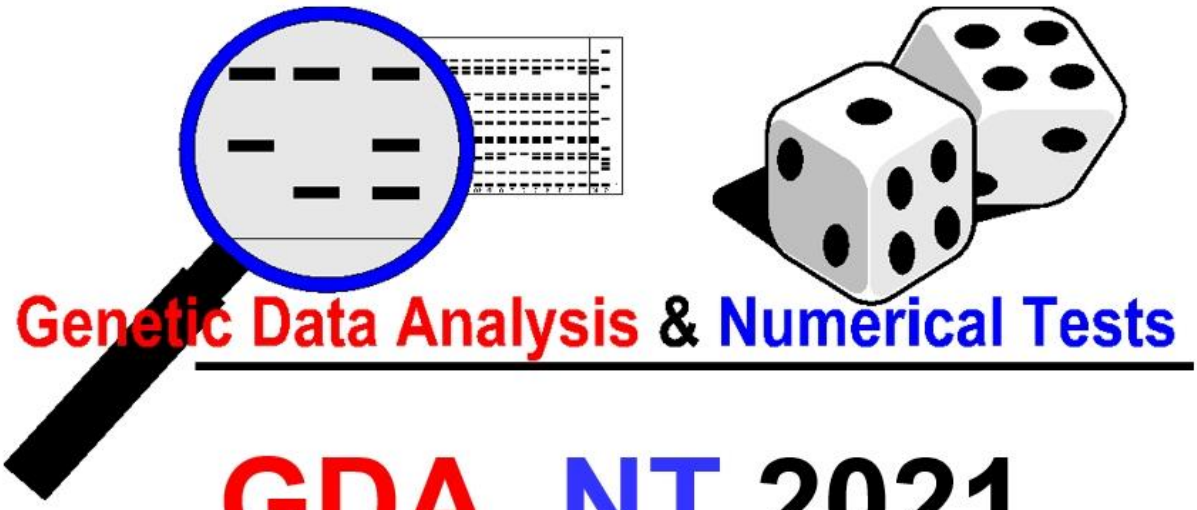


GDA_NT 2021

User's manual

From:

Bernd DEGEN

Thuenen-Institute of Forest Genetics
Sieker Landstrasse 2
22927 Grosshansdorf
Germany
e-mail: bernd.degen@thuenen.de

Last updated: 5th of August 2022

How to cite the program?

Degen, B. (2022): GDA-NT 2021 a computer program for population genetic data analysis and assignment. Conservation Genetics Resources. DOI: 10.1007/s12686-022-01283-2

Contents

1 Introduction

The program GDA_NT 2021 can be used for three kinds of applications:

a) Analysis of population genetic measures
b) Assignment of genotypes of unknown origin to reference populations
c) Quality check of gene markers for population genetic studies

For each of these applications short videos are placed at the web-page of the program.

The population genetic measures are subdivided in parameters on genetic composition of populations (frequencies of alleles, single locus genotypes and multilocus genotypes), genetic differentiation (genetic distance and measures on population differentiation and fixation) and genetic variation. One speciality of the program is that I added to the broadly used measures such as F_{ST} a few measures developed H.-R. Gregorius such as the genetic distance D_0 (Gregorius, 1984), population differentiation *Dj* (Gregorius, 1987) and evenness *E* (Gregorius, 1990) and the standardized F_{ST} of Hedrick (Hedrick, 2005; Meirmans and Hedrick, 2011). Further I added the option to measure the kinship-coefficient among pairs of individuals within populations (Loiselle *et al.*, 1995). The statistical significance of the measures on genetic differentiation among populations and the significance of the departure from Hardy-Weinberg-heterozygosity can be evaluated with numerical tests. A few measures have the option to include information on geographic positions into the calculations in order to define geographically defined sub-populations (e.g. for *Dj* and the computation of allele frequencies for the kinship coefficient).

The part on the genetic assignment has elements that can also be find in the program GeneClass (Piry *et al.*, 2004) such as the Bayesian method (Rannala and Mountain, 1997) and the allele frequency based approach (Paetkau *et al.*, 1995). I added other options on the genetic distance based assignment and the possibility to combine diploid and haploid gene loci. GDA_NT 2021 further offers more options to simulate genotypes for the calculation of exclusion probabilities and to do self-assignment tests.
GDA_NT 2021 offers the possibility to run a quality check of gene markers in order to select those gene markers that are suitable and informative for the population genetic studies. This can be used to select or exclude gene loci that show questionable values (e.g. high level of missing data, low level of polymorphism or excess of homozygotes or heterozygotes compared to Hardy-Weinberg expectations). This option is also interesting in frame of the development of new gene markers. Using next generation DNA-sequencing techniques it has become

relatively easy to develop new sets of gene markers – in most cases SNPs – for population genetic studies (Pakull *et al.*, 2016; Degen *et al.*, 2021).

The program works with data of co-dominant gene markers of diploid organisms (e.g. nSNPs, nSSRs) and haplotype data (e.g. cpSNPs, cpSSRs).

2 Obtaining the program and manual

GDA_NT 2021 has been programmed in visual basic and compiled as 32-bit versions for the operating system Microsoft Windows (Windows 10 and earlier versions). The program, a zip-file with different input data files for demonstration, different videos that explain the use of the program and the User's manual are available on our internet homepage:

https://www.thuenen.de/en/institutes/forest-genetics/software/gda-nt-2021

3 Installation

Download the file "GDA_NT_2021_setup.exe". You start the setup procedure with double mouse click on the file "GDA_NT_2021_setup.exe". After successful setup the program "GDA_NT 2021" will be added to your program list.

Input data for demonstration

The zip-file “Demo_files.zip” has input files for the different applications. These files are used in this manual and in the videos.

4 Structure of the program

The program has five different menus (see figure 1):

- "File" => open the input files, import data into the format of input files, save sub-selections of populations and loci as new input files, export data to the program SGS and exit the program
- "Analysis" => change selection on populations and loci, define mode of inheritance of loci, analysis of population data, genetic assignment
- "Results" => save frequencies of alleles as csv-files, open text editor to visualise the results of the output files, save allele frequencies as input to the program SGS

- “Test” => start numerical tests: a) within populations (Test on Hardy-Weinberg heterozygosity), b) between pairs of two populations (genetic distances), and c) among all populations (measures of differentiation and population fixation)
- "Help" => to be informed about the program and to get examples for the format of the input files

The structure represents the order of user's activities. First, you have to load the input file (one file for the analysis of population data, two files for the genetic assignment) in the menu ("File"). Then in the menu ("Analysis") you might modify the selection of populations and loci and call the window for the data analysis. Thereafter, you want to view the results ("Results") or run numerical tests (“Test”). The program follows these steps by activating (black colour of letters) or deactivating its functions (grey colour of letters).

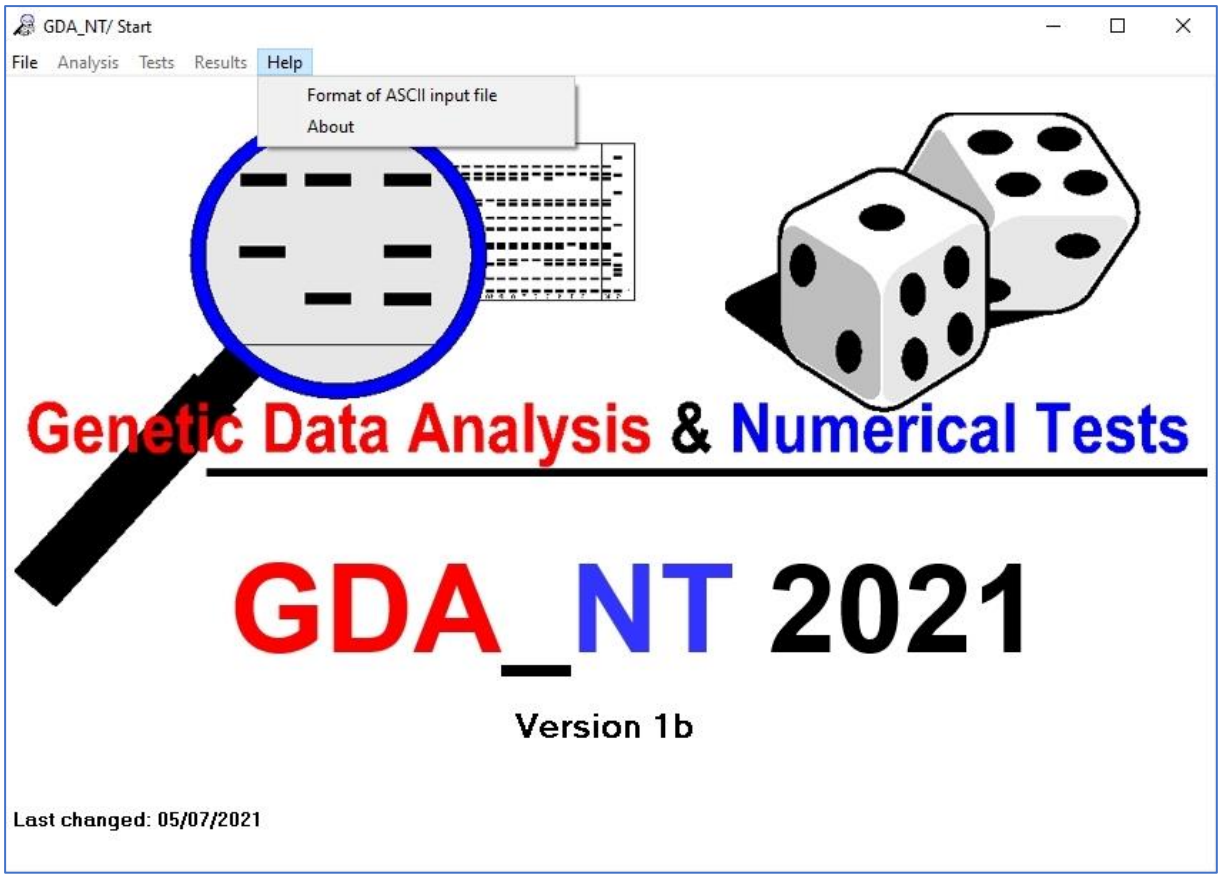


Figure 1: Structure of the program GDA_NT 2021

5 Menu “File”

5.1 Input data

The first thing to do is open an input file. Depending on the application the user needs to open one or two input files. Only one file needs to be opened for the genetic analysis of population data (figure 2). The genetic assignment requests two input files: one with the data of the reference populations (“Open file with reference pops for assignment”) and another on with the genetic data of the tested groups (“Open file with test groups for assignment”). The format of the input files is always the same (figure 3). The two files for the genetic assignment need to have the same names and order of the gene loci.

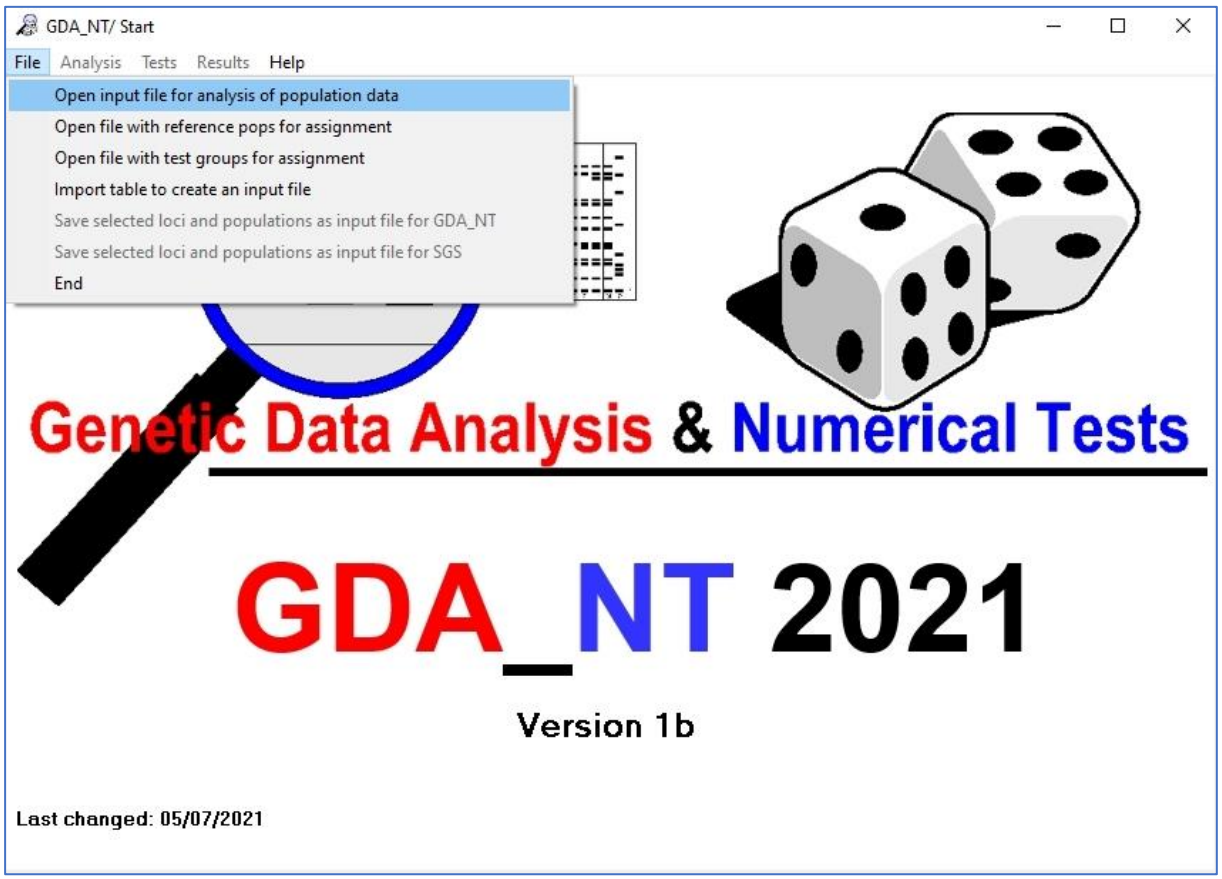


Figure 2: Open the input files

Note: The different alleles must be coded by numerical values. There is no restriction in terms of number of alleles. Haplotypes must be coded as homozygotes.

```
Demo Populations                    <= title data set
3                                   <= number of populations
5                                   <= number of gene loci
4                                   <= largest number to code an allele
SNP_001                             <= name of first locus
SNP_002
SNP_003
SNP_004
SNP_005
Pop_1,28.0252,53.8763               <= name of the first population
3,3,3,3,2,2,3,3,-1,-1               <= first allele of first locus
3,3,4,3,2,2,3,3,-1,2                <= missing data
3,3,4,4,2,2,3,3,2,2

Pop_2,23.7928,53.6858               <= x- and y-position of second population
3,3,4,3,2,2,3,3,2,2                    (or longitude & latitude)
3,3,4,3,2,2,3,3,2,2
3,3,4,3,2,2,3,3,2,2
                                    <= blanc line to separate two populations
Pop_3,46.4746,52.455
3,3,4,3,2,2,3,3,2,2
3,3,4,3,2,2,3,3,2,2
3,3,4,4,2,2,3,3,2,2
END                                 <= indication of end of data set
```

Figure 3: Format "Multilocus Genotypes/ Haplotypes of Individuals"

5.2 Import table to create an input file

In order to simplify the creation of input files, the user can import tables with the data created by other programs such as EXCEL. If you use EXCEL, save the data as a “csv-file” and make sure that “,” has been used to separate the data fields and that “.” is used as decimal symbol. The structure of such a table is shown in figure 4. Each individual is a row in the table. The first column gives the ID of the populations, the second column is a name of the group to which the population belongs, followed with a column on the ID of the individuals, two columns on the geographic position of the individual (can be the same of all individuals of the same population) and then for each allele a column. The first row includes the names of the gene loci.

If you go to “File/Import table to create an input file” a new window is opened (figure 5). First add in the text field a title of the data set. Then press the bottom “Import data” and open the csv-file with the data. After successful reading of the table, you can save the data as an input file for GDA_NT 2021 either structured as data for each population or with data aggregated according to the groups.

	A	B	C	D	E	F	G	H	I	J	K
1	Pop_ID	Country	Ind_ID	Longitude	Latitude	SNP_001	SNP_001	SNP_002	SNP_002	SNP_003	SNP_003
2	Pop_1	Belarus	1	28.0252	53.8763	3	3	3	3	2	2
3	Pop_1	Belarus	2	28.0252	53.8763	3	3	4	3	2	2
4	Pop_1	Belarus	3	28.0252	53.8763	3	3	4	4	2	2
5	Pop_1	Belarus	4	28.0252	53.8763	3	3	3	3	2	2
6	Pop_1	Belarus	5	28.0252	53.8763	3	3	4	4	2	2
7	Pop_1	Belarus	6	28.0252	53.8763	3	3	4	3	2	2
8	Pop_1	Belarus	7	28.0252	53.8763	3	3	4	3	2	2
9	Pop_1	Belarus	8	28.0252	53.8763	3	3	4	3	2	2
10	Pop_1	Belarus	9	28.0252	53.8763	3	3	3	3	2	2
11	Pop_1	Belarus	10	28.0252	53.8763	3	-1	3	3	2	2
12	Pop_2	Belarus	11	30.6759	55.072	3	3	4	3	2	2
13	Pop_2	Belarus	12	30.6759	55.072	3	3	4	3	2	2
14	Pop_2	Belarus	13	30.6759	55.072	3	3	4	3	2	2
15	Pop_2	Belarus	14	30.6759	55.072	3	3	4	3	2	2
16	Pop_2	Belarus	15	30.6759	55.072	3	3	4	3	2	2
17	Pop_2	Belarus	16	30.6759	55.072	3	3	4	-1	2	2
18	Pop_2	Belarus	17	30.6759	55.072	3	3	4	3	2	2
19	Pop_2	Belarus	18	30.6759	55.072	3	3	4	3	2	2
20	Pop_2	Belarus	19	30.6759	55.072	3	3	3	3	2	2
21	Pop_2	Belarus	20	30.6759	55.072	3	3	4	4	2	2
22	Pop_3	Belarus	21	24.7367	53.604	3	3	4	4	2	2
23	Pop_3	Belarus	22	24.7367	53.604	3	3	4	3	2	2
24	Pop_3	Belarus	23	24.7367	53.604	3	3	4	3	2	2
25	Pop_3	Belarus	24	24.7367	53.604	3	3	4	3	2	2

Figure 4: Example for a table with data ready for the import into GDA_NT 2021

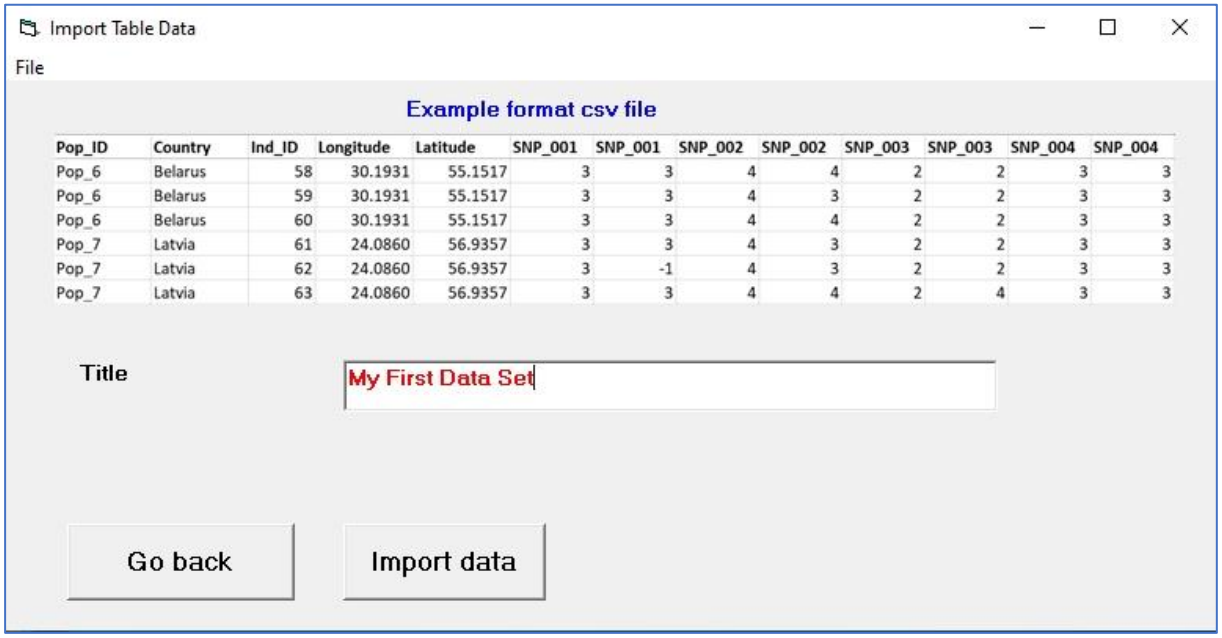


Figure 5: Window to import data from a csv-file to create a new input file

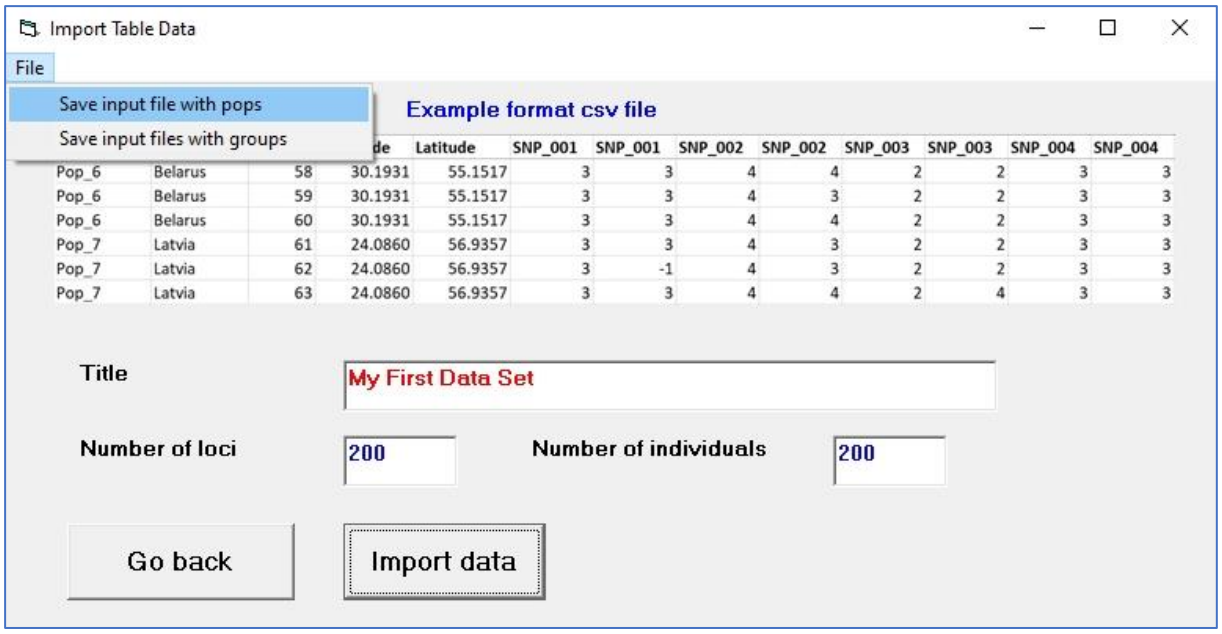


Figure 6: Save options in the window to import data from a csv-file to create a new input file

5.3 Export and saving as new input file

Once the user has modified the selection of populations and/or loci the selected part of the data set can be saved as a new input file ("Save selected loci and populations as input file for GDA_NT"). Further there is the option to export the input data into the input format of the program SGS ("Save selected loci and populations as input file for SGS") which allows detailed analysis of the spatial genetic structure. The program SGS is available at:

https://www.thuenen.de/en/fg/software/sgs/

6 Menu "Analysis"

In the menu "Analysis" the user can select populations and loci, define the mode of inheritance of the loci (only relevant for routines of the assignment) and call the windows "Analysis of population data" and "Assignment" (figure 6).

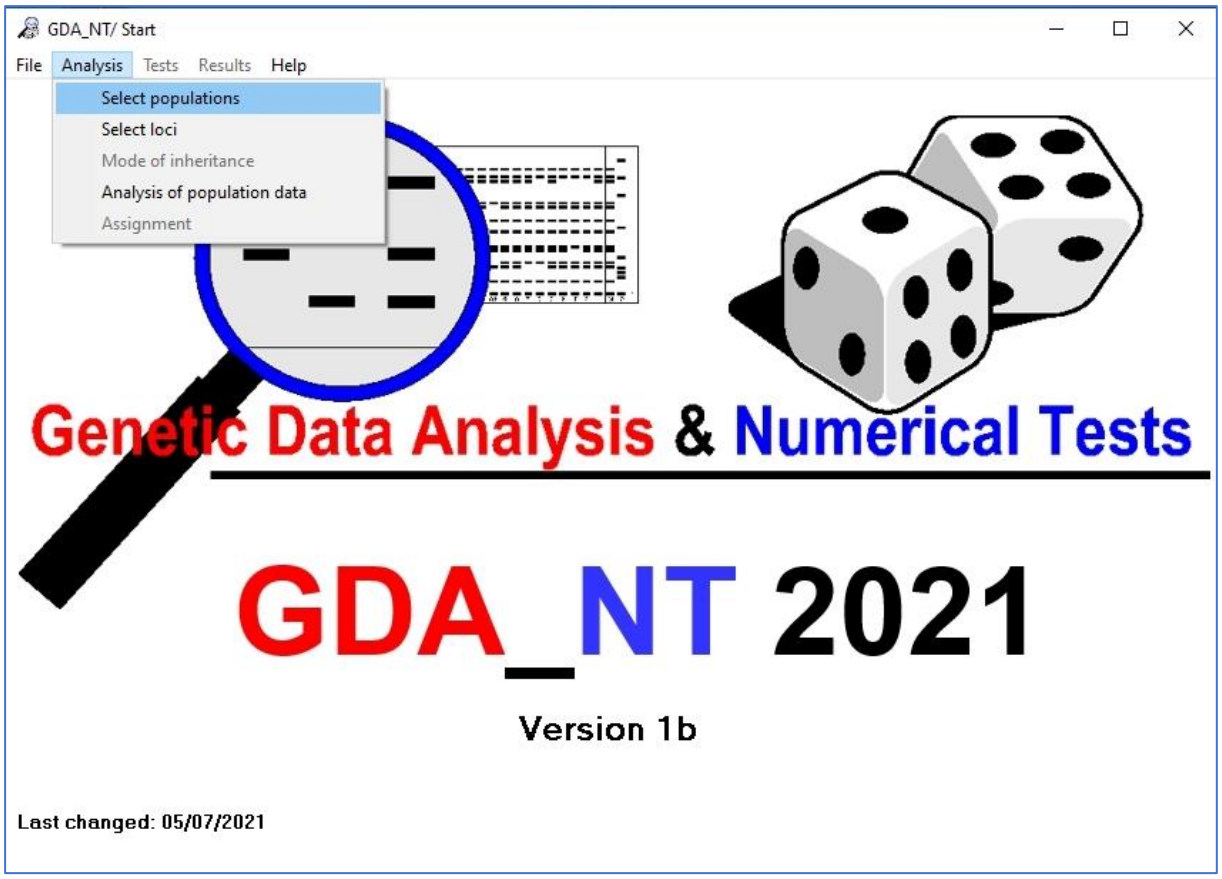


Figure 7: The menu “Analysis”

6.1 Selection of populations

You should select the population in the list and then click on the arrow in order to select or deselect a population (figure 7). All further calculations will be done only with the selected populations. In addition, there are the options to “include all populations” or to “exclude all populations”.

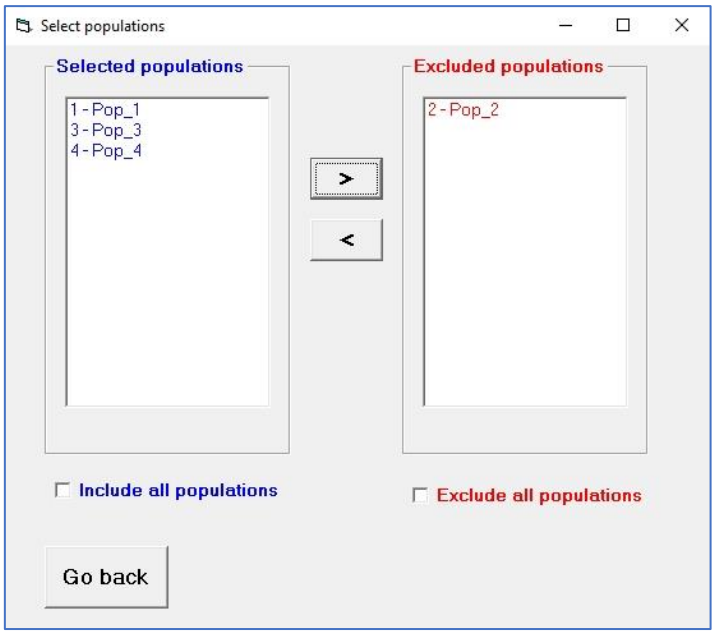


Figure 8: Window to select populations

6.2 Selection of loci

The window “Selection of loci” is similar to the window “Selection of populations”. Here you have also the possibility to select or deselect a particular range of loci.

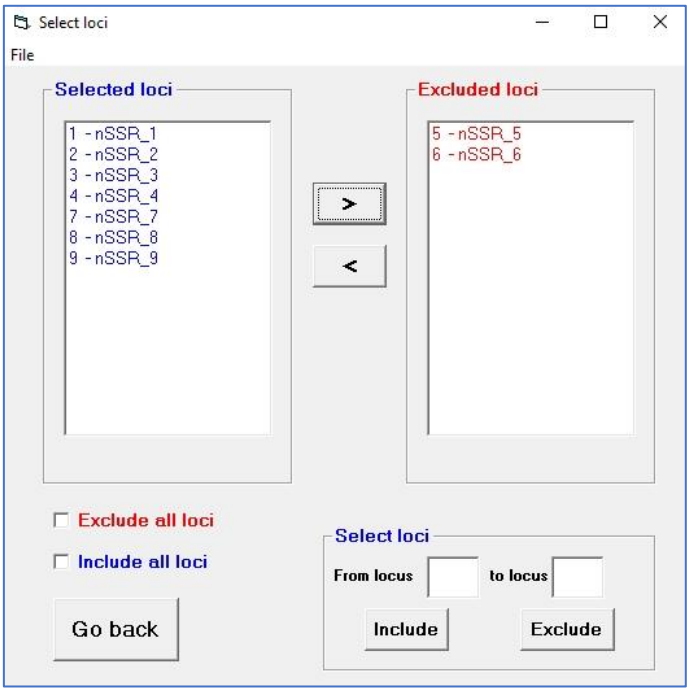


Figure 9: Window to select loci

6.3 Mode of inheritance

The window “Mode of inheritance” is of relevance for the calculation of allele frequencies and frequencies of haplotypes in frame of the assignment algorithms. Thus, the selection of this window will only be enabled if you have opened input files for the assignment. By selecting loci in the lists and clicking on the arrows you can change the mode of inheritance for the loci (figure 10).

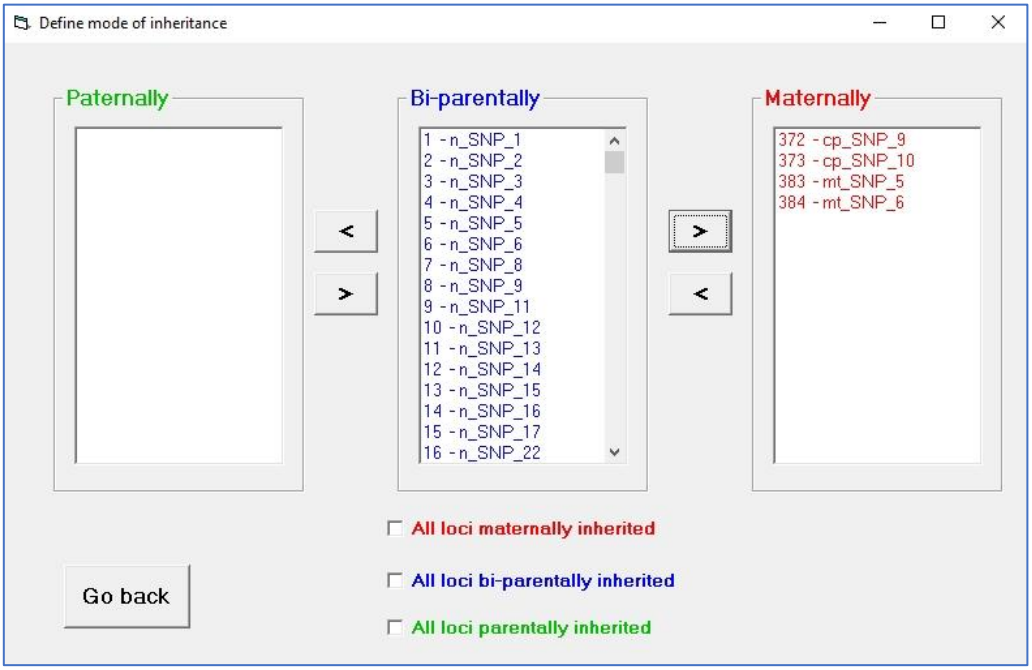


Figure 10: Window to select the mode of inheritance for the loci

6.4 Analysis of population data

The options and measures to be selected are organised in six frames (figure 11):

- Geographic co-ordinates
- Genetic composition
- Genetic differentiation
- Genetic variation
- Genetic distance and kinship among individuals
- Quality check gene loci

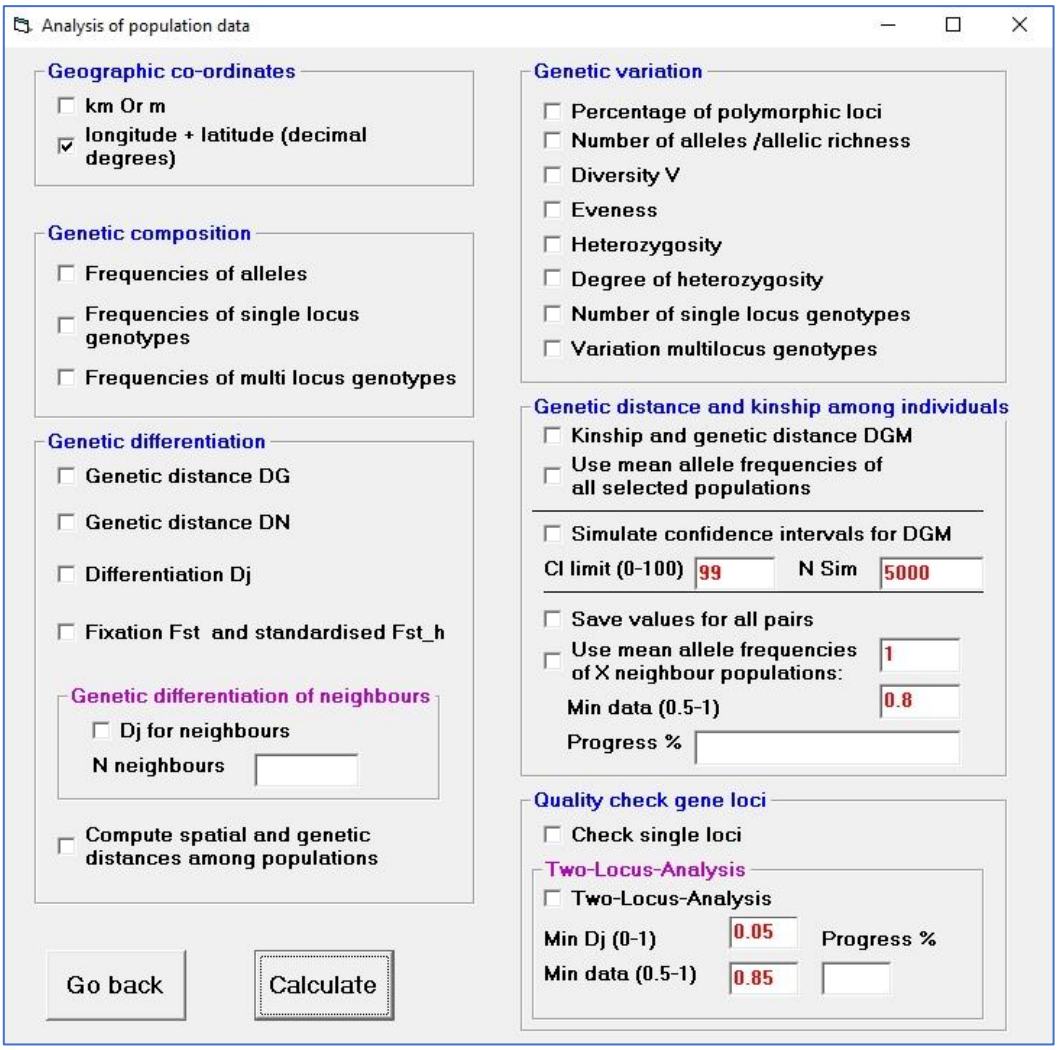


Figure 11: Window “Analysis of population data”

6.4.1 Geographic co-ordinates

In the input file the header of each population includes data on the X- and Y-position of the population. The user can define if these values are representing co-ordinates in km or m or if these values represent values of longitude and latitude in decimal degrees. The spatial distance among points is calculated as Euclidian distance for X- and Y-positions given in km or m. The spatial distance *Dis* between two geographic co-ordinates [(lon1, lat1), (lon2, lat2)] is computed in km using the formula for the distance calculation on the sphere surface:

$$Dis = 6378.388 \times \cos^{-1}(\sin lat1 \times \sin lat2 + \cos lat1 \times \cos lat2 \times \cos(lon2 - lon1)) \qquad (1)$$

6.4.2 Genetic composition

The user can select the calculation of relative frequencies of alleles, single locus genotypes and multi locus genotypes. The genotypes are computed as unordered genotypes and for the calculation of multi locus genotypes all selected loci are included. Thus, using the window "Selection of loci" you can select other combinations loci for the computation of multilocus genotype frequencies.

6.4.3 Genetic differentiation

The user can compute the two genetic distances D_G (Gregorius, 1974) or D_N (Nei, 1972) among allele frequencies of two populations:

$$D_G(i,j) = \frac{1}{2} \cdot \sum_{k=1}^{n} \left| p_{ik} - p_{jk} \right| \tag{2}$$

$$D_N(i,j) = -\ln\left(\frac{\sum_{k=1}^{n}(p_{ik} \cdot p_{jk})}{\sqrt{\sum_{k=1}^{n} p_{ik}^2 \cdot \sum_{k=1}^{n} p_{jk}^2}} \right) \tag{3}$$

where *i* and *j* represent two populations, *n* is the number of alleles, p_{ik} is the relative frequency of the *k*-th allele.

The genetic differentiation D_j (Gregorius and Roberds, 1986) is calculated as:

$$D_j = \frac{1}{2} \cdot \sum_{i=1}^{n_k} \left| p_i^{(j)} - \bar{p}_i^{(j)} \right| \tag{4}$$

where *j* is the population, p_i is the frequency of allele *i* at locus *k* and $\bar{p}_i$ is the frequency at allele *i* of all other populations (complement). For the option "D_j for neighbours" the allele frequencies $\bar{p}_i$ are computed for the **x** geographically closest neighbour populations.

The population fixation index F_{ST} (Wright, 1965) for a given locus is measured as:

$$F_{ST} = 1 - \frac{H_S}{H_T} \tag{5}$$

Here H_S is the average of the expected heterozygosities at the locus of all populations and H_T is the total expected heterozygosity calculated using the allele frequencies at all populations.

F_{ST} is not independent of the level of genetic diversity. For highly variable gene markers such as nuclear microsatellites the level of population fixation F_{ST} is per se small. The $F_{ST\,H}$ (Hedrick, 2005) considers the influence of the level of genetic variation on F_{ST} by performing a standardisation:

$$F_{ST\,H} = \frac{F_{ST}}{F_{ST\,max}} \tag{6}$$

$$F_{ST\,max} = \frac{(N_p - 1) \times (1 - H_S)}{(N_p - 1 + H_S)} \tag{7}$$

Here $F_{ST\,max}$ is the maximal possible F_{ST} and N_p is the number of populations.

The option “compute spatial and genetic distances among populations” produces a matrix of genetic distances D_G and spatial distances.

6.4.4 Genetic variation

The “percentage of polymorphic loci” counts each locus that is not fixed to a single allele and calculates the percentage among all loci. Number of alleles A_i counts the number of different alleles at locus i, Ar_i allelic richness corrects the number of alleles with rarefaction using the smallest sample size (El Mousadik and Petit, 1996):

$$Ar_i = \sum_j \left[\frac{1 - \binom{n}{N - N_j}}{\binom{n}{N}} \right] \tag{8}$$

Here j are the different alleles at locus i. N is the sample size in the given population for locus i; N_j is the absolute frequency of allele j and n is the smallest sample size at that locus among all populations.

The diversity v_i at locus i is according to Gregorius (1987):

$$1 \leq v_i = \frac{1}{\sum p_j^2} \leq A_i \tag{9}$$

here p_j is the relative frequency of allele j. The evenness E_i (Gregorius, 1990) measures the similarity of the observed distribution of allele frequencies to an equal distribution of alleles:

$$E_i = 1 - \frac{1}{2} \sum_{j=1}^{n} \left| p_j - \bar{p}_j \right| \tag{10}$$

Here $\bar{p}_j$ is the frequency of allele *j* for an equal distribution of all observed alleles *A* at the locus *i*. Please note that GA_NT 2021 uses the observed number of alleles and not all possible numbers of alleles as indicated in Gregorius (1990).

The option “Heterozygosity” enables for each locus the calculation of the observed heterozygosity H_o, the expected heterozygosity under Hardy-Weinberg assumptions H_e and the fixation index F_{IS}:

$$F_{IS} = 1 - \frac{H_o}{H_e} \tag{11}$$

The “degree of heterozygosity” measures the relative frequencies of individuals with different levels of heterozygosity, that is the number of loci at which an individual is heterozygote. The “number of single locus genotypes” counts the number of unordered genotypes at each locus. The “variation of multilocus genotypes” measures the absolute number of different multilocus genotypes and their effective numbers analogous to formula 9.

6.4.5 Genetic distance and kinship among individuals

This frame includes different options to analyse the genetic data of pairs of individuals within each population. The frequency of an allele of a diploid individual can have the values 0 (absent), 0.5 (heterozygote with one copy of the allele) and 1 (homozygote with two copies of that allele). The program computes the kinship coefficient *Kin* according to Loiselle *et al.* (1995):

$$Kin_{ij} = \frac{\sum(p_i - \bar{p}) \times (p_j - \bar{p})}{k \times \bar{p} \times (1 - \bar{p})} + \frac{2}{(8k+1)^{0.5} - 1} \tag{12}$$

$$k = \frac{n \times (n-1)}{2} \tag{13}$$

Here *i* and *j* are the two individuals, p_i is the frequency of the allele *p* of individual *i*; $\bar{p}$ is the frequency of allel *p* in the population; *n* is the number of individuals in the population, and *k* is the number of possible pairs of individuals. The values of the kinship coefficient get largely influenced by the precision of the measured allele frequencies in the population $\bar{p}$. Because of that the user has different option how to calculate the allele frequencies in the population:

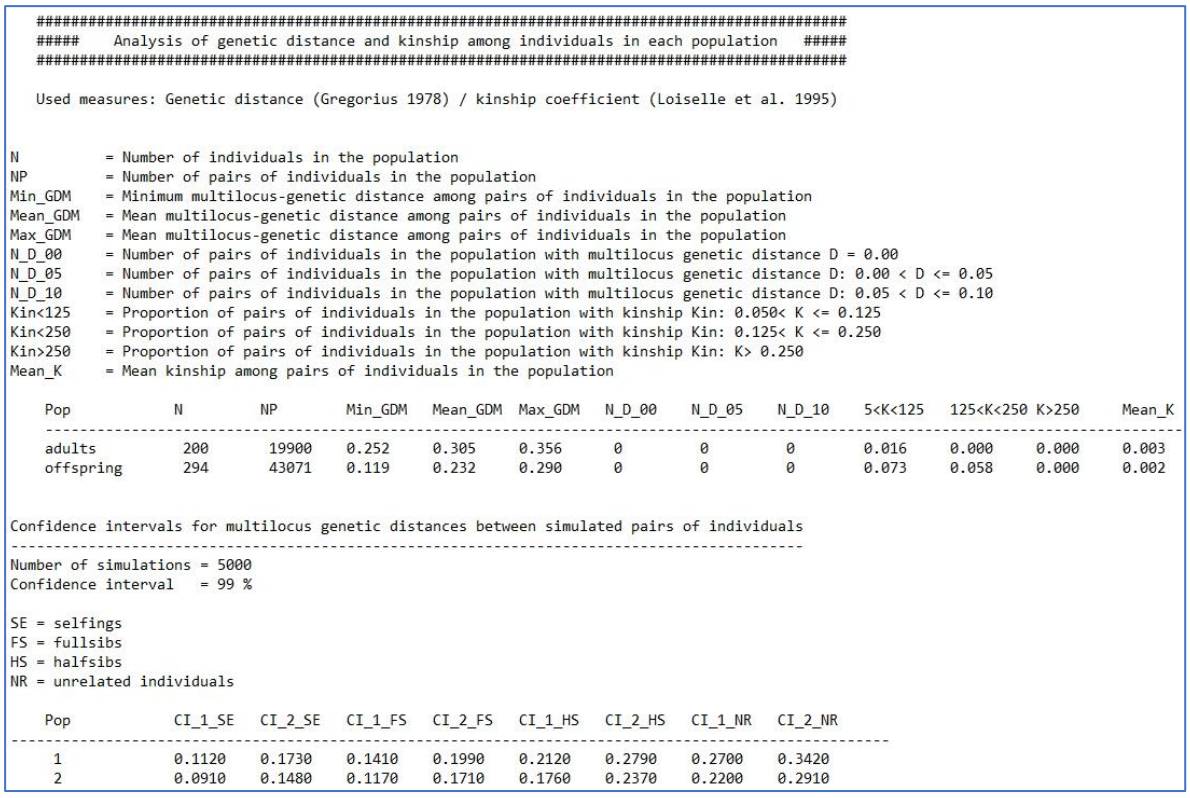
- the standard setting is that $\bar{p}$ gets calculated for each population, but if the sample size is small the values are unprecise
- the user can calculate $\bar{p}$ using all individuals in all selected populations (“Use mean allele frequencies of all selected populations”)

- or you use the individuals of the given population plus the individuals of **X** spatially next neighboured populations (“Use mean allele frequencies of X neighbour populations”). For this the user can specify the number of neighbour population for the calculation. This option makes sense if you have a data set with many populations (>10) and a large-scale genetic structure.

The expected value of the kinship coefficient *Kin* of non-related individuals is 0, for half-sibs it is 0.125 and for full-sibs it is 0.25.

In addition to the kinship coefficient the program computes for each pair of individuals, the multilocus genetic distance *DGM*. This genetic distance is calculated in analogy to formula 2 but with allele frequencies of individuals. The program also simulates confidence intervals of DGM for individuals that are a) the result of selfing (SE), b) half sibs (HS), c) full sibs (FS) and d) unrelated individuals (NR). For data sets with large numbers of gene loci these confidence intervals have no overlapping. In that case the classification of each pair of individuals in the respective confidence interval can be used to estimate the proportion of selfings, half sibs, full sibs and unrelated individuals in the populations.

The calculation of the kinship coefficient Kin and the multilocus genetic distance might be unprecise for individuals with a lot of missing data. Here the user can set a minimum proportion of complete data (“Min data (0.5-1)”. As a standard setting, parameters that describe the distribution of *Kin* and *DGM* are added to the output file (figure 12).



```
   ##############################################################################################
   #####    Analysis of genetic distance and kinship among individuals in each population   #####
   ##############################################################################################

   Used measures: Genetic distance (Gregorius 1978) / kinship coefficient (Loiselle et al. 1995)


N          = Number of individuals in the population
NP         = Number of pairs of individuals in the population
Min_GDM    = Minimum multilocus-genetic distance among pairs of individuals in the population
Mean_GDM   = Mean multilocus-genetic distance among pairs of individuals in the population
Max_GDM    = Mean multilocus-genetic distance among pairs of individuals in the population
N_D_00     = Number of pairs of individuals in the population with multilocus genetic distance D = 0.00
N_D_05     = Number of pairs of individuals in the population with multilocus genetic distance D: 0.00 < D <= 0.05
N_D_10     = Number of pairs of individuals in the population with multilocus genetic distance D: 0.05 < D <= 0.10
Kin<125    = Proportion of pairs of individuals in the population with kinship Kin: 0.050< K <= 0.125
Kin<250    = Proportion of pairs of individuals in the population with kinship Kin: 0.125< K <= 0.250
Kin>250    = Proportion of pairs of individuals in the population with kinship Kin: K> 0.250
Mean_K     = Mean kinship among pairs of individuals in the population

    Pop            N         NP        Min_GDM   Mean_GDM  Max_GDM   N_D_00    N_D_05    N_D_10    5<K<125   125<K<250 K>250     Mean_K
    ---------------------------------------------------------------------------------------------------------------------------------------
    adults          200       19900    0.252     0.305     0.356      0         0         0        0.016     0.000     0.000     0.003
    offspring       294       43071    0.119     0.232     0.290      0         0         0        0.073     0.058     0.000     0.002


Confidence intervals for multilocus genetic distances between simulated pairs of individuals
-----------------------------------------------------------------------------------------------
Number of simulations = 5000
Confidence interval   = 99 %

SE = selfings
FS = fullsibs
HS = halfsibs
NR = unrelated individuals

    Pop            CI_1_SE   CI_2_SE   CI_1_FS   CI_2_FS   CI_1_HS   CI_2_HS   CI_1_NR   CI_2_NR
-------------------------------------------------------------------------------------------------
     1             0.1120    0.1730    0.1410    0.1990    0.2120    0.2790    0.2700    0.3420
     2             0.0910    0.1480    0.1170    0.1710    0.1760    0.2370    0.2200    0.2910
```

Figure 12: Output for the analysis of “Genetic distance and kinship among individuals”

But the user can also store the results for each individual pair in a csv-file (“Save values for all pairs”, figure 13).

Pop	Ind_1	Ind_2	N_loci	DGM	CI	F_D0	F_D1	Kin
adults	1	2	500	0.304	NR	0.486	0.094	-0.0165
adults	1	3	500	0.31	NR	0.47	0.09	-0.0289
adults	1	4	500	0.31	NR	0.458	0.078	-0.0004
adults	1	5	500	0.314	NR	0.454	0.082	-0.0145
adults	1	6	500	0.309	NR	0.456	0.074	-0.0142
adults	1	7	500	0.314	NR	0.45	0.078	-0.0143
adults	1	8	500	0.292	NR	0.496	0.08	0.0252
adults	1	9	500	0.325	NR	0.432	0.082	-0.0442
adults	1	10	500	0.304	NR	0.472	0.08	-0.0092
adults	1	11	500	0.286	NR	0.512	0.084	0.0272
adults	1	12	500	0.3	NR	0.476	0.076	0.0205
adults	1	13	500	0.312	NR	0.466	0.09	-0.0041
adults	1	14	500	0.307	NR	0.476	0.09	-0.0029
adults	1	15	500	0.292	NR	0.494	0.078	0.0225
adults	1	16	500	0.309	NR	0.468	0.086	-0.0115

Figure 13: Output for the analysis of “Genetic distance and kinship among individuals” stored for each pair of individuals in a csv-file; “Pop” = name of the population, “Ind_1” and “Ind_2” numbers of the individuals that builds the pair, “N_loci” = number of gene loci that were used for the calculation, “DGM” = multilocus genetic distance among the two individuals, “CI” classification of the pair of individuals into classes of simulated confidence intervals => in the table all pairs are classified as “NR” = not related; “F_D0” = proportion of gene loci with a genetic distance of 0; “F_D1” = proportion of gene loci with a genetic distance of 1; “Kin” = kinship coefficient for the two individuals

6.4.6 Quality check gene loci

Often a quick evaluation on the quality of gene markers for population genetic studies is useful before you start a deeper data analysis or a broader genetic inventory. So, you can concentrate on the informative gene markers and exclude those that have no variation or other problematic characteristics. Also, the level of redundancy among different gene markers is of interest.

By clicking “Check single loci” the user selects the option to compute the following parameters for each locus (table 1). The definition of critical values for these measures depends on the planned application of the gene markers. Some measures such as the completeness of data and a minimum level of genetic diversity *v* is relevant for all applications, but other aspects

such as the correlation among spatial and genetic distances might be irrelevant, if either the spatial genetic structure is not a question to be addressed or if the data on the geographic position of the populations is unknown or very unprecise.

Abbreviation	Definition	Critical values	Meaning
data	average percentage of complete data	< 80%	too much missing data
Fis	average fixation index F_{IS}	< -0.3 < 0.3	Strong departure of heterozygosity from Hardy-Weinberg expectations
v	genetic diversity *V*	1	No allelic variation
D j	genetic differentiation D_j	< 0.1	Not much genetic differentiation
F_ST	population fixation index F_{ST}	< 0.05	Not much population fixation
r_DG_SD	Pearson's correlation coefficient among matrices of genetic and spatial distances of populations	< 0.1	Nearly no spatial genetic structure

Table 1: Overview on different measures that are calculated as quality check of single loci and an example for critical values and their meaning

The user calls for an analysis of two locus combinations by clicking on "Two-Locus-Analysis". But certain loci get excluded from this analysis:

- loci that are monomorphic ($v = 1$)
- loci that do not have a minimum level of genetic differentiation *Dj* ("Min Dj") defined by the user
- loci that have too much missing data ("Min data"), threshold value defined by the user

This procedure analysis a few measures for all combinations of two loci (table 2). The main objectives of this analysis are to identify redundant and problematic loci. That are loci that are showing extremely similar or equal results potentially due to very close linkage. Or these are loci that have genotyping errors. Further, the procedure looks for particular powerful combinations of loci that have a very strong spatial genetic pattern.

Abbreviation	Definition	Comment
Fis_A	mean fixation index F_{IS} of first locus	Redundant loci have very similar or identical F_{IS}-values
Fis_B	mean fixation index F_{IS} of second locus	
N_MG	number of unordered multilocus-genotypes for combinations of alleles at both loci	
V_MG	mean diversity *V* multilocus-genotypes, uses the relative frequencies of the two-locus genotypes in formula 9	
V_R	Mean of the quotient of observed diversity V and expected diversity V (random combination of alleles)	Values < 0 give indication of linkage among loci, departure from Hardy-Weinberg-expectations or genotyping errors
DG_M	mean genetic distance *DG* among populations based on frequencies of multilocus-genotypes	Indication of the level of genetic differentiation
r_DGM_SD	Pearson's correlation among genetic (multi-locus) and spatial distances	Values > 0 indicate a spatial genetic structure
r_DG_GD	Pearson's correlation coefficient among matrices of genetic distances among populations for both loci	Redundant loci have very similar or identical values
DG_RND	mean genetic distance *DG* among observed and expected frequencies of multilocus-genotypes at the two loci	Linked loci and loci with null alleles have values close to 0.4 and bigger

Table 2: Overview on different measures that are calculated to check two-locus combinations

An example for a quality check of gene markers is demonstrated in a video:

https://www.thuenen.de/media/institute/fg/Software/Videos/Quality_Check_Loci_GDA_NT.mp4

6.5 Assignment

The user needs to open two input files in the start window in order to run a genetic assignment (figure 14). The format of these files is the same as for all input files of the program (see 5.1 “Input data”). The first file includes the data of the reference populations (“Open file with reference pops for assignment”). The second file has the data of the test groups that are the subject of the assignment (“Open file with test groups for assignment”). The test groups could also be different single individuals. In this case they should be formatted as different populations. First, open the reference populations and then open the test groups. The program requires that the number of gene loci, their names and their order is the same in both input files. You get a message if the reading of the input files was successful. The file with the reference populations and the file with the test groups can be the same if you just want to do a self-assignment test. The option “Assignment” in the menu “Analysis” is enabled only if both files were successfully loaded into the program.

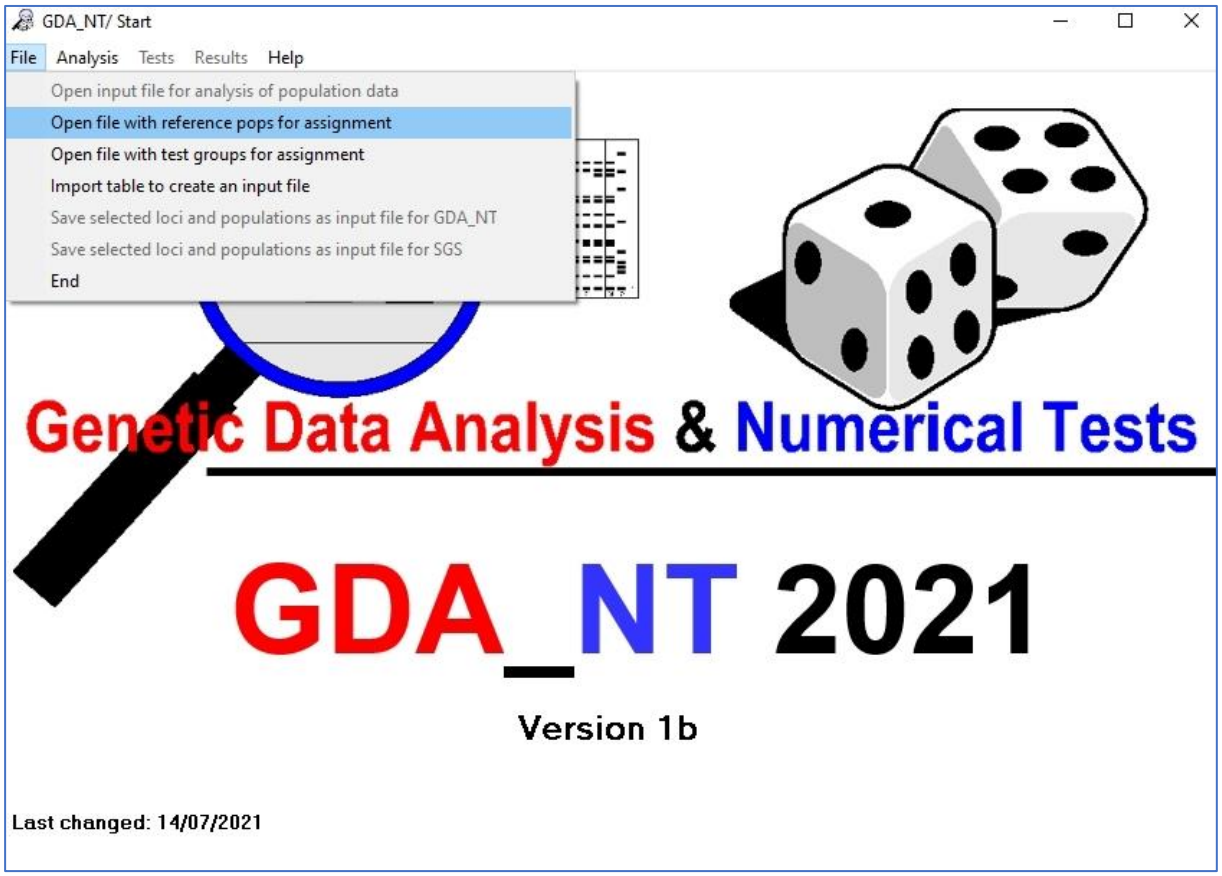


Figure 14: Start window, menu “File” with the option to open a reference data set and a test data set for the genetic assignment.

The application of the genetic assignment with GDA_NT 2021 gets demonstrated in the following video:
https://www.thuenen.de/media/institute/fg/Software/Videos/Genetic_Assignment_GDA_NT.mp4

By clicking on “Assignment” the window “Assignment/ exclusion of groups” opens (figure 15). The window consists of three frames:

1. “Method for assignment” => here the user can select which algorithm / method is used for the assignment
2. “Calculation of exclusion probabilities” => here are the setting for the simulation of genotypes that are used for the calculation of exclusion probabilities
3. “Self-assignment” => parameters that specify the way of the self-assignment

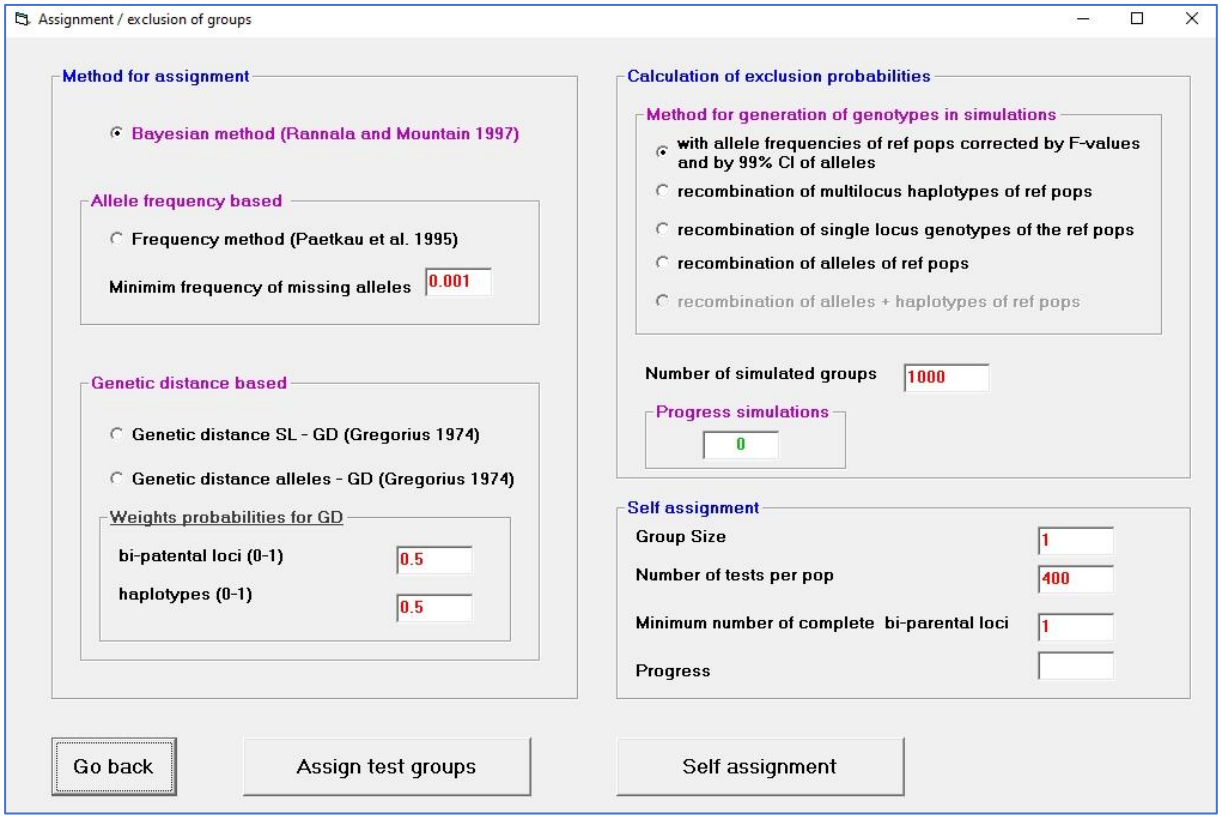


Figure 15: The window “Assignment /exclusion of groups”

6.5.1 Methods of assignment

The standard setting is the use of the Bayesian method following Rannala and Mountain (1997). The approach concerns the derivation of probability density of population allele frequencies from the frequencies in samples (Cornuet *et al.*, 1999). The approach assumes

an equal probability density of the allele frequencies of each locus in each reference population. The marginal probability of observing an individual with genotype $A_k A_{\acute{k}}$ at the locus *j* in population *i* was equal to (formula 9 of Rannala and Mountain (1997)):

$$\frac{(n_{ijk} + \frac{1}{K_j} + 1)(n_{ijk} + \frac{1}{K_j})}{(n_{ij} + 2)(n_{ij} + 1)} \; if \; k = \acute{k} \tag{14}$$

$$\frac{2(n_{ijk} + \frac{1}{K_j})(n_{ij\acute{k}} + \frac{1}{K_j})}{(n_{ij} + 2)(n_{ij} + 1)} \; if \; k = \acute{k} \tag{15}$$

Here n_{ijk} is the number of allele *k* sampled at gene locus *j* in population *i*, n_{ij} is the sample size of alleles at locus *j* in population *i* and K_j is the total number of alleles observed in all reference populations.

Then the assignment is done in three steps:

1. Computing the allele frequencies in all reference populations
2. Computing the likelihoods of the individual genotypes of the test groups occurring in each reference population (the likelihoods of all individuals in a test group gets summed up)
3. Assigning the test group to the reference population with the highest likelihood

The Bayesian approach assumes random mating and no linkage disequilibrium among different loci.

The allele frequency based method assigns an individual or a group of individuals to the reference population in which the genotypes of the test group members are most likely to occur (Paetkau *et al.*, 1995). The frequency of allele *k* at locus *j* in reference population *i* is p_{ijk}. Under the assumption of Hardy-Weinberg- equilibrium, the likelihood of a genotype $A_k A_{\acute{k}}$ to be present in the *i*th reference population at the *j*th locus is p_{ijk}^2 if $k = \acute{k}$ and if $k \neq \acute{k}$ then it is 2 p_{ijk} $p_{ij\acute{k}}$. And further assuming independence of the loci the likelihood of a multilocus genotypes gets computed as the product of the likelihood at each locus. For the group of individuals, the likelihood is the product of the likelihoods of the individuals. To avoid the problem of excluding a reference population as a candidate only because a rare allele is absent (which could be only a sampling effect) the user defines a “Minimum frequency of missing alleles”. The

standard setting for that is 0.001. Then the assignment follows the same three steps as described above for the Bayesian approach.

As a third option the genetic distance based approach is offered using the genetic distance DG (formula 2) and either the frequencies of single genotypes or the frequencies of alleles in the reference populations and test groups. The distance then is transformed into similarities: $S = 1 - DG$. The test group gets assigned to the reference population with the highest similarity.

The advantage of the distance based approach is that it does not require Hardy-Weinberg equilibrium or independence of the loci (Cornuet *et al.*, 1999).

In all three methods maternally and paternally inherited haplotypes are treated as one additional locus in the formulas. And for the distance based approach the weight for bi-parental inherited loci and the haplotypes can be modified (standard setting is 0.5 for both).

6.5.2 Calculation of exclusion probabilities

All three methods of assignment always find a reference population as the most likely for each test group. But it is also possible that the true population of origin is not among the alternative reference populations. Here we need a measure of confidence that the test group truly belongs to a given reference population. One way to get an estimation on that is to compare the likelihood values of the test group with a distribution of likelihood values of individuals or groups of individuals drawn directly from the reference population. The proportion of values with lower likelihood (Bayesian and allele frequency approach) or higher genetic distances compared to the value of the test group can be interpreted as a measure that the test group truly belongs to the reference population (Cornuet *et al.*, 1999). For example, if the genetic distance of the test group to the reference group is 0.18 and 98% of all genetic distances of groups sampled from the reference population to the reference population are smaller, then the probability that the test group truly belongs to the reference population is 2%.

The program offers different options how to simulate multilocus genotypes drawn from each reference population (figure 15):

1. “with allele frequencies of ref pops corrected by F-values and by 99% CI of alleles” => Here alleles of the simulated multilocus genotypes are selected proportional to their frequencies in the reference population. The program considers the fact that these allele frequencies are erroneous because of sampling effect. Thus, for each genotype

the allele frequencies are selected randomly within their 99% confidence interval based on the sample size of the reference population. Further the program corrects the proportion of homozygotes and heterozygotes according to the F_{IS} values of each locus in the reference population.

2. “recombination of multilocus haplotypes of ref pops” => Here multilocus genotypes are generated by a) randomly drawing multilocus haplotypes from the existing multilocus genotypes of the samples in the reference population, and b) rearranging them randomly with haplotypes of other individuals of the reference population. This is a re-sampling without replacement. This method includes also the missing values and keeps a potential linkage among loci.
3. “recombination of single locus genotypes of the ref pops” => This is simply a creation of new multilocus genotypes by randomly selecting single locus genotypes from individuals of the reference population. Also, this is a sampling without replacement and the departure from Hardy-Weinberg heterozygosity gets considered in the sampling.
4. “recombination of alleles of ref pops” => This method just takes randomly alleles from individuals of the reference population as a unit of resampling (again without replacement).
5. “recombination of alleles + haplotypes of ref pops” => This option gets only enabled if you have specified bi-parentally inherited loci and uni-parentally inherited loci. The alleles of the simulated multilocus genotypes get sampled as in 4 but the multilocus haplotypes are sampled from the individuals of the reference populations and no recombination among loci is applied for these uni-parentally inherited loci.

6.5.3 Self-assignment

As an indicator of the performance of the assignment method but also of the performance reference data, the user can compute the proportion of correctly assigned individuals in self-assignments tests (Cornuet *et al.*, 1999). Here the individuals of the reference population were self-classified to the sampled groups using the leave-one-out approach (Efron, 1983). The user can specify three parameters for these self-assignment tests:

1. “Group size” => This is the number of individuals that were taken as unit of the self-assignment tests. The standard setting is one individual. In this case each individual of the reference population is subject of the self-assignment. But you can also select more than one individual.

2. “Number of tests per pop” => If the group size is bigger than 1, you need to specify how many random samples of groups should be done in each reference populations. Thus not all combinations are tested but a high simulated sample of them.
3. “Minimum number of complete bi-parental loci” => With this parameter you define the minimum requirement for data completeness for individuals to be included into self-assignment tests.

7 Menu “Tests”

The user has the possibility to do numerical tests to evaluate the statistical significance of different measures selected in the window “Analysis of population data” (figure 16). These are permutation tests using Monte-Carlo simulations comparing the observed value of the measure with the distribution of this measure in the permutations (Manly, 1997).

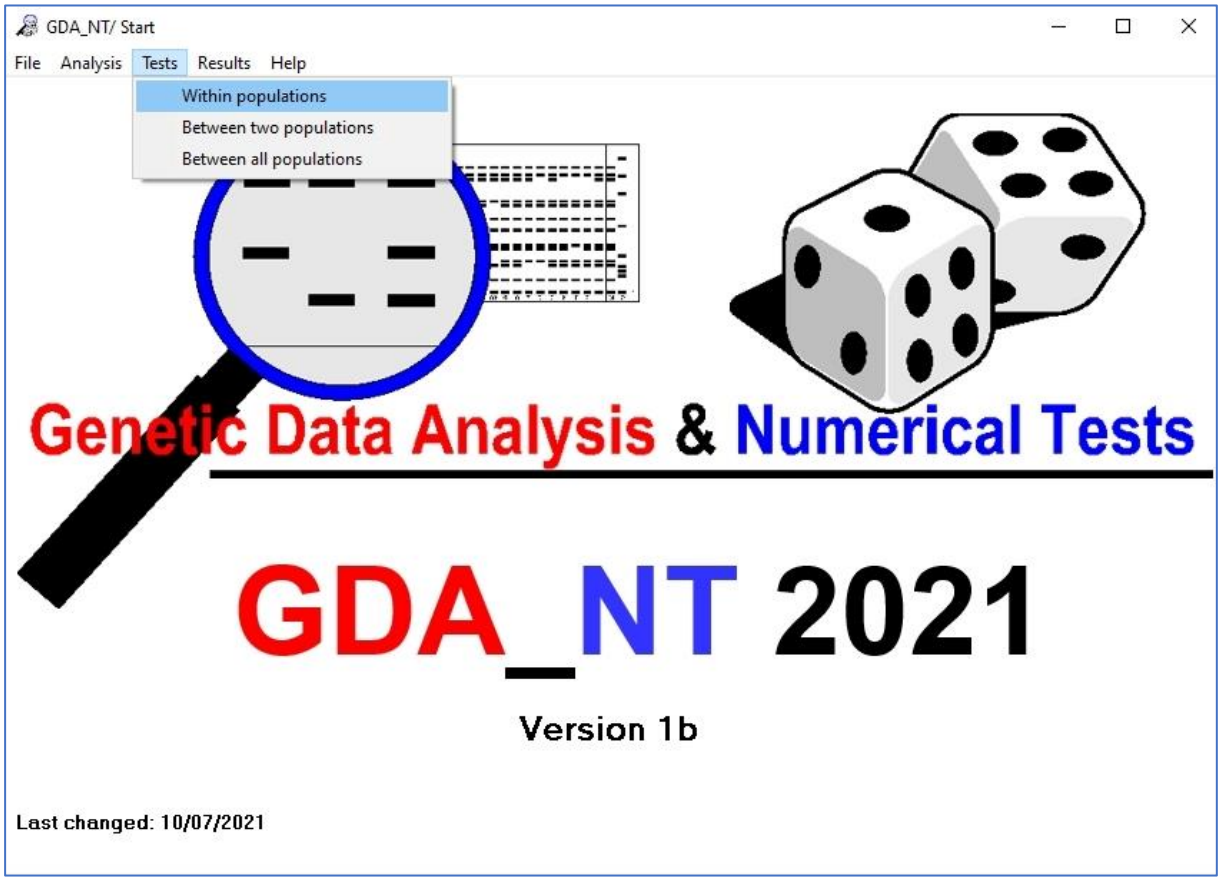


Figure 16: Menu “Test”

The user can run tests “Within populations” if measures of heterozygosity has been computed. The option for tests “Between two populations” is enabled after calculation of genetic distances

and the option for tests “Between all populations” is enabled after computation of measures *Dj* and F_{ST}.

7.1 Tests “within populations”

By clicking on this element of the test menu the window for tests within populations opens (figure 17). The permutation tests checks whether the observed fixation indices F_{IS} are significant different to the value expected for random mating. You can select a particular population to be tested by clicking on the population name in the list. The standard setting is that all populations will be tested. Further you can modify the number of permutations. Here the standard setting is 1000. In each of those permutations the alleles of the selected population at a given locus are randomly recombined (sampling without replacement). Then the F_{IS}-value of the permutated genotypes gets calculated and compared with the observed value. The probability of rejecting the null hypothesis of Hardy-Weinberg heterozygosity is equal the cases when the observed negative F_{IS} values is smaller compared to the value of the permutations, or when the observed positive F_{IS} value is bigger than the value of the permutations. Thus, this is a directional test: significant positive means than significant excess of homozygotes and significant negative F_{IS} value means excess of heterozygotes.

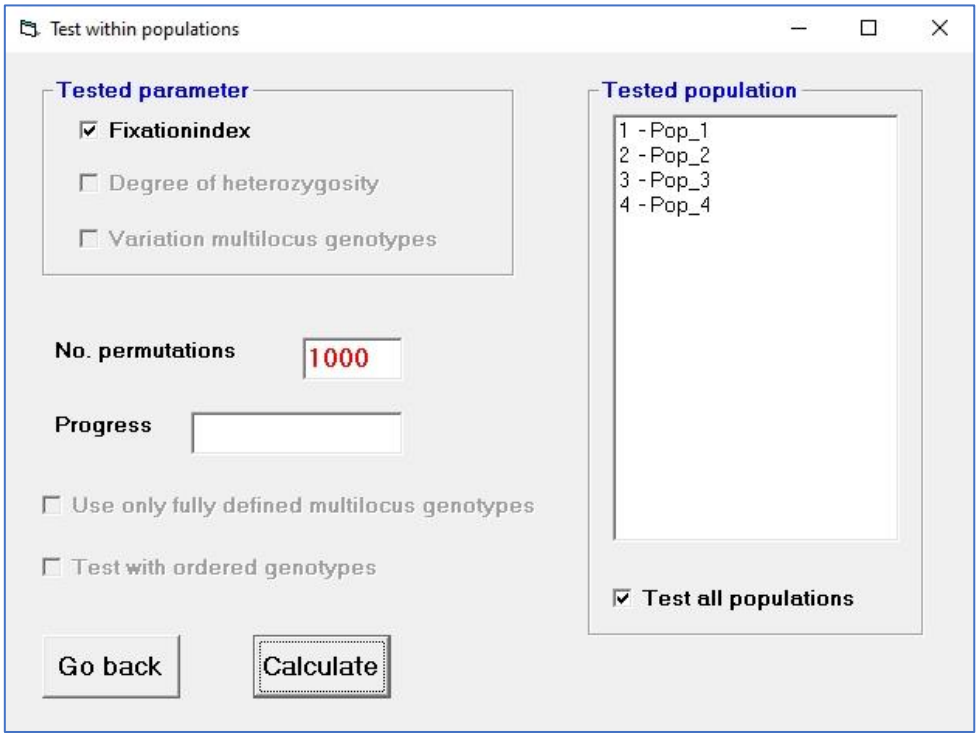


Figure 17: Window tests ”Within populations”

In the example (figure 18) the gene loci nSSR_5, nSSR_7 and the mean value over all loci have statistically significant negative F_{IS} values (Prob < 0.05).

```
#######################################################
####        Results of Numerical Tests           ####
#######################################################


 Tested Population     = Pop_1

 Number of Permutation = 1000

   *********************
   **  Fixationindex  **
   *********************

   Prob = pobability that Fis is in the range for Hardy-Weinberg-expectation

   Locus          F         Prob
   -----------------------------------
   nSSR_1        -0.016    0.329
   nSSR_2        -0.021    0.413
   nSSR_3        -0.014    0.406
   nSSR_4        -0.070    0.063
   nSSR_5        -0.162    0.010
   nSSR_6        -0.142    0.399
   nSSR_7        -0.151    0.018
   nSSR_8         0.054    0.230
   nSSR_9         0.060    0.359

   Mean          -0.051    0.027
```

Figure 18: Example with results of a permutation test of the fixation indices F_{IS}

7.2 Tests between two populations

The user has the possibility to test the statistical significance of genetic distances (*DG*, *DN*). These permutation tests are done by randomly shifting alleles, single locus genotypes or multilocus genotypes between a selected pair of populations. The units of resampling can be defined by the user (figure 19). In each permutation a new combination of the genetic information of the two population gets generated. Then the allele frequencies in the permutated two populations are calculated and based on these are frequencies the genetic distances are computed. The null hypothesis is that the samples of the two populations are taken from the same population (= no genetic difference). The frequency of genetic distances in the permutations that are bigger than the observed genetic distance is the probability of the null hypothesis. If this percentage is small (<0.05) we need to reject the null hypothesis and should consider the samples taken from two genetically different populations.

Besides the “Unit of resampling” the user can specify the “Number of permutations” (standard setting = 1000) and can select the pairs of population. As the standard setting all pairs of populations get tested and a matrix of probabilities is added to the result file (figure 20).

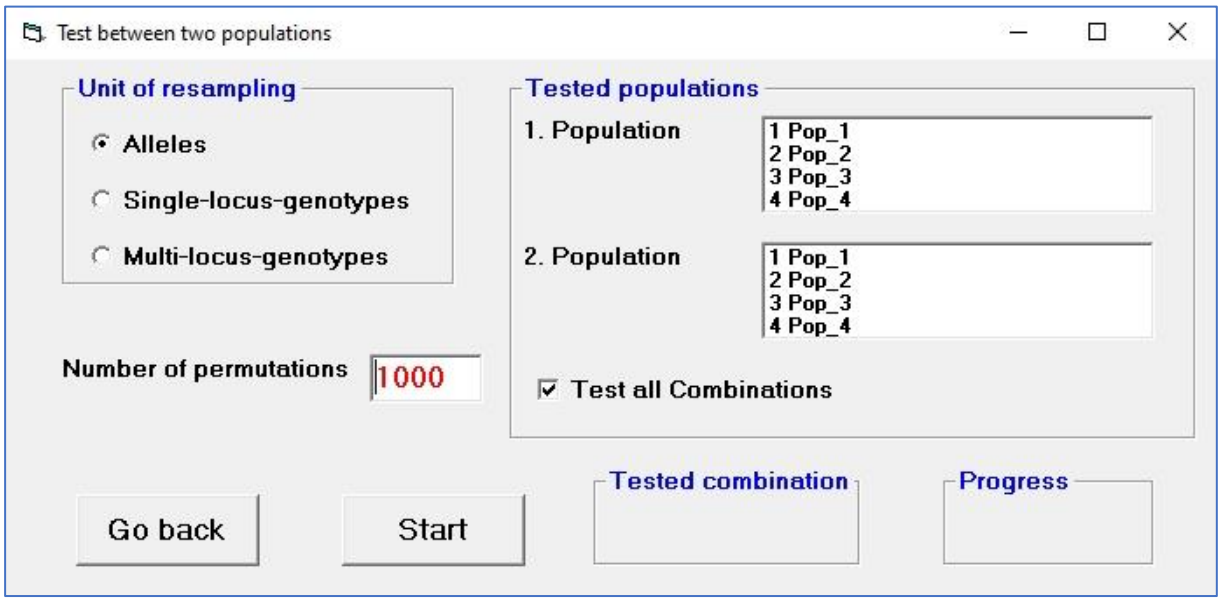


Figure 19: Window tests ”between two populations”

```
#########################################################
####                 SUMMARY                         ####
####        Results of Numerical Tests               ####
#########################################################

 Number of Permutation = 1000
 Unit of Permutation   = Alleles

  MATRIX of Pobabilities for Genetic Distance DG
  ----------------------------------------------

  Values = probabilities for collection in same population


  nSSR_1 (Alleles)


    1      2      3      4
   --------------------------------
1          0.318  0.456  0.117
2                 0.097  0.010
3                        0.111

  ###################################
```

Figure 20: Example for the results of permutation tests between two populations using the demo file; population 2 and 4 are significant different at the locus nSSR_1

7.3 Test between all populations

The user can do permutation tests to check if the measures of differentiation (*Dj*) and population fixation (F_{ST}, F_{ST_H}) are statistically significant. That is to say if the null hypothesis "all populations are genetically identical" should be rejected. Here the complete multilocus genotypes of an individual is the unit of permutation. Thus, each permutation is a new arrangement of individuals in the populations. The user can select the number of permutations (figure 21). The standard setting is 1000. After each permutation the measures of differentiation and population fixation are calculated and compared to the observed values. The result of each permutation is added to the result file and at the end the frequencies of permutation results (P) > and < compared to the observed values (O) are added as two rows (figure 22).

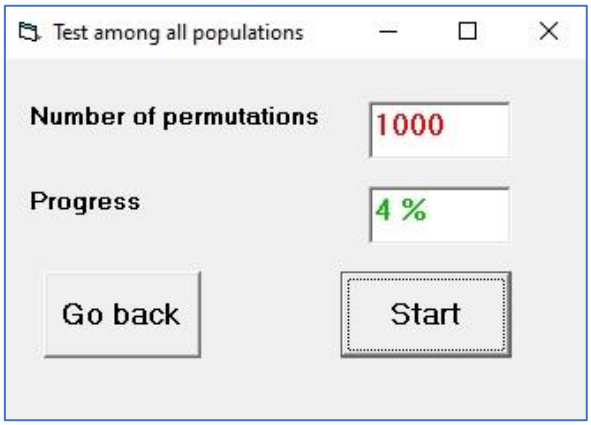


Figure 21: Window tests "between all populations"

Per	Dj(nSSR_1)	Dj(nSSR_2)	Dj(nSSR_3)	Dj(nSSR_4)	Dj(nSSR_5)	Dj(nSSR_6)	Dj(nSSR_7)	Dj(nSSR_8)	Dj(nSSR_9)	Dj_T
995	0.2691	0.1721	0.2161	0.3225	0.2315	0.2835	0.256	0.4281	0.3167	0.2773
996	0.2254	0.1449	0.2192	0.3244	0.2234	0.3502	0.1929	0.383	0.4371	0.2778
997	0.2701	0.1324	0.1421	0.3449	0.2571	0.3648	0.2525	0.3912	0.3643	0.2799
998	0.2531	0.1756	0.2094	0.3261	0.3002	0.3801	0.205	0.4365	0.3933	0.2977
999	0.2646	0.1541	0.1723	0.3275	0.304	0.3712	0.1916	0.3389	0.368	0.2769
1000	0.2416	0.0932	0.2185	0.2801	0.2112	0.414	0.2452	0.4398	0.4019	0.2828
P<O	1	0.996	1	1	0.999	0.991	1	0.994	0.82	1
P>O	0	0.004	0	0	0.001	0.009	0	0.006	0.18	0

Figure 22: Example for the results of a permutation test among all populations; here for the genetic differentiation of each single locus and the total differentiation over all loci (*Dj_T*).

8 Menu “Results”

The results of the analysis of population data and the genetic assignment are stored as text files. Using the option “Editor” in the result menu you can open the Windows text editor to visualise these text files. Other options are to “Save results of allele frequencies as a csv-file” and to “Save the allele frequency as input file for SGS” (figure 23). The last option is of interest if you have spatial explicit data and more than 20 populations studied. In that case you can use the program Spatial Genetic Structure (Degen *et al.*, 2001) to make a detailed analysis. This program is available at:

https://www.thuenen.de/en/fg/software/sgs/

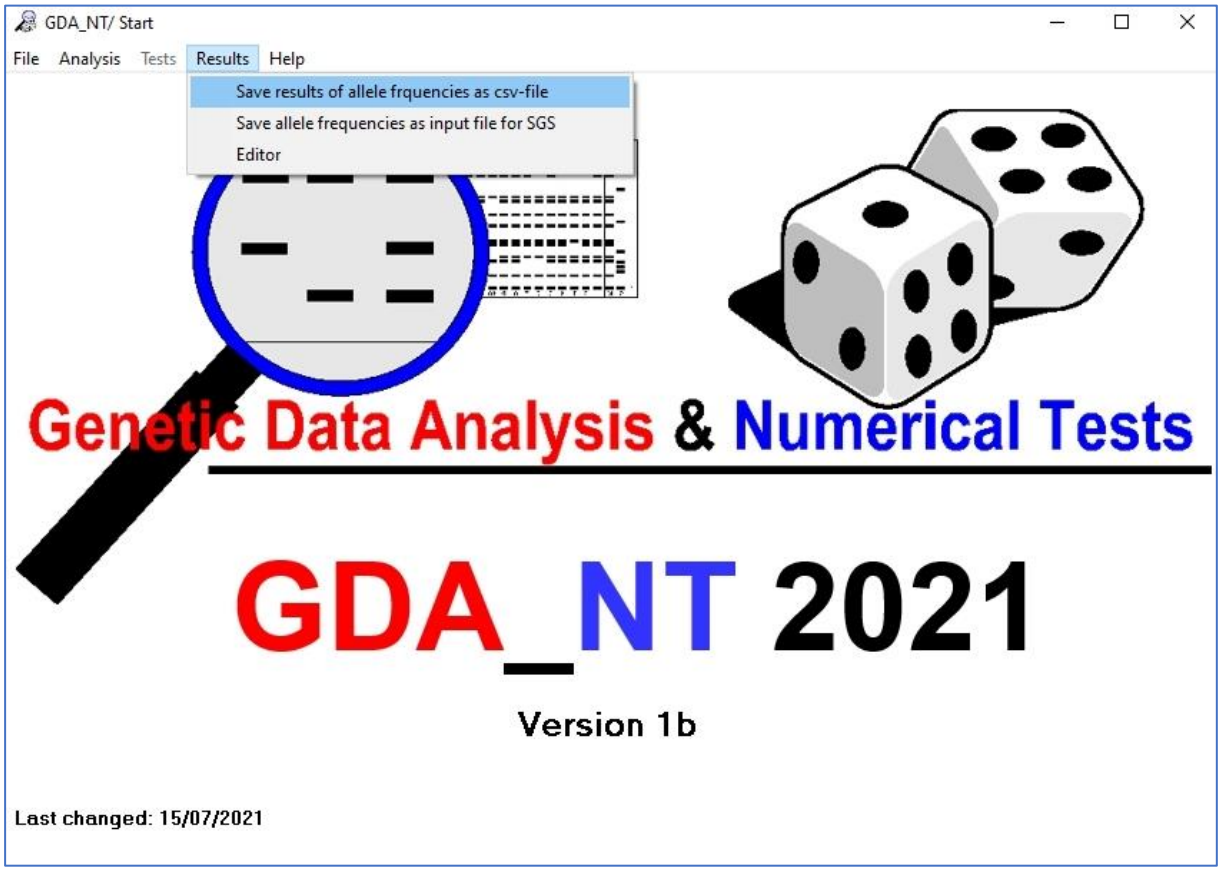


Figure 23: Options in the menu “Results”

9 Validation of computed results

For most calculated measures the correctness of the results has been checked by comparing with results of other software (GenePop, PopGen, GeneClass, GSED, PAST) using the same input data. For measure unique to GDANT simulated data have been used to check the reliability of the results.

10 Acknowledgement

I would like to thank Céline Blanc-Jolivet and Malte Mader for critical testing of the program and for helpful suggestions of its improvement.

11 Literature

Cornuet, J.-M., Piry, S., Luikart, G., Estoup, A., Solignac, M., 1999. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics 153, 1989-2000.

Degen, B., Blanc-Jolivet, C., Bakhtina, S., Ianbaev, R., Yanbaev, Y., Mader, M., Nurnberg, S., Schroder, H., 2021. Applying targeted genotyping by sequencing with a new set of nuclear and plastid SNP and indel loci for Quercus robur and Quercus petraea. Conserv. Genet. Resour., 3.

Degen, B., Petit, R.J., Kremer, A., 2001. SGS - Spatial genetic software: A computer program for analysis of spatial genetic and phenotypic structures of individuals and populations. Journal of Heredity 92, 447-449.

Efron, B., 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. Journal of the American Statistical Association 78, 316-331.

El Mousadik, A., Petit, R., 1996. High level of genetic differentiation for allelic richness among populations of the argan tree [Argania spinosa (L.) Skeels] endemic to Morocco. Theor. Appl. Genet. 92, 832-839.

Gregorius, H.-R., 1990. A diversity-independent measure of evenness. The American Naturalist 136, 701-711.

Gregorius, H.-R., Roberds, J., 1986. Measurement of genetical differentiation among subpopulations. Theor. Appl. Genet. 71, 826-834.

Gregorius, H., 1974. Genetic distance between populations. The concept of measurement of genetic distance. Silvae Genetica 23, 22-27.

Gregorius, H.R., 1984. A unique genetic distance. Biom. J. 26, 13-18.

Gregorius, H.R., 1987. The relationship between the concepts of genetic diversity and differentiation. Theor. Appl. Genet. 74, 397-401.

Hedrick, P.W., 2005. A standardized genetic differentiation measure. Evolution 59, 1633-1638.

Loiselle, B.A., Sork, V.L., Nason, J., Graham, C., 1995. Spatial genetic-structure of a tropical understory shrub, Psychotria officialis (Rubiaceae). Am. J. Bot. 82, 1420-1425.

Manly, B., 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology. Chapmann & Hall, London.

Meirmans, P.G., Hedrick, P.W., 2011. Assessing population structure: F-ST and related measures. Mol. Ecol. Resour. 11, 5-18.

Nei, M., 1972. Genetic distance between populations. The American Naturalist 106, 283-292.

Paetkau, D., Calvert, W., Stirling, I., Strobeck, C., 1995. Microsatellite analysis of population structure in Canadian polar bears. Molecular Ecology 4, 347-354.

Pakull, B., Mader, M., Kersten, B., Ekue, M.R.M., Dipelet, U.G.B., Paulini, M., Bouda, Z.H.N., Degen, B., 2016. Development of nuclear, chloroplast and mitochondrial SNP markers for Khaya sp. Conserv. Genet. Resour. 8, 283-297.

Piry, S., Alapetite, A., Cornuet, J.M., Paetkau, D., Baudouin, L., Estoup, A., 2004. GENECLASS2: A software for genetic assignment and first-generation migrant detection. Journal of Heredity 95, 536-539.

Rannala, B., Mountain, J.L., 1997. Detecting immigration by using multilocus genotypes. Proceedings of the National Academy of Sciences 94, 9197-9201.

Wright, S., 1965. The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution, 395-420.