****

****** readme file of FreeNA******

F_{ST} Refined Estimation by Excluding Null Alleles

****

Standard disclaimer

This program is provided "as-is". The authors and providers disclaim all warranties, expressed or implied, regarding the performance of this software. You may distribute this program freely in any format, so long as the following conditions are met: the program remains intact without modification, the readme file is included without modification, no fee of any kind is charged.

Description

FreeNA is a PC computer program which performs three major tasks:

1) it estimates null allele frequencies for each locus and population analysed following the Expectation Maximization (EM) algorithm of Dempster, Laird, and Rubin (1977) as described in the Supplementary Material of Chapuis and Estoup (2007). This estimator of null allele frequency was chosen because it shows better performance than the estimators of Chakraborty et al. (1997) and Brookfield (1996) (see Chapuis and Estoup (2007) for a comparative study).
2) it estimates F_{ST} following Weir (1996) from any microsatellite dataset harboring null alleles; the method used for such estimation is the so-called *ENA* method described in Chapuis and Estoup (2007). The *ENA* correction method was found to efficiently correct for the positive bias induced by the presence of null alleles on F_{ST} estimation

and provide accurate estimation of F_{ST} in presence of null alleles. This method is based on the following steps:

(i) estimation of null and visible allele frequencies for each locus and population;

(ii) allocation of a single allele size not present in the original dataset to null alleles;

(iii) adjustment of genotype frequencies based on the previous estimations of allele frequencies (for details see the '*Step 2*' of the section '*Simulation method*' in Chapuis and Estoup (2007));

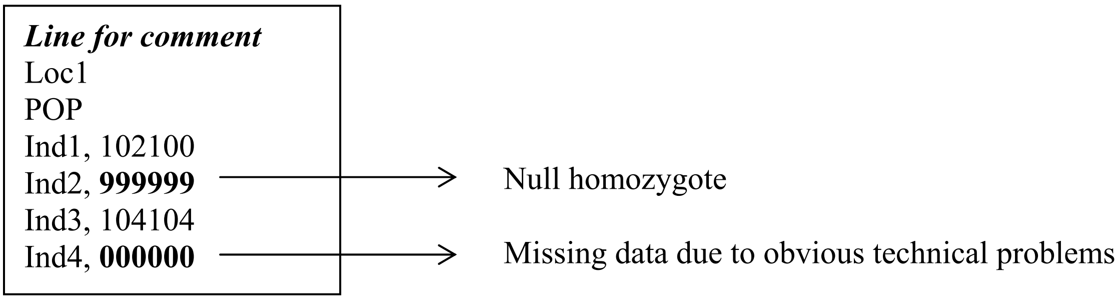
(iv) calculation of Weir's (1996) F_{ST} restricting the computation to visible states. The null allele state is hence ignored in computation and the sums of allele and genotype frequencies are not adjusted to 1. This is feasible because F_{ST}-estimate at a given locus is the appropriate combination of single allele estimates of F_{ST} (Weir 1996).

3) it calculates the Cavalli-Sforza and Edwards' (1967) genetic distance from any microsatellite dataset harboring null alleles; the method used for such estimation is the so-called *INA* method (Chapuis and Estoup 2007). The *INA* correction method provided slightly biased estimates of Cavalli-Sforza and Edwards' (1967) genetic distance, but give better estimates than those obtained without correction. The *INA* method follows the steps (i) to (ii).

The input file

FreeNA is compiled for a minimum of one allele and a maximum of 98 alleles per locus. The input file is a text file with a format identical to that of the population genetics package GENEPOP (alleles coded with 2 or 3 digits) (Raymond and Rousset 1995; see also "Genepop on the web"). Name your file simply (e.g., accents are not tolerated). The file name may have an extension or not. The title line cannot be empty. In the Dempster et al.' (1977) approach, genotypes that exhibit no bands (i.e., null genotypes) are assumed to represent genotypes homozygous for a null allele (i.e., null homozygotes). In FreeNA, non visible genotypes due to obvious technical problems (e.g., degraded or low quantity of DNA; any PCR problems during PCR reaction) can be specified in the dataset, and hence distinguished from null homozygous genotypes. In the input file, a genotype ‘9999’ (for 2 digits allele coding) or ‘999999’ (for 3 digits allele coding) represents a null homozygote while a genotype ‘0000’ (for 2 digits allele coding) or ‘000000’ (for 3 digits allele coding) represents missing data due to the above mentioned technical problems (see Figure 1).

Figure 1. Example of input file.



```
Line for comment
Loc1
POP
Ind1, 102100
Ind2, 999999
Ind3, 104104
Ind4, 000000
```

If FreeNA failed to read your input file, you can convert it in a GENEPOP format using the computer package GENECLASS2 (free of charge downloading from http://www.montpellier.inra.fr/URLB/) and try again.

The output files

The outputs of the program are provided in five different files. For all five files, populations were re-named by their filing number and allele sizes were re-numbered as 2-digit numbers (GENEPOP format; note that the correspondence between allele sizes and allele numbers is given in the 2nd file).

- The 1st file (your_output_file_name.r) gives the estimate of null allele frequency using the EM algorithm of Dempster, Laird, and Rubin (1977), for each population and locus.

- The 2nd file (your_output_file_name.fr) gives the allele frequencies and genotype numbers estimated using the EM algorithm of Dempster, Laird, and Rubin (1977), for each population and locus. These values are those used for computing corrected F_{ST} values.
- The 3rd file (your_output_file_name.gFst) gives the global F_{ST} values (for all loci and for each locus separately) computed both using and without using the *ENA* correction method described in Chapuis and Estoup (2007). If the number of loci are sufficient (more than four), bootstrapping over loci are automatically performed. The resampling is repeated until *n* global F_{ST} values, computed both using and without using the *ENA* correction method, are obtained. *n* is an integer between 1,000 to 10,000 specified by the user in the menu. The lower and upper bounds of the 95% confidence intervals are obtained by taking the 2.5% and 97.5% quantiles of the distributions of bootstrapped global F_{ST} values, respectively. See Raymond and Rousset (1995) for a discussion of some problems encountered with bootstrapping when there is not a sufficient number of loci.
- The 4th file (your_output_file_name.pFst) gives for each pair of analyzed populations the F_{ST} value (for all loci and for each locus separately) computed both using and without using the *ENA* correction method described in Chapuis and Estoup (2007). If the number of loci are sufficient (more than four), bootstrapping over loci are automatically performed as described for global F_{ST} values in the 3rd file description.
- The 5th file (your_output_file_name.dc) gives for each pair of analyzed populations the value of the Cavalli-Sforza and Edwards' (1967) genetic distance (for all loci and for each locus separately), computed both using and without using the *INA* correction method described in Chapuis and Estoup (2007). If the number of loci are sufficient (more than four), bootstrapping over loci are automatically performed as described for global F_{ST} values in the 3rd file description.

How to run FreeNA

Put the file FreeNA.exe (executable file) in the same directory as your data file. Double click on FreeNA.exe. Following the menu, enter the names of your input and output files. Choose an output file name without extension. Fix the number of replicates (an integer between 1,000 and 50,000) for the computation of the bootstrap 95% confidence intervals automatically performed for the global F_{ST} statistics computed both using and without using the *ENA* correction method described in Chapuis and Estoup (2007). Note that bootstrapping is performed only when there is more than four loci in the data set.

FreeNA Citation

Please cite the following reference if you use FreeNA:

Chapuis, M.P., and A. Estoup. 2007. Microsatellite null alleles and estimation of population differentiation. *Mol. Biol. Evol.* 24(3): 621-631.

Contact

In case of trouble, please send an e-mail to: chapuimp@supagro.inra.fr.

References

Brookfield, J. F. Y. 1996. A simple new method for estimating null allele frequency from heterozygote deficiency. *Mol. Ecol.* 5:453-455.

Chakraborty, R., M. De Andrade, S. P. Daiger, and B. Budowle. 1992. Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. *Ann. Hum. Genet.* 56:45-57.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J Roy. Stat. Soc.* B 39:1-38.

Nei, M. 1987. Molecular Evolutionary Genetics. Columbia University Press.

Raymond, M., and F. Rousset. 1995. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Heredity* 86:248-249.

Weir, B. S. 1996. Genetic Data Analysis II. Sinauer Associates, Sunderland, Mass.