*AncestryPainter*:

a program for displaying ancestry composition of populations and individuals

Qidi Feng1

CAS-MPG Partner Institute for Computational Biology (PICB)  
Shanghai Institutes for Biological Sciences (SIBS)  
Chinese Academy of Sciences (CAS)  
320 Yueyang Road

Shanghai 200031, P.R.China

The AncestryPainter software2 is available at

http://www.picb.ac.cn/PGG/resource.php

Dec 26, 2017

1Comments can be emailed to me at [fengqidi@picb.ac.cn](mailto:fengqidi@picb.ac.cn).

2This manual is the companion to version 1.1 of the software.

Table of Contents

1 Introduction 3

1.1 Availability and Citation 3

1.2 Basic Overview 3

2 Inputs 5

2.1 Individual Q-matrix file (Q file, required) 5

2.2 Individual information file (ind file, required) 5

2.3 Target Population/Individual (population/individual name, optional) 6

2.4 Population order file (pop.order file, Optional) 8

2.5 Color order file (color.file, Optional) 9

3 Outputs 11

3.1 PDF/PNG Figure 11

3.2 R script 11

3.3 Ancestry file 11

3.4 Color file 11

4 Command line arguments 12

5 Example 13

6 Reference 14

# 1 Introduction

Estimation of ancestry for individuals is a key question in population genetics, which was tried to be solved by model-based approaches like structure[1](#_ENREF_1), ADMIXTURE[2](#_ENREF_2). These clustering software concerned with estimating the proportion of K ancestries (K clusters) for each individual, where K is a parameter given in advance, but can be varied in independent runs. Individuals are given proportion for each ancestry, such that the estimated proportions of an individual sum to one across the K ancestries. The matrix of ancestry proportion is referred to here as the individual Q-matrix (the Q out file of ADMIXTURE software).

A convenient way to display the ancestry of individuals is to show each individual as a line segment, which is partitioned into K colored components, with the length of each colored component represents the proportion of its corresponding ancestry. Such graphic display can be achieved by existing programs like *Distruct[3](#_ENREF_3" \o "Rosenberg, 2003 #569)*. However, all existing programs as far as we know display populations in a horizontal direction. If there were too many individuals or populations, the graphic would be extremely long and difficult to display in a single line. Besides, display all individuals in a single line is not convenient for comparison if they are relatively distant in display.

To solve these drawbacks, our program uses a circular way to display the ancestry of individuals, with advances in that: 1. The circular display avoid the awkward situation of too many populations in order, thus takes a better advantage of space; 2. For a better comparison, the width of each population was set to be equal regardless of its sample size; 3. A pie chart could be optionally added in the center of the circular graph to highlight the ancestry proportion of the target population/individual the user particularly concerned; 4. For a better view purpose, the order of population as well as individuals within populations in display could be automatically sorted according to their ancestry proportion.

## 1.1 Availability and Citation

The AncestryPainter was first used to generate Figure2A in Feng et al. (2017)[4](#_ENREF_4). The website for AncestryPainter is: http://www.picb.ac.cn/PGG/resource.php

*ADMIXTURE* or *structure* software could be applied to get the ancestry estimation of individuals.

The *ADMIXTURE* software is available at <https://www.genetics.ucla.edu/software/admixture/>.

The *structure* software is available at <http://web.stanford.edu/group/pritchardlab/structure.html>.

## 1.2 Basic Overview

The AncestryPainter program can be used under Linux, MacOS and Windows. The program is written in Perl. It automatically produces and calls an R script to draw the figure. The download website for Perl and R is shown below. The program takes individual-Q matrix file and population information file as inputs, with/without the indication of target population, it outputs the pdf/png figure directly. Example data are included in the program package. User could test the program with example data by a simple command:

perl AncestryPainter.pl -i ./example/data.ind -q ./example/data.Q -t Uygur

The program would produce the figure out.pdf, the R script used to draw the figure out.r, and out.ancestry that integrate the information of population, individuals and ancestry proportion.

The program allows additional options that permit the user to control the target population or individual the pie chart display, the color of each ancestry, the order of populations in display, the format of output figure. (please refer to 4. Command-line arguments).

Website to download Perl:

<https://www.perl.org/get.html>

Website to download R:

<https://www.r-project.org>

**Notice**: In order to run *AncestryPainter*, please add the path of Perl and R to your environment variable so that the produced R script could be automatically called.

How to add Perl/R to your environment variable under Windows system?

1. Type ‘sysdm.cpl’ in cmd, then an environment variable graphical interface would appear.
2. Add Perl/R\_installed\_path to the beginning of PATH variable;
3. Reboot cmd, and then we can use R/Perl anywhere.

If failed to add the path of Perl and R, then you need run ‘*perl AncestryPainter.pl -i ./example/data.ind -q ./example/data.Q –o output’* under where you installed Perl, and then copied the outputs to the path where R installed, and run ‘*Rscript output.r output.ancestry output.pdf’.* Figure would be produced.

# 2 Inputs

The program takes a file with individual Q-matrix (Q file, required), and a separate file of individual information file (ind file, required) to get the corresponding population information of individual Q-matrix. Q file is the output of *ADMIXTURE* software. The order of each individual in Q file should be the same with that in individual information file. AncestryPainter program would integrate these two files to produce out.ancestry file.

## 2.1 Individual Q-matrix file (Q file, required)

Individual Q-matrix file is the ancestry proportion (by column) of each individual (by row) output by clustering programs like *ADMXITURE* or *structure* software. The format is exactly the same with Q output of *ADMIXTURE* program. Each row represents an individual. Each column represents the proportion of each ancestry. Columns are separated by tab or spaces. Table1 shows the proportion of 4 ancestries of 15 individuals. **For each individual, the sum of ancestry proportions should be 1**.

Table1: Individual Q-matrix (Q file)

|  |  |  |  |
| --- | --- | --- | --- |
| 0.287624 | 0.184745 | 0.365431 | 0.1622 |
| 0.272924 | 0.145596 | 0.460998 | 0.120482 |
| 0.322236 | 0.198415 | 0.347277 | 0.132072 |
| 0.314911 | 0.20066 | 0.340199 | 0.14423 |
| 0.232732 | 0.180639 | 0.425543 | 0.161085 |
| 0.292692 | 0.214244 | 0.341407 | 0.151656 |
| 0.26415 | 0.188119 | 0.404649 | 0.143082 |
| 0.289151 | 0.213711 | 0.367376 | 0.129762 |
| 0.395426 | 0.221496 | 0.26315 | 0.119928 |
| 0.260558 | 0.146546 | 0.416873 | 0.176023 |
| 0.30849 | 0.156664 | 0.387158 | 0.147688 |
| 0.883945 | 0.100641 | 0.00429 | 0.011124 |
| 0.894186 | 0.090524 | 0.015279 | 0.00001 |
| 0.912589 | 0.080458 | 0.006609 | 0.000344 |
| 0.880817 | 0.109171 | 0.010002 | 0.00001 |

## 2.2 Individual information file (ind file, required)

The ind file needs the first two columns to be population and individual information, separately. Columns are separated by tab or spaces. The order of individuals in ind file should be exactly the same with that in Q file. Table2 indicate the 15 individuals from two populations, which corresponds to the ten individuals in Table1.

Users could also use ped file or fam file in PLINK[5](#_ENREF_5) (which are also inputs of ADMIXTURE program) instead of the ind file, since the first two columns of ped/fam file could also provide the population and individual information our program required.

Table2: Individual information file (ind file)

|  |  |
| --- | --- |
| Uygur | 1A4-5 |
| Uygur | 1A8-9 |
| Uygur | 1B1-14 |
| Uygur | 1D11-50 |
| Uygur | 1F1-71 |
| Uygur | 1F12-86 |
| Uygur | 1F3-73 |
| Uygur | 1F8-78 |
| Uygur | 1G1-87 |
| Uygur | 1G4-91 |
| Uygur | 1G5-92 |
| Albanian | ALB191 |
| Albanian | ALB202 |
| Albanian | ALB212 |
| Albanian | ALB213 |

## 2.3 Target Population/Individual (population/individual name, optional)

User could define one and only one target population/individual to be plotted as a pie chart in the center of the circle figure. The target population/individual could be the candidate the user particularly concerned. Name of target population/individual must be included in the ind file. Figure1 shows the figure drawn without any target population/individual indicated, while Figure2 shows that drawn with ‘Uygur’ as the target population by ‘-t Uygur’. Figure3 show that drawn with ‘’ as the target individual by ‘-t ‘.

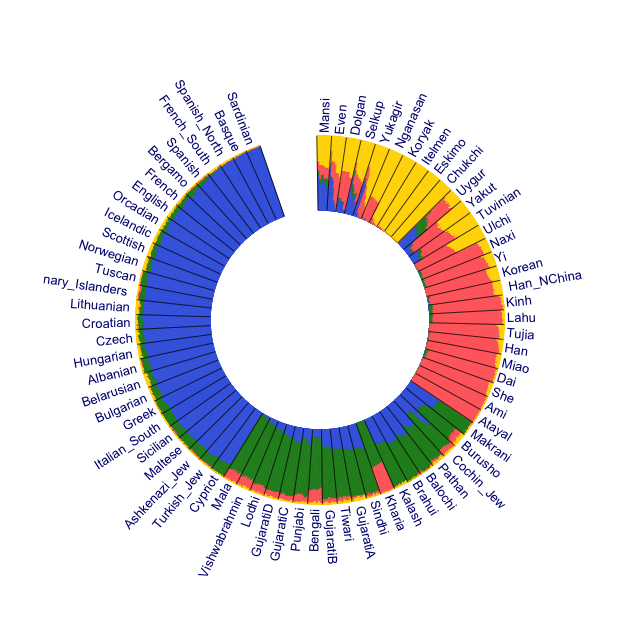


Figure1 AncestryPainter without target population indicated

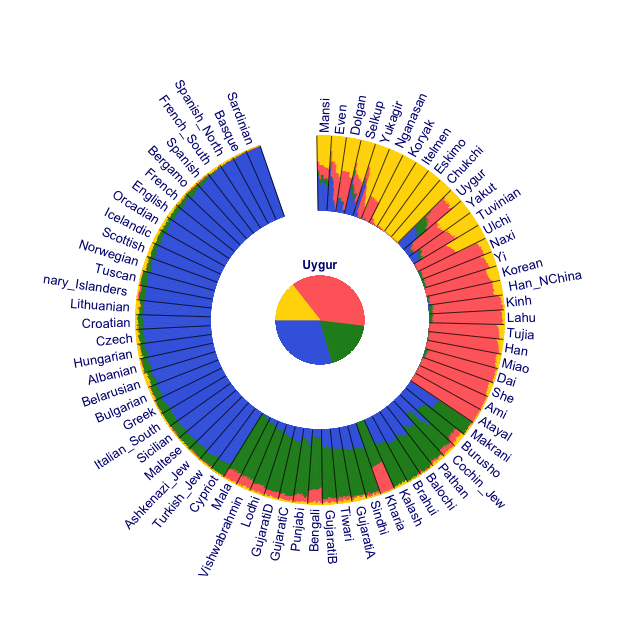


Figure2 AncestryPainter with ‘Uygur as target population

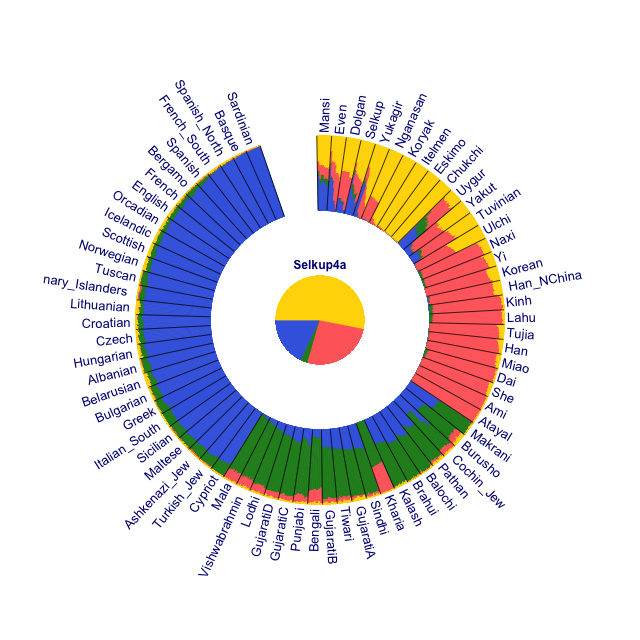


Figure3 AncestryPainter with ‘Selkup4a’ (from population Selkup) as target individual

## 2.4 Population order file (pop.order file, Optional)

By default, *AncestryPainter* program would automatically display the input population in an order with their representative proportion (the largest ancestry for each population) from highest to lowest for each K (Figure2). To get the population order file in the automatically produced figure, user could use Linux command ‘cut -f 2 out.ancestry | uniq > pop.order’, where ‘out.ancestry’ is one of the output files.

User could further modify, or exclude some populations in the pop.order, and re-run *AncestryPainter* program by adding ‘-p pop.order’ to adjust the final figure. ‘/example/pop.order’ shows an example, with one population every single line. Population name included in pop.order file should be included in the population information file.

Figure4 shows the figure produced by example data with the modified population order file indicated by ‘-p pop.order’. Figure4 shows the figure drawn with a reduced population set.

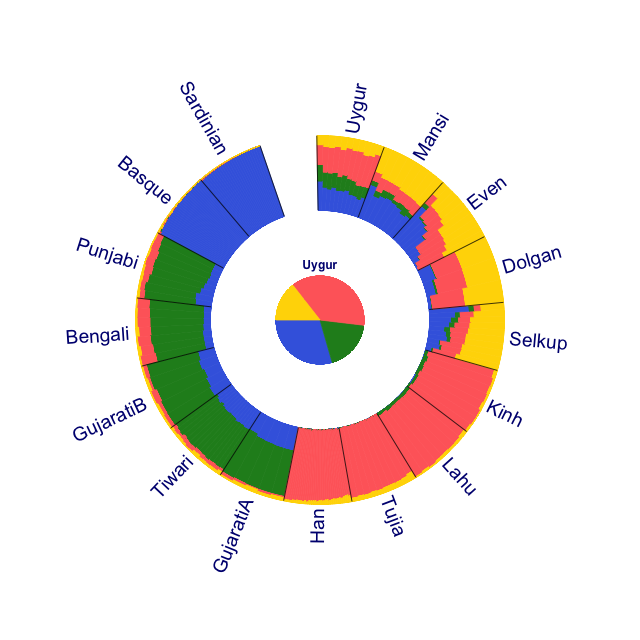


Figure4. AncestryPainter with reduced population list (‘Uygur’ as target population).

## 2.5 Color order file (color.file, Optional)

User could decide the colors used for each ancestry in the figure by using color order file (-c color.file). /example/color.file shows an example of color order file, with one color every single line. Color of Nth line corresponds to the ancestry of Nth column in the Q file.

Users could use default colors in the program without defining color order file. There are 78 default colors in AncestryPainter program. Therefore, if the number of ancestries is more than 78, the user must give their own color order file. One of the output is the color file used, based on which users could modify to modulate the color in the figure.

Figure5 shows the figure drawn with ‘-c color.file’ indicated.

Figure6 shows the figure draw with black lines removed with ‘-l noline’ indicated.

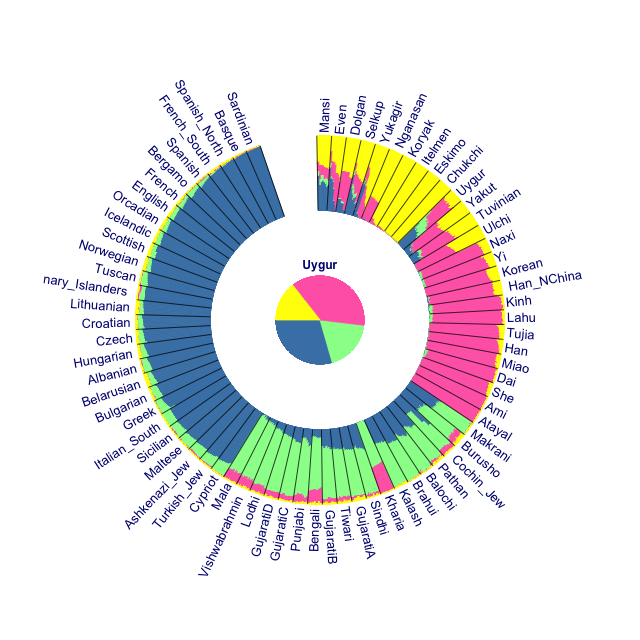


Figure5. AncestryPainter with ‘-c color.file’ option (Uygur as target population)

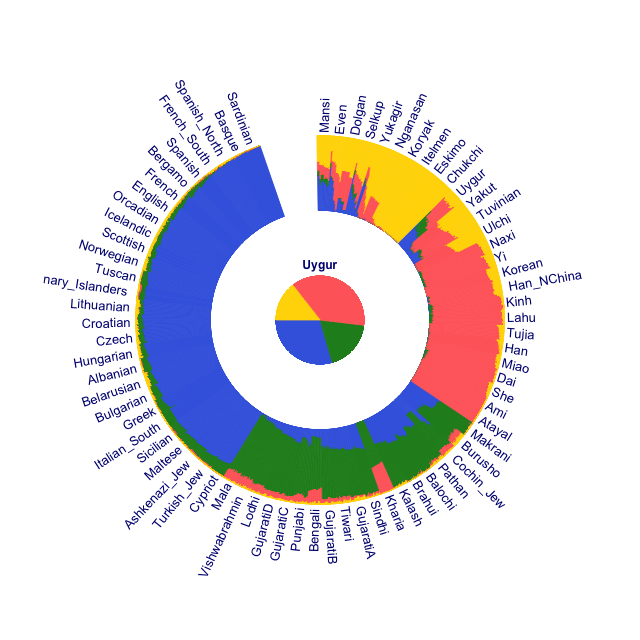


Figure6 AncestryPainter with ‘-l noline’ option (Uygur as target population)

# 3 Outputs

## 3.1 PDF/PNG Figure

Our program would automatically produce the out.pdf (default) or out.png (‘-f png’) figure under the current directory.

## 3.2 R script

Our program would also outputs the R script (out.r) which drawn the figure. User could also modify the R script to improve the figure by running ‘Rscript out.r’.

## 3.3 Ancestry file

Ancestry file is the integrated file of ancestry proportions and individuals information with populations been sorted (data.ancestry). User could directly check the ancestry proportions of particular individual from ancestry file.

## 3.4 Color file

Color file contains the colors used in figure (out.color), based on which users could modify and then used as input to update the figure.

# 4 Command line arguments

Command-line arguments enable the user to indicate certain options from the command line.

‘-q’: (required, input file) Define the individual Q-matrix file.

‘-i’: (required, input file) Define the population information of each individual.

‘-t’: (optional, population name) Define the only target population to be plotted as a pie chart in the center of the circle figure.

‘-p’: (optional, input files) Define the populations to be included in the figure, and the display order of the populations in the figure.

‘-c’: (optional, input files) Define the colors of each ancestry in the figure.

‘-o’: (optional) Define the suffix of output files. ‘out’ by default.

‘-l nolines’: (optional) remove the black lines between populations. Default is keeping black lines (Figure6).

‘-f png’: (optional) output png figure instead of pdf figure. Default is pdf figure.

# 5 Example

The example input files used in this program package is modified from K=4 graph shown for populations of East Asians, West Eurasians, Central Asian/Siberians, and South Asians, with Uygur as target population in Figure 2A of Feng et al. (2017)[4](#_ENREF_4). The example directory include data.ind, data.Q (output of *ADMIXTURE* program), color.file (to define the color of each ancestry), and pop.order (to define the population order in figure display).

When we type the simplest command:

perl AncestryPainter.pl -i ./example/data.ind -q ./example/data.Q

We can get the figure out.pdf (Figure1). the R script produced the figure out.r, the color used to draw the figure (out.color) and the ancestry file out.ancestry.

We could add the target population to be plotted in center, population order, colors, and define the output suffix by added the options, and get Uygur.pdf (Figure5), Uygur.r and Uygur.ancestry:

perl AncestryPainter.pl -i ./example/data.pop -q ./example/data.Q -t Uygur -c color.file -p pop.order -o Uygur

# 6 Reference

1. Jonathan K. Pritchard, M.S.a.P.D. (2000). Inference of Population Structure Using Multilocus Genotype Data. Genetics 155: 945–959.

2. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome research 19, 1655-1664.

3. Rosenberg, N.A. (2003). distruct: a program for the graphical display of population structure. Molecular Ecology Notes 4, 137-138.

4. Feng, Q., Lu, Y., Ni, X., Yuan, K., Yang, Y., Yang, X., Liu, C., Lou, H., Ning, Z., Wang, Y., et al. (2017). Genetic History of Xinjiang's Uyghurs Suggests Bronze Age Multiple-Way Contacts in Eurasia. Molecular biology and evolution 34, 2572-2582.

5. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81, 559-575.