep this in mind when exporting as a PNG image.

5.4 Error messages

- 1: Only 1 column in input file – not allowed
- 2: The admix or phenotpe file has a varying number of columns

- 3: Java gives a NumberFormat error in the input file.
- 5: File is not found
- 102: I think you have a plink eigenvector file but a line does not parse.
- 300: PCA read is not in SMARTPCA format.
Hint: if you have a PLINK file make sure that your fam file is in the directory.
- 400: Error in data file: file may be empty

6 Conclusion

Any technical questions on the use of **Genesis** should be sent to genesis@bioinf.wits.ac.za. ■

6.1 Literature cited

- G Abraham, M. Inouye (2014). Fast principal component analysis of large-scale genome-wide data. *PLoS One* **9**(4) p. e93766, doi:10.1371/journal.pone.0093766
- Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9), 165564. doi:10.1101/gr.094052.109
- Jakobsson, M., and Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, 23(14), 18011806.
- Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, 2(12), e190. doi:10.1371/journal.pgen.0020190
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 94559.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, 81.
- S. Purcell and C. Chang 2014. PLINK. <https://www.cog-genomics.org/plink2>
- Xiuwen Zheng, David Levine, Jess Shen, Stephanie M. Gogarten, Cathy Laurie, Bruce S. Weir. A High-performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. *Bioinformatics* 2012; doi:10.1093/bioinformatics/bts606

6.2 Copyright

Copyright (C) 2014, Robert W Buchmann, University of the Witwatersrand

This program is free software: you can redistribute it and/or modify it under the terms of the GNU Affero General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU Affero General Public License for more details.

You should have received a copy of the GNU Affero General Public License along with this program. If not, see <http://www.gnu.org/licenses/>