Documentation of the GRAF-pop Feature

of the GRAF software package

Yumi (Jimmy) Jin, PhD

[jinyu@ncbi.nlm.nih.gov](mailto:jinyu@ncbi.nlm.nih.gov)

National Center for Biotechnology Information

National Library of Medicine

National Institutes of Health

4/17/2018

**Summary**

A new feature GRAF-pop was added to the GRAF software package in version 2.0, which included a *-pop* option of the C++ program *graf* and a Perl script *PlotPopulations.pl* to further process the results.

The *-pop* option is used to tell *graf* to attempt to assign subjects to populations and to save the results in the output text file specified by this option. For example, the following command reads genotypes from the PLINK file collection with prefix G1000FpGeno and calculates the genetic distance scores and ancestry proportions for subjects and saves results to the file G1000\_sbj\_scores.txt:

graf -plink G1000FpGeno -pop G1000\_sbj\_scores.txt

The results generated by the *graf -pop* option can be passed to *PlotPopulations.pl* for further processing. *PlotPopulations.pl* takes two required parameters and some optional parameters. The first parameter required is the file name of above mentioned *graf -pop* results. The second required parameter is the name of an output file, which should be either a .png or a .txt file. When the second parameter is a .png file, the script processes the data and plots the scores of each subject on a graph and saves the results in the output file, e.g.,

PlotPopulations.pl G1000\_sbj\_scores.txt G1000\_sbj\_pops.png

When ancestries are available, perhaps by self-report from the subjects, this information can be saved to a two-column text file and passed to *PlotPopulations.pl,* so that the script can plot the ancestries on a graph, e.g.,

PlotPopulations.pl G1000\_sbj\_scores.txt G1000\_sbj\_pops.png

-spf G1000SbjSuperPop.txt

The abbreviation spf stands for “Self-reported Population File”. If the second parameter is a .txt file, *PlotPopulations.pl* saves subject genetic distance scores and the calculated ancestry components to the output file, e.g.,

PlotPopulations.pl G1000\_sbj\_scores.txt G1000\_sbj\_list.txt

-spf G1000SbjSuperPop.txt

In comparison with version 2.0, versions 2.1 and higher use slightly different statistical metrics to determine the genetic distance between each subject and each reference population. In addition, GRAF versions 2.1 and higher normalize the statistical scores on the SNPs with genotypes for each sample, which makes the results comparable among genotype datasets obtained using different methods. The GRAF-pop feature tolerates not only very high genotype missing rates but also non-randomly missing genotypes.

See the following sections for descriptions of the GRAF-pop algorithm, explanations of genetic distance scores and ancestry components, as well as possible usages of other program options.

**GRAF-pop algorithm**

GRAF-pop is a feature that was added to the software package GRAF (Jin et al, 2017) in version 2.0. The documentation for GRAF-pop is currently separate from the main documentation for GRAF, which is in the file GRAF\_ReadMe\_YYYYMMDD.docx, where YYYYMMDD represents a date. Some passages below assume familiarity with this ReadMe file, so users of the GRAF-pop feature should read the general ReadMe file before proceeding to read more of this file.

The GRAF-pop feature can be used to estimate subject ancestry compositions from the genotype data obtained using different methods. Inherent to the GRAF-pop method is that the 1000 Genomes Project collected its initial subjects from three populations: Asian, European, and African. Therefore, for many single nucleotide polymorphisms (SNPs), the frequencies of the two alleles are known with most accuracy in these three populations. The frequencies of SNP alleles can be used in various ways to define distances between populations, including the original three populations and other populations. In general, dbGaP users have expressed interest in being able to estimate the population ancestry compositions of subjects based on genotype data.

The algorithm implemented in GRAF-pop works as follows. To determine the subject ancestries, we first check the dbGaP Fingerprint Collection (See GRAF ReadMe file) and find all subjects whose ancestries are self- reported by the subjects (and relayed to dbGaP by the data submitters) as European (EUR), African or African American (AA), or Asian (ASN, not including South Asian). The allele frequencies of the 10,000 fingerprint SNPs are calculated for the three populations using the fingerprint genotypes of these subjects, after closely related subjects and outliers are filtered out, and these frequencies are used as references to determine ancestries for other subjects.

Suppose a sample is genotyped for *S* independent, biallelic SNPs, and the count of the reference allele of SNP *j* is *gj0*, *1*, *2*}. Also suppose for a comparison population *c*, the frequencies of the reference and alternative alleles are *pj,* and *qj.* Define a statistical score *D* between sample *s* and population *c* as:

Generally, *D* is greater when the genotype of the sample *s* is more different from the genotypes of most of the subjects in the selected comparison population *c*. We refer to *D* as the “genetic distance” of genotypes between *s* and *c*.

To determine the ancestry for each subject *s*, GRAF-pop calculates *D* scores between *s* and each of the above three comparison populations. Each subject is represented by a point in a 3-dimensional Cartesian coordinate system defined by its three *D* scores. Subjects from same populations form clusters in space.

We select dbGaP subjects of the following three sub-populations as reference populations to estimate the ancestry proportions of other subjects: Self-reported CEU, or Utah Residents with Northern and Western European Ancestry to represent European (E), Ghanaian to represent African (F), and Chinese and Japanese to represent East Asian (A). These three populations are different from the previous three comparison populations (EUR, AA, ASN) whose allele frequencies are used for calculating *D* scores because the EUR, AA, and ASN populations rely on self-report and cover much larger geographic regions, whereas the E, F, and A subjects were selected from narrow regions.

When being connected by line segments, the centroids of the three reference populations form a triangle ΔEFA. We translate and rotate all the points in the space so that ΔEFA is parallel to the x-y plane and the FA side of ΔEFA is parallel to the x-axis. Next, the points are projected and plotted on the x-y plane. After transformation, the *x, y, z* coordinates of each subject are called GD1, GD2, GD3 scores, respectively. The nomenclature is by analogy to principal component (PC) scores used by other methods for assigning populations. Fig. 1 shows an example graph obtained using HapMap subjects.

Assuming all subjects are admixtures of the three ancestries *E*, *F*, and *A*, we calculate the proportions of these three ancestries of each subject based on the barycentric coordinates (Coxeter, 1969) of this subject; details of how to compute the barycentric coordinates are below. If a subject is inside the reference triangle ΔEFA, then the three normalized barycentric coordinates are within the interval (0, 1). The physics intuition is that the point is the center of mass resulting from putting three unequal masses at the three vertices; if those masses are normalized to sum to 1, then they are the proportion of the mass at each vertex, or equivalently the proportion of each reference population in the ancestry of this subject. In this case we directly use these coordinates to estimate the three ancestry proportions. If a subject is located outside the triangle, then we set each coordinate to 0 if it is less than 0, and calculate the proportion of each ancestry as the percentage of the corresponding coordinate. In general, we do the following to estimate the proportions *Pe*, *Pf* and *Pa* of ancestry *E, F, A* for each subject:

1. Calculate the barycentric coordinates (*α, β,* *γ*) for this subject with respect to the reference triangle ΔEFA.
2. Let *α’ =* max(0, α), *β’* = max(0, β), *γ’* = max(0, γ).
3. Calculate ancestry proportions as:

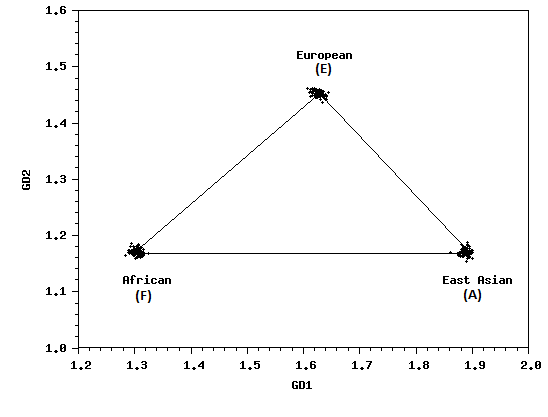


Fig. 1. A graph showing the GRAF-pop *D* scores of the HapMap subjects. The *GD1, GD2* values are the *D* scores projected on the x-y plane after translations and rotations (without scaling or shearing).

The above method assumes that same set of SNPs are used for calculating the *D* scores of the subjects being checked and those in the reference populations *E, F, A*. In GRAF-pop, assuming each subject can have different, non-random missing genotypes, we calculate the expected mean *D* scores of subjects in reference population *E, F, A* for the specific set of genotyped SNPs of each subject being checked, and use these scores to construct reference triangle ΔEFA.

Suppose there is a population *H* in Hardy-Weinberg equilibrium with *M* independent, biallelic SNPs genotyped for all subjects. Let the frequencies of the reference and alternative alleles at SNP *j* be *uj* and *vj*. If we calculate the *D* score from each subject in *M* to a comparison population *c* *EUR*, *AA*, *ASN*}, whose allele frequencies at SNP *j* are *pj* and *qj*, the mean *D* score is expected to be:

.

For each subject, GRAF-pop finds the set of fingerprint SNPs with genotypes, and calculates the three *D* scores to comparison populations EUR, AA, and ASN. Then it calculates the expected *D* score for each population *H* *E*, *F, A*} to each comparison population *R* *EUR*, *AA*, *ASN*}, using the same set of SNPs. The subject being checked is represented with a point *P* in the Cartesian coordination system, and the expected *D* scores of populations *E, F, A* are used to create a reference triangle ΔEFA for the subject. Point *P* and the three vertices of ΔEFA are rotated and translated so that ΔEFA is parallel to the x-y plane and side EF is parallel to the x-axis. GRAF-pop then uses the *x, y* coordinates of point *P* and the three vertices to calculate the normalized barycentric coordinates *(α, β, γ*) of *P* against ΔEFA using the following equations:

where (*xp, yp*), (*xe, ye*), (*xf, yf*), (*xa, ya*) are the Cartesian *x, y* coordinates of *P* and the three vertices *E, F, A*, and ***T*** is the matrix

.

In practice, det(***T***) is never close to 0 because the three angles of the triangle are neither close to 0° nor close to 180°. Equivalent equations were given for a similar purpose of determining ancestry proportions in (Chen et al. 2013), but those authors did not recognize that the proportions determined by these equations have a geometric interpretation as barycentric coordinates.

After barycentric coordinates are calculated for all subjects, they are converted back to the Cartesian coordinates (*xp0, yp0*) using the reference triangle calculated by assuming all fingerprint SNPs have genotypes:

where (*xe0, ye0*), (*xf0, yf0*), (*xa0, ya0*) are the Cartesian coordinates of the three vertices *E, F, A* calculated by assuming all fingerprint SNPs have genotypes. For each subject with missing genotypes, the normalized coordinate *xe0, ye0* are called GD1 and GD2 scores, respectively. The distance from point *P* to the plane containing ΔEFA is called GD3 score. All subjects are plotted on the same graph using their (*xp0, yp0*) values, and each subject’s ancestry proportions *Pe*, *Pf*, *Pa* are estimated based on the barycentric coordinates (*α, β, γ*) using the method mentioned above.

In addition to reference populations EUR, AA, ASN, the following two groups of subjects are also used as reference populations to calculate genetic distances for each subject group of subjects self-reported as Mexicans or Latinos, and group of subjects self-reported as Asian India or Pakistanis. The difference of the two *D* scores from each subject to the above reference population, called GD4 score, is also used to determine the genetic populations of this subject.

Based on the demographics of other populations in dbGaP, we partitioned the data into eight defined populations and “Other” using the GD1, GD2 and GD4 scores. Subjects are partitioned in two steps. First, they are separated into seven groups using GD1 and GD2 scores as shown indirectly in Table 1. The definitions are indirect because the *Pe*, *Pf*, *Pa* values are derived from GD1 and GD2. Second, populations 6, 7, 8 are separated from one another using GD1 and GD4 scores (Table 2). GRAF-pop provides options to show cutoff lines on the graphs it generated (see instructions below). It also lets users to select different cutoff values.

Table 1. Grouping subjects based on the ancestry proportions

|  |  |  |
| --- | --- | --- |
| **PopID** | **Population** | **Cutoff standard** |
| 1 | European | *Pe* ≥ 87% |
| 2 | African | *Pf* ≥ 95% |
| 3 | East Asian | *Pa* ≥ 95% |
| 4 | African American | 40%≤ *Pf* < 95% and *Pa* < 13% |
| 5 | Hispanic1 | *Pf* < 40% and *Pe* < 87% and *Pa* < 13% and *Pf* ≥ *Pa* |
| 6,7,8 | (Three populations) | *Pa* < 95% and *Pe* < 87% and *Pf* < 13% and *Pf* < *Pa* |
| 9 | Other | *Pa* ≥ 13% and *Pf* ≥ 13% |

Table 2. Separating Asians and Hispanics using GD1 and GD4 scores

|  |  |  |
| --- | --- | --- |
| **PopID** | **Population** | **Cutoff standard** |
| 7 | Other Asian | GD1 > 30 × (GD4)2 + 1.73 |
| 8 | South Asian | GD4 > 5 × (GD1 -1.69)2 + 0.042 |
| 6 | Hispanic2 | GD4 < 0 and PopID is not 7 |

**Running GRAF-pop**

The population feature is implemented in the *-pop* option of the C++ program *graf* and the perl script *PlotPopulations.pl*. The genotype data of the fingerprint SNPs of the subjects of the 1000 Genome Project are saved to a PLINK set and added to the package so that users can test run the program conveniently.

The *-pop* option has been added to the instructions, which can be seen by running *graf* without parameters:

graf

Usage: graf [options]

…

-pop output file: Check subject populations and save

results to the output file

…

Example: run the following command to check populations of the 1000 Genome Project subjects and save results to output file G1000\_sbj\_dists.txt

graf -plink G1000FpGeno -pop G1000\_sbj\_scores.txt

The results generated by *graf* can be passed to *PlotPopulations.pl* for further processing. The following instructions are displayed on the screen when the script is run without parameters:

PlotPopulations.pl

Usage: PlotPopulations.pl <input file> <output file> [Options]

Note:

Output file should be either a .png file or a .txt file.

If the output file is a .png file, the script will plot the results

to a graph and save the graph to the file.

If the output file is a .txt file, the script will save the

calculated subject ancestry components to the file.

Options:

Set window size in pixels

-gw graph width

Set graph axis limits

-xmin min x value

-xmax max x value

-ymin min y value

-ymax max y value

Set a rectangle area to retrieve subjects for graph of GD1 vs. GD2

-xcmin min x value

-xcmax max x value

-ycmin min y value

-ycmax max y value

-isByd 0 or 1

0: retrieve subjects whose values are within the above

rectangle (default value)

1: retrieve subjects whose values are beyond the above

rectangle

Set population cutoff lines

-ecut proportion: cutoff European proportion dividing Europeans

from other population. Default 87%.

-fcut proportion: cutoff African proportion dividing Africans

from other population. Default 95%.

Set it to -1 to combine African and African

American populations

-acut proportion: cutoff East Asian proportion dividing East

Asians from other populations. Default 95%.

Set it to -1 to combine East Asian and Other Asian

populations

-ohcut proportion: cutoff African proportion dividing Hispanics

from Other population. Default 13%.

-fhcut proportion: cutoff African proportion dividing Hispanics

from African Americans. Default 40%.

Select some self-reported populations (by IDs) to be highlighted on

the graph

-pops comma separated population IDs,

e.g., -pops 1,3,4 -> highlight populations #1, #3 and #4

Select self-reported populations (by IDs) to show areas including 95%

dbGaP subjects with genotypes of at least 4000 fingerprint SNPs

-areas comma separated dbGaP self-population IDs,

e.g., -areas 1,3

-> show areas that include 95% dbGaP subjects with

self-reported populations #1 and #3

1: European/White/Caucasian

2: African (Ghana/Yoruba)

3: East Asian (Chinese/Japanese)

4: African American/Black

5: Puerto Rican/Dominican

6: Mexican/Latino

7: Asian/Pacific Islander

8: Asian Indian/Pakistani

Select which score to show on the y-axis

-gd4 1 or 0. 1: show GD4 on y-axis; 0: show GD2

Set population cutoff lines

-cutoff 1 or 0. 1: show cutoff lines; 0: hide cutoff lines

Rotate the plot with respect to the x-axis by a certain angle

-rotx angle in degrees

Set the size (diameter) of each dot that represents each subject

-dot pixels

The input file with self-reported subject race information

-spf a file with two columns: subject and self-reported

population

The script takes two required parameters, which must be the first two arguments and are not preceded by flags, unlike all the optional arguments, which are preceded by a flag. The first parameter should be the name of the file that is generated by *graf -pop* option and contain subject genetic distance scores. The second parameter is the output file, expected to be either a .png or .txt file. If the output file is a .png file, the script processes the scores and saves the results to the output file. The default graph is GD1 vs. GD2, e.g.,

PlotPopulations.pl G1000\_sbj\_scores.txt G1000\_sbj\_pops.png

When option gd4 is set to 1, the script generates a graph of GD1 vs. GD4:

PlotPopulations.pl G1000\_sbj\_scores.txt G1000\_sbj\_pops.png

-spf G1000SbjSuperPop.txt -gd4 1

If the output file is a .txt file, the script processes the data and saves the results to the output file in a format of a rectangular table.

PlotPopulations.pl G1000\_sbj\_scores.txt G1000\_sbj\_list.txt

In the output file, columns P\_e, P\_f, P\_a show each subject’s African, European, and East Asian proportions *Pe*, *Pf*, *Pa*, in percentages. The populations determined by GRAF-pop are included in the last two columns as an identifier and as the full name of the population.

When self-reported ancestries are available, the information can be passed to the script with *-spf* option so that the script can color-code the subjects using the self-reported ancestries, e.g.,

PlotPopulations.pl G1000\_sbj\_scores.txt G1000\_sbj\_pops.png

-spf G1000SbjSuperPop.txt

The input ancestry file should have two columns without column header, containing the sample IDs and self-reported ancestries, in this order. In the graph generated by the script, the ancestries are numbered and color coded.

The cutoff lines used to partition the subjects are drawn on the graphs when option -cutoff is set, e.g.,

PlotPopulations.pl G1000\_sbj\_scores.txt G1000\_sbj\_pops.png

-spf G1000SbjSuperPop.txt -cutoff 1

PlotPopulations.pl G1000\_sbj\_scores.txt G1000\_sbj\_pops.png

-spf G1000SbjSuperPop.txt -gd4 1 -cutoff 1

If multiple subjects appear at the same locations in the x-y plane, the user can use option *-pops* to bring some ancestries to the front, while setting some ancestries to the back and fade out them in the graph. For example, the following command generates a graph with the ancestry No. 5 (AMR, standing for Ad Mixed American) in the back and colored yellow: The assignments of colors to populations are currently hard-coded.

PlotPopulations.pl G1000\_sbj\_scores.txt G1000\_sbj\_pops.png

-spf G1000SbjSuperPop.txt -pops 1,3,2,4

The ancestry numbers following -pops should be separated by commas without spaces.

One can also use the -*rotx* option to rotate the graph of GD1 vs. GD2 around *x*-axis by a certain angle specified in degrees (can be any real number). For example, the following command generates a graph showing the subjects rotated by 90o:

PlotPopulations.pl G1000\_sbj\_scores.txt G1000\_sbj\_pops.png

-spf G1000SbjSuperPop.txt -rotx 90

Options *-gw, -xmin, -xmax, -ymin, -ymax, -dot*, similar to those in *PlotGraf.pl*, can be used to adjust the graph size, specify axis limits, and set the dot size, e.g.,

PlotPopulations.pl G1000\_sbj\_scores.txt G1000\_sbj\_pops.png

-spf G1000SbjSuperPop.txt -gw 800 -ymin 1.1 -dot 5

PlotPopulations.pl G1000\_sbj\_scores.txt G1000\_sbj\_pops.png

-spf G1000SbjSuperPop.txt -gw 800 -gd4 1 -ymin -0.2

One can use the option -*areas* to select populations to show the expected oval areas that include 95% of dbGaP subjects with at least 4000 fingerprint SNPs with genotypes, e.g.,

PlotPopulations.pl G1000\_sbj\_scores.txt G1000\_sbj\_pops.png

-spf G1000SbjSuperPop.txt -areas 1,4,7

The integers in the comma delimited string represent the eight self-reported ancestry groups in dbGaP, with most common ancestry terms in each group shown below:

1: European/White/Caucasian

2: African (Ghana/Yoruba)

3: East Asian (Chinese/Japanese)

4: African American/Black

5: Puerto Rican/Dominican

6: Mexican/Latino

7: Asian/Pacific Islander

8: Asian Indian/Pakistani

GRAF-pop uses the ancestry proportions shown in Table 1 as default cutoff values. The user can use options *-ecut, -fcut, -acut, -ohcut, -ahcut, -fhcut* to set the cutoff values to different numbers, e.g.,

PlotPopulations.pl G1000\_sbj\_scores.txt G1000\_sbj\_pops.png

-spf G1000SbjSuperPop.txt -cutoff 1

-fcut 85 -ahcut 80 -ohcut 15.5

When -*fcut* or -*acut* are set to negative values, the African or East Asian cutoff line is not plotted on the graph, and the script does not distinguish Africans from African Americans, or East Asians from Other Asians, e.g.,

PlotPopulations.pl G1000\_sbj\_scores.txt G1000\_sbj\_pops.png

-spf G1000SbjSuperPop.txt -fcut -1

As mentioned above, when the second parameter (the output file) is a .txt file, the script saves subjects and the ancestry proportions into a rectangular table. Options *-xcmin, -xcmax, -ycmin, -ycmax, -isByd* can be used to specify a rectangular area and let the script to retrieve subjects whose *x*(GD1)*, y* (GD2) scores are either within or beyond this area. For example, the following command saves all subjects with 1.8 < GD1 < ∞ and -∞ < GD2 < 1.2, which are all the EAS (East Asian) subjects:

PlotPopulations.pl G1000\_sbj\_scores.txt G1000\_sbj\_list.txt

-spf G1000SbjSuperPop.txt -xcmin 1.8 -ycmax 1.2

When option *-isByd* is set to 1, the script retrieves subjects whose value are beyond rectangular area specified by options *-xcmin, -xcmax, -ycmin, -ycmax*. For example, the following command excludes most of the 1000 Genome Subjects with super populations AMR (Ad Mixed American) and SAS (South Asian):

PlotPopulations.pl G1000\_sbj\_scores.txt G1000\_sbj\_list.txt

-spf G1000SbjSuperPop.txt -xcmin 1.64 -xcmax 1.8

-ycmin 1.24 -ycmax 1.36 -isByd 1

**References**

1. [Jin, Y](https://www.ncbi.nlm.nih.gov/pubmed/?term=Jin%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=28609482), [Schäffer, A.A](https://www.ncbi.nlm.nih.gov/pubmed/?term=Sch%C3%A4ffer%20AA%5BAuthor%5D&cauthor=true&cauthor_uid=28609482)., [Sherry, S.T](https://www.ncbi.nlm.nih.gov/pubmed/?term=Sherry%20ST%5BAuthor%5D&cauthor=true&cauthor_uid=28609482)., [Feolo M](https://www.ncbi.nlm.nih.gov/pubmed/?term=Feolo%20M%5BAuthor%5D&cauthor=true&cauthor_uid=28609482). Quickly identifying identical and closely related subjects in large databases using genotype data. [*PLoS One*](https://www.ncbi.nlm.nih.gov/pubmed/?term=Jin+shaffer+GRAF)**12(6):**e0179106 (2017).
2. Coxeter, H. S. M. Introduction to geometry (2nd ed.). John Wiley and Sons. pp. 216–221 (1969).
3. Chen, C.-Y., Pollack, S., Hunter, D.J., Hirschhorn, J.N., Kraft, P., Price, A.L. Improved ancestry inference using weights for external reference panels. Bioinfomatics **29**:1399-1406 (2013).