User Guide for GenoGeographer (0.3.1)

Claus Børsting, Vania Pereira, Torben Tvedebrink

Introduction

GenoGeographer (Tvedebrink et al., 2017; 2018; 2019) is a software used for population assignment of an individual in a forensic genetic context. Typical ancestry inference studies of a population or individual rely, to some extent, on the knowledge of the history of the population and its parental ancestry components. The ancestry investigations aim to ascertain the proportion of each parental ancestry component in the population or individual. In a forensic context, however, the biogeographic ancestry of an individual is generally unknown, and a population assignment is carried out instead. Here, the profile is evaluated in several reference populations present in a database to assess the most likely biogeographical origin of the profile.

GenoGeographer provides two types of analyses after the DNA profile with pre-selected Ancestry Informative Markers (AIMs) from the person of interest is uploaded to the software:
First, it performs an outlier test (z-score test) of the AIM profile in all reference populations, where the null hypothesis H_0 = 'The AIM profile originates from the reference population' is either accepted or rejected using a one-sided 95% confidence level. An AIM profile with z-score ≤1.64 lies within the confidence interval, and H_0 is accepted, whereas an AIM profile with z-score >1.64 is considered an outlier and H_0 is rejected.
Second, it performs calculations of the evidential weights under user-defined hypotheses in the form of likelihood ratios (LRs). $LR = P(E|H_1)/P(E|H_2)$, where H_1 = the tested individual belongs to population A, and H_2 = the tested individual belongs to population B. LRs should only be calculated if the outlier tests indicate that the database with reference populations includes at least one likely population of origin and population A must be the most likely population (z-score ≤1.64).

GenoGeographer includes reference population databases for four panels of AIMs: The Precision ID Ancestry panel (Pereira et al., 2017; Mogensen et al., 2020; Köksal et al., 2023), the Seldin Panel (Kosoy et al., 2009), The Kidd panel (Kidd et al., 2014), and the VISAGE basic tool (Xavier et al., 2020).

A stepwise protocol with screen-dump examples from GenoGeographer is given below:

Work-flow

1. Open GenoGeographer (Genogeographer (aau.dk)).
2. Click on 'Analyse AIMs profile' (top of the screen) and upload the AIM profile (.csv format) by clicking on 'Browse…' and selecting the AIM profile of interest. The csv file should contain one column with locus names (e.g. rs10007810) and one column with genotypes (e.g. AA or CT). The locus and genotype columns may be defined after uploading the file using the drop-down menus on the left of the screen.
3. Define the settings for the analysis (Figure 1) and click 'Analyse!'.

Figure 1.

Grouping type: 'Meta-populations' should be chosen, since the AIMs were originally selected for separation of continental populations. However, selection of individual 'Populations' may provide the experienced analyst with some hints to the bio-geographic origin of the person of interest. Meta-populations are groups of individual populations that were grouped according to the STRUCTURE analysis shown under 'Reference Databases' (top of the screen). There are six major meta-populations: Sub-Saharan Africa, North Africa, Europe, Middle East, South-Central Asia, and East Asia.

'Minimum sample-size' of the reference populations should be 75 individuals to allow reasonable estimates of allele frequencies.

Simulated two-way 1:1 admixtures of the meta-populations may be considered as reference populations in the analysis by ticking the 'Admixture' box. This should not be the first choice, but may be useful depending on the results from the outlier test or if specific hypotheses involving admixed populations need to be considered.

The confidence interval of the calculated population likelihoods and LRs should be set to 95%. If two population likelihoods are not significantly different according to the confidence interval, this will be indicated in the LR table (see Figure 6).

The proper reference database should be selected in the 'Select reference database' drop-down menu.

If needed, adjust the p-values by exponential tilting by ticking the 'Adjust' box. This will slow the analysis and is generally not necessary.

4. Graphics (Figure 2) and tables (Figure 3 and 4) of the GenoGeographer analysis are available for interpretation.

Graphics

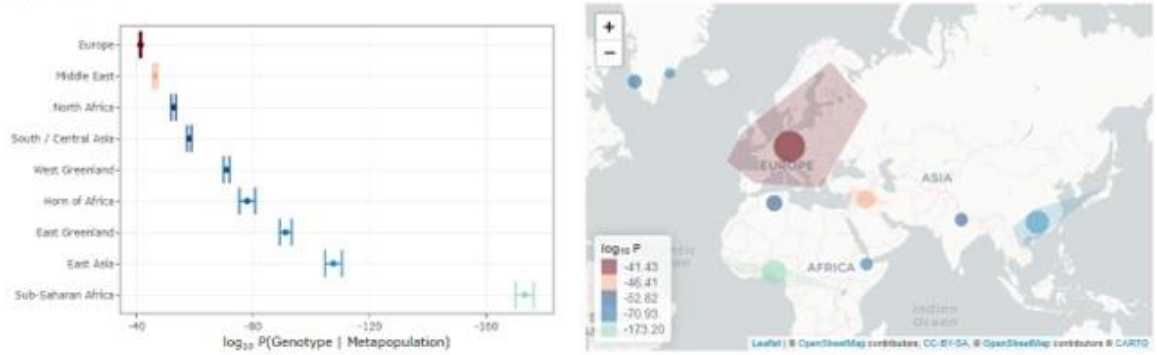

Figure 2. Log10 to the population likelihoods of the six metapopulations and three admixed populations (West Greenland, East Greenland, and Horn of Africa) are indicated with confidence intervals on the left. The three admixed populations are shown here by default, because they are not included in any of the metapopulations. On the right is shown a map that indicate the most likely population of origin using the same colours as in the diagram on the left.
In the example shown, the most likely population of origin is Europe.

Tables

Copy CSV Excel PDF Print Show 10 entries Search:

metapopulation	$\log_{10}$ P(G|Metapopulation)	CI[$\log_{10}$ P(G|Metapopulation)]	z-score	p-value
Europe	-41.434	[-41.717; -41.152]	0.219	0.414
Middle East	-46.409	[-47.009; -45.808]	1.196	0.116
North Africa	-52.818	[-53.603; -52.033]	2.969	0.001
South / Central Asia	-58.15	[-58.868; -57.433]	4.253	0
West Greenland	-70.929	[-71.914; -69.945]	7.057	0
Horn of Africa	-78.024	[-80.723; -75.325]	11.304	0
East Greenland	-91.169	[-93.218; -89.121]	14.519	0
East Asia	-107.606	[-110.465; -104.748]	20.767	0
Sub-Saharan Africa	-173.202	[-176.302; -170.102]	51.077	0

Figure 3. Log10 to the population likelihoods of the six metapopulations and three admixed populations are indicated with confidence intervals. The results from the outlier test (z-score test) are shown in the two right-most columns. The colours are the same as in figure 2.
If the z-score ≤1.64, the AIM profile is not considered an outlier in the tested population. The p-value is the probability that the null hypothesis of the z-score test is true. If the p-value <0.05 (z-score >1.64), the null hypothesis is rejected and the AIM profile is considered an outlier in that reference population.
In the example shown, the AIM profile may originate from two metapopulations (Europe and Middle East), whereas it is unlikely that the AIM profile belongs to any of the other populations in the table. Thus, it is reasonable to proceed with the calculation of evidential weights of population assignments (Figure 4). If there were no likely populations among the reference populations (z-score >1.64 for all reference populations), population assignment would be impossible and should not be attempted (with Genogeographer, STRUCTURE, or other tools).

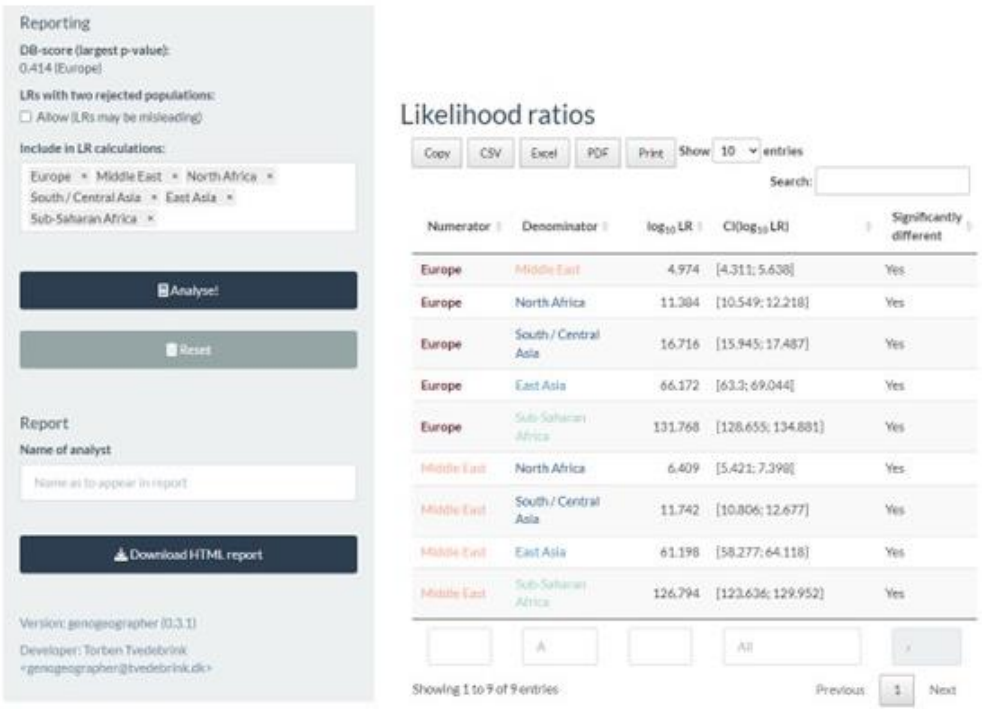

Figure 4. Log10 to LRs are shown with confidence intervals.
LR = $P(E|H_1)/P(E|H_2)$, where H_1 = the tested individual belongs to population A, and H_2 = the tested individual belongs to population B. Relevant populations may be included using the 'Include in LR calculation' box on the left of the screen. By default, population A must be a likely population (z-score <1.64). This may be circumvented by ticking the 'LRs with two rejected populations' box (not recommended).
In the example shown, nine LRs were calculated; five, where Europe was population A and one of the other metapopulations was population B, and four, where Middle East was population A and one of the remaining metapopulations (except Europe) was population B. The smallest LR = 9.42E+4 was obtained when Europe was population A and Middle East was populations B. This was expected since Europe and Middle East were the only two likely populations of origin (Figure 3).

5. The results may be presented in a report in various ways, e.g. in the form of a table, as shown below, in the form of text, or a combination of both.
Some of the LRs in the table are astronomical and it may be useful to define a maximum reported LR, e.g. LR >10,000 or LR >100,000. The LRs with Middle East as population A are not shown in the table, because they do not include the most likely population of origin (Europe) and could be misleading. F.ex. it is > 1,000,000 times more likely to have the AIM profile, if the DNA originates from from an individual that belongs to the Middle Eastern metapopulation than if the DNA originates from an individual that belongs to the North African metapopulation (Figure 4). True, but irrelevant and misleading, because Europe is not considered as a possible population.

The text may be phrased as follows: "*The AIM profile indicates that the DNA originates from an individual from the European metapopulation.*" and "*It is more than 10,000 times more likely to have the AIM profile, if the DNA originates from an individual that belongs to the European metapopulation than if the DNA originates from an individual that belongs to the Middle Eastern metapopulation, the North African metapopulation, the*

South-Central Asian metapopulation, the East Asian metapopulation, or the Sub-Saharan African metapopulation."

LR = $P(E\|H_1)/P(E\|H_2)$		H_1 = Population A Europe
H_2 = Population B	Middle East	9.42E+04
	North Africa	2.42E+11
	South / Central Asia	5.20E+16
	East Asia	1.49E+66
	Sub-Saharan Africa	5.86E+131

6. The example shown above was fairly straightforward and the individual most likely had a European background.

The example below is more complicated as it is an example of admixed population background. The z-score test (not shown) indicated only one likely reference population: North Africa (z-score = 0.545 (p-value: 0.293)). Figure 5 shows the calculated LRs. The smallest LR was as low as 1,510 (Population A: North Africa; Population B: Middle East), and since both hypotheses involved an admixed population (North Africa and Middle East), an additional analysis was performed, where the box 'Admixture' was ticked off (Figure 6).

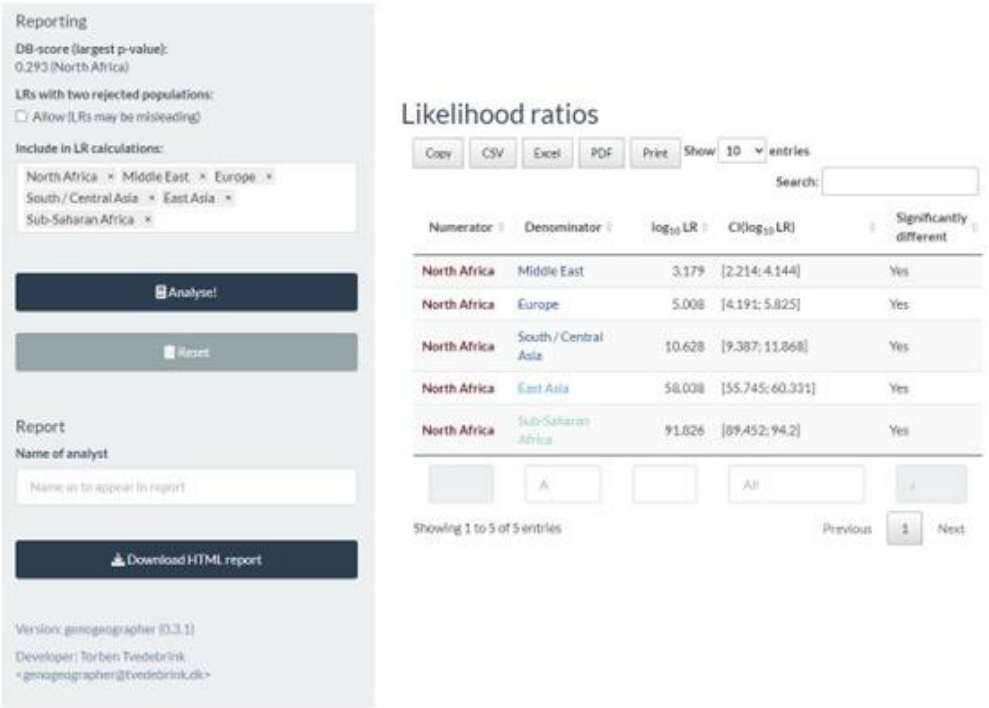

Figure 5. Log10 to LRs are shown with confidence intervals.
In the example shown, five LRs were calculated; five, where North Africa was population A and one of the other metapopulations was population B. The smallest LR = 1.51E+03 was obtained when North Africa was population A and Middle East was populations B.

The outlier tests revealed that three two-way 1:1 admixture populations were possible populations of origin (North Africa & South/Central Asia (z-score = 0.85 (p-value: 0.198)), Middle East & North Africa (z-score = 1.20 (p-value: 0.114)), and Europe & North Africa (z-score =1.554 (p-value: 0.06)). Thus, a total of 26 LRs were calculated with all combinations of hypotheses involving the six metapopulations and the three likely two-way 1:1 admixture populations (Figure 6).

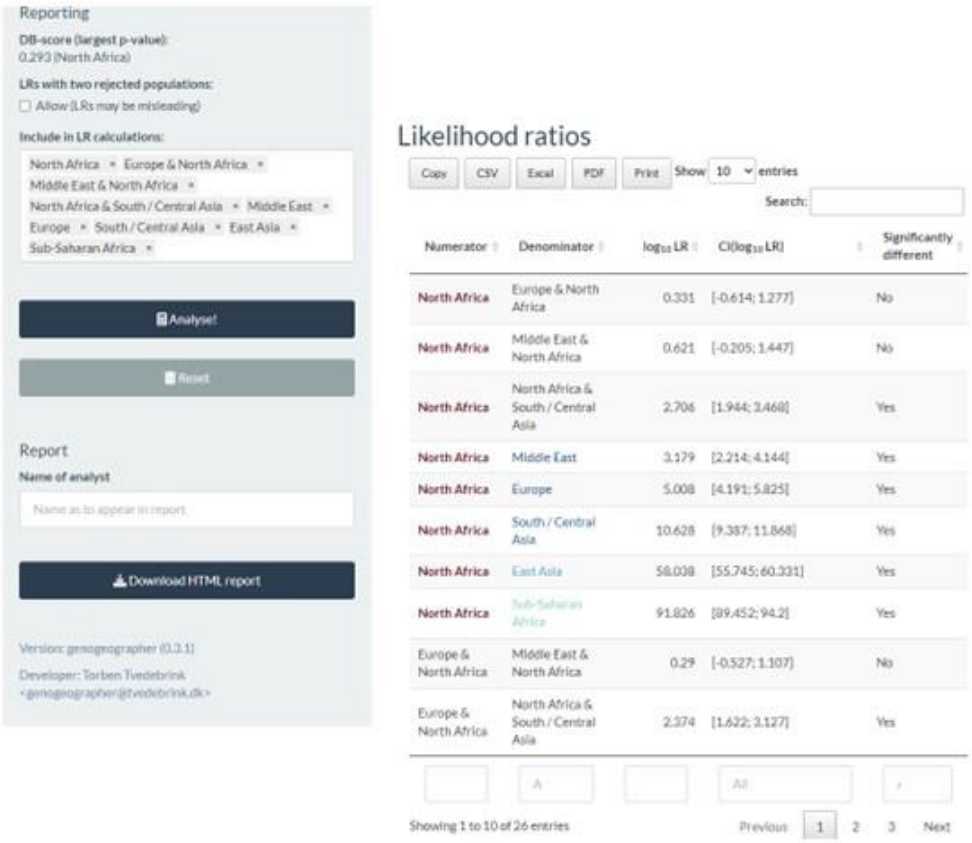

Figure 6. Log10 to LRs are shown with confidence intervals.
Only 10 of the 26 calculated LRs are shown. Of the eight LRs where North Africa was selected as population A, the population likelihoods of two alternative hypotheses (Europe & North Africa and Middle East & North Africa) were not significantly different from the North African population likelihood (right side of table). The LRs were 2.1 and 4.2, respectively.

The results clearly indicated that there were alternative admixed populations that were as likely an origin as the North African metapopulation. This should be clearly explained in the report both in the table with LRs and in the complimentary text. However, an LR table with 26 LRs in four columns of likely populations (with z-score <1.64) may be overwhelming to interpret. It may be considered to reduce the table to the LRs involving the most likely population as population A, e.g. North Africa (see below).

The result summary may be phrased as: "*The AIM profile indicates that the DNA originates from an individual from the North African metapopulation or from another admixed population with contributions from North Africa, Middle East, South/Central Asia or Europe*."

LR = $P(E\|H_1)/P(E\|H_2)$		H_1 = Population A North Africa
H_2 = Population B	Europe & North Africa	Inconclusive*
	Middle East & North Africa	Inconclusive*
	North Africa & South / Central Asia	508†
	Middle East	1510
	Europe	1.02E+05
	South / Central Asia	4.25E+10
	East Asia	1.09E+58
	Sub-Saharan Africa	6.70E+91

*The population likelihoods were equally likely

†The weight of the evidence is considered to be low, when the LR<1000

Literature.

R Kosoy, R Nassir, C Tian, PA White, LM Butler, G Silva, R Kittles, MERiquelme, PK Gregersen, JW Belmont, FM De La Vega, MF Seldin (2009). *Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America.* Human Mutatation 30, 69-78.

KK Kidd, WC Speed, AJ Pakstis, MR Furtado, R Fang, A Madbouly, M Maiers, M Middha, FR Friedlaende, JR Kidd (2014). *Progress toward an efficient panel of SNPs for ancestry inference.* Forensic Science International: Genetics 10, 23-32.

T Tvedebrink, PS Eriksen, HS Mogensen, N Morling (2017). *GenoGeographer - A tool for genogeographic inference.* Forensic Science International: Genetics Supplement Series 6, e463-e465.

V Pereira, HS Mogensen, C Børsting, N Morling (2017). *Evaluation of the Precision ID Ancestry Panel for crime case work: A SNP typing assay developed for typing of 165 ancestral informative markers.* Forensic Science International: Genetics 28, 138-145.

T Tvedebrink, PS Eriksen, HS Mogensen, N Morling (2018). *Weight of the evidence of genetic investigations of ancestry informative markers.* Theoretical Population Biology 120, 1-10.

T Tvedebrink, PS Eriksen (2019). *Inference of admixed ancestry with Ancestry Informative Markers.* Forensic Science International: Genetics 42, 147-153.

HS Mogensen, T Tvedebrink, C Børsting, V Pereira, N Morling (2020). *Ancestry prediction efficiency of the software GenoGeographer using a z-score method and the ancestry informative markers in the Precision ID Ancestry Panel.* Forensic Science International: Genetics 44, 102154.

C Xavier, M de la Puente, A Mosquera-Miguel, A Freire-Aradas, V Kalamara, A Vidaki, TE Gross, A Revoir, E Pośpiech, E Kartasińska, M Spólnicka, W Branicki, CE Ames, PM Schneider, C Hohoff, M Kayser, C Phillips, W Parson, VISAGE Consortium (2020). *Development and validation of the VISAGE AmpliSeq basic tool to predict appearance and ancestry from DNA.* Forensic Science International: Genetics 48, 102336.

Z Köksal, OL Meyer, JD Andersen, L Gusmão, HS Mogensen, V Pereira, C Børsting (2023). *Pitfalls and challenges with population assignments of individuals from admixed populations: applying GenoGeographer on Brazilian individuals.* Forensic Science International: Genetics 67, 102934.